Linking the Austrian Standards-Based Test for English to the CEFR: The Standard-Setting Process

Technical Documentation – Standards-based Proficiency Test in English, 8th grade, 2013
Linking the Austrian Standards-Based Test for English to the CEFR: The Standard-Setting Process
Technical Documentation – Standards-based Proficiency Test in English, 8th grade, 2013
BIFIE | Department Bildungsstandards & Internationale Assessments (BISTA),
Salzburg 2013

# Inhaltsverzeichnis

## 1 General Introduction
- 1.1 The Austrian Level-System
- 1.2 Relevant Standard-Setting Procedures
  - 1.2.1 The Contrasting-Groups Method
  - 1.2.2 The Bookmark Method
  - 1.2.3 The Item-descriptor Matching method (IDM)
  - 1.2.4 The Basket method
- 1.3 Familiarisation and Training of the Panelists

## 2 Productive Skills: CEFR-Linking for Writing
- 2.1 Workshop Schedule and Participants
- 2.2 Familiarisation
- 2.3 Script Selection and Rating Process
- 2.4 Definition of Cut-Scores
- 2.5 Conclusion and Final Cut-Scores
- 2.6 Evaluation

## 3 Productive Skills: CEFR-Linking for Speaking
- 3.1 Workshop Schedule and Participants
- 3.2 Familiarisation
- 3.3 Video Selection and Rating Process
- 3.4 Definition of Cut-Scores
- 3.5 Conclusion and Final Cut-Scores
- 3.6 Evaluation

## 4 Receptive Skills: CEFR-Linking for Reading and Listening
- 4.1 Workshop Schedule and Participants
- 4.2 Item Selection
- 4.3 Methodological Considerations
- 4.4 Familiarisation and Training
- 4.5 Standard-Setting Rounds
- 4.6 Rater Analyses
- 4.7 Final Cut-Scores for the Receptive Skills
- 4.8 Evaluation

## 5 Summary and Conclusion

## References

## 6 Annexes
1 General Introduction

This technical documentation provides information on the process of linking the Austrian standards-based proficiency test in English for the 8th grade to the Common European Framework of Reference for Languages (CEFR). The implementation of the linking process is primarily based on the suggestions of the manual of the Council of Europe (Council of Europe, 2009). The CEFR-linking process involves – as proposed in the manual – five steps or activities. These steps encompass the whole process of test development and are not necessarily in a hierarchical or chronological order (cf. Figure 1):

1. **Familiarisation:** This step circumscribes a set of activities for item-writers and standard-setting judges to familiarise with the CEFR and the national standards test. In the standard-setting workshop, it is necessary for participants to share a common knowledge of CEF-levels, information of test and item development and their task in the standard-setting.

2. **Specification:** The content analysis of the examination to elaborate a linking claim of how the tests relate to the CEFR. This step needs to be performed prior to the standard-setting and involves internal and external validation of the test. This process is strongly inherent during item development. We conducted several trials and a pilot study to examine the psychometric quality of the test and its items (details of this process are not part of this technical report).

3. **Standardisation training / benchmarking:** For standardizing the understanding of CEF-levels and for training of the task required during standard-setting, participants received several materials, for example, calibrated test-items and benchmark performances.

4. **Standard-setting:** The main focus of this report is on the the standard-setting. In general, standard-setting circumscribes a multidisciplinary decision-process encompassing people from a broad range of professional fields (stakeholders, didacticians, test-developers, psychometricians, politicians, teachers, etc.). One thing that all of the standard-setting methods have in common are workshop meetings of expert judges (Cizek & Bunch, 2007; Cizek, Bunch & Koons, 2004). Within these workshops, the judges try to reach consensus about setting cut-off-scores (or cut-scores) between performance levels by application of standardized methods (standard-setting methods). Thus, the primary goal of a standard-setting workshop is to establish cut-scores that enable the allocation of student's performances to discrete performance levels (Cizek & Bunch, 2007; Hartig, 2007).

5. **Validation:** This phase involves all steps of validating the complete process of CEFR-linking and also the validation of the test itself. With respect to the standard-setting this includes rater-analysis, quantitative and qualitative evaluation of the appropriateness of the cut-scores.

---

1 A preliminary version of this manuscript was reviewed by Dr. Neus Figueras, Departament d’Educació, Generalitat de Catalunya, Spain.
2 English as first foreign language only.
1.1 The Austrian Level-System

The Austrian standards test for English as first foreign language assesses the four subskills writing, speaking, listening and reading in three-year intervals. The first nation-wide assessment started in April 2013. In the writing test, students are asked to produce one short (40–70 words) and one long script (120–180 words) in 30 minutes (Gassner, Mewald, Brock, Lackenbauer & Siller, 2011). The speaking test is only administered to a sample of eighth grade students in 120 Austrian schools. The students are tested in randomly selected pairs and the tasks include one monologue and two dialogues. The tests for the receptive skills listening and reading mainly involve closed-format items3. More detailed information on the item-development process and test specifications can be found in Gassner et al. (2011) and BIFIE and ÖSZ (2011).

---

3 Sample items can be found on the BIFIE-website: www.bifie.at.
For the standard-setting, we followed a compensatory, in contrast to a conjunctive strategy (Haladyna & Hess, 2000) within each language skill. Usually, the reliability of total test-scores is increased in compensatory models under the assumption that the total score best reflects the test construct. As a consequence, the standard-setting was based on the total score metric within the four different language skills, reading, listening, writing, and speaking. Hence, CEFR-linking for subskills within those skills was not being approached.

The Austrian national educational standards for English are primarily based on A2 and B1 descriptors of the CEFR (BIFIE & ÖSZ, 2011, p. 11). As a consequence, a competency model with three levels was proposed with levels A1 or below, A2 and B1 or above. Performance level descriptors (PLDs) were developed by our consultancy groups and the BIFIE for levels A2 and B1 or above. The contents of the PLDs were cross-validated by didacticians from the BIFIE and external consultants. In regard to the standard-setting workshop, two cut-scores needed to be established, between A1 or below vs. A2 and A2 vs B1 or above.

### 1.2 Relevant Standard-Setting Procedures

Methods for standard-setting can either be categorized as examinee-centered or test-centered (Kane, 1998). Examinee-centered approaches contain the rating of student’s work (e.g., scripts, dialogues), whereas decision processes within test-centered approaches are focused on the test-items. On the basis of such expert-ratings (either of items or student’s work), the determination of cut-scores is grounded.

In the Austrian standards-based performance test, we decided to use test-centered approaches to link to the CEFR-levels for the receptive skills reading and listening. This line of reasoning is simply due to the fact that most of the items used in these domains are closed-format items (e.g., multiple-choice items). For the productive skills writing and speaking, on the other hand, we applied an examinee-centred approach, as student’s work is directly available.

Based on either examinee- or test-centered approaches, there are different methods to choose from for arriving at cut-scores. The contrasting-groups method is adequate for an examinee-centered approach, whereas the bookmark method, the item-descriptor matching method and the basket method are suitable for a test-centered approach. In the following, these methods are described in more detail.

#### 1.2.1 The Contrasting-Groups Method

As we needed an examinee-centered approach for establishing a cut-score in writing and speaking, we used the contrasting-groups method (Bechger, Kuijper & Maris, 2009; Berk, 1976; Cizek & Bunch, 2007; Council of Europe, 2009). In the simplest case where one cut-off-score has to be determined (e.g., between pass and fail or B1 and B2), the judges are asked to indicate for each piece of student’s work (e.g., a script) to which of the two groups it belongs to. For example, if a cut-score between B1 and B2 was the goal, the question posed to the judges would be “To which performance level would you assign the student’s work?”. Each script would thus receive an individual mark from a respective judge (or some judges, depending on the rating design) relating to the respective competence group (B1 vs. B2). Next, two frequency distributions of total test scores, representing either the group members rated B1 or B2, can

---

4 Kaftandjieva (2010) lists 62 different standard-setting methods, extensive reviews on standard-setting methods are provided by Cizek and Bunch (2007); Cizek et al. (2004); Hambleton (2001, 2000).
be plotted. Figure 2 is adopted from the COE manual and shows hypothetical data of two groups rated B1 and B2, respectively. Here, the judges rated students’ work on a test where a maximal test score of 50 was possible. Two frequency distributions can be generated representing the two performance groups. As can be seen, the distributions show a clear overlap, this means, some students scoring higher receive a rating belonging to the lower performance group and vice versa. To determine a cut-score, one could calculate the mean of each group distribution and subsequently the mean of these two group means. In this example, a score near 25 would best reflect the cut-score. However, using this score generates misclassifications, as can be seen in the table of Figure 2. For example, 4 students with a test score of 27 were classified as B1 (false positive) and 6 with a score of 23 as B2 (false negative). 19 false negatives (FNs) and 18 false positives (FP) would result in a total of 9.3% misclassifications. One primary goal of standard-setting could be to find a cut-score that best reduces the sum of false positives and false negatives and thus the percentage of misclassifications.

It is noteworthy to mention that the method of reducing the total percentage of misclassification to define a cut-score is not appropriate for every testing purpose. It is important to challenge the question of the potential impact of classifying pupils as either false negatives or positives. Alternative approaches for finding the cut-scores are presented in the manual of the COE and also by Bechger et al. (2009). In our approach, we did not use a special adjustment (e.g., weighting) of cut-scores to control for misclassifications. This means that FNs and FPs were regarded as equally critical. The methods (e.g., ROC, logistic regression) we implemented for the standard-setting of the productive skills, implicitly search for an optimal classification of cut-scores, but we did not a priori define a criterium for reducing either FNs or FPs.

1.2.2 The Bookmark Method

The main feature of the bookmark method, which was used for setting cut-scores for the receptive skills, is an Ordered Item Booklet (OIB), a booklet where all items are arranged in increasing order of their difficulty (Mitzel, Lewis, Green & Patz, 1999; Mitzel, Lewis, Patz & Green, 2001). This indicates that items that have already been pilot-tested are required for this procedure. Item-difficulty can be derived from concepts of classical test theory as well as item response theory. On every single page of the booklet a single item is depicted. Scale location (i.e., item-difficulty), answer-keys, notes, etc. can be printed additionally along with the item-text.

The bookmark method incorporates the concept of borderline knowledge or borderline persons. The task of the panel members is to imagine a group of pupils at the border of two proficiency levels that would answer a given test-item with a certain chance or probability. Usually, people are asked to imagine a group of pupils that would answer a question in two out of three possibilities, corresponding to a probability of 2/3. This so-called Response Probability (RP) can be defined arbitrarily, but, most critically, the final cut-scores will be affected by the choice of the RP (Wyse, 2011). Hence, it is inevitable that panel members show a clear understanding of the concept of the RP, which has to be considered for training of the participants in the standard-setting workshop.
Linking the Austrian Standards-Based Test for English to the CEFR: The Standard-Setting Process

Figure 2: The contrasting groups method. A) Judges are asked to rate student's work as either B1 or B2 performances. Two performance groups, plotted as frequency distributions, result from this rating process. To define a cut-score the midpoint of the two distributions can be calculated reflected by the mean of both group-means. B) The cut-score defined in A) lies at a test-score value of 24, for example. This cut-score leads to 19 false negatives (FNs) and 18 false positives (FP) and a total of 9.3 % misclassifications. Figure adapted from the COE manual.
If panelists are asked to set a cut-score between level A1 and A2, they work through the OIB, constantly answering the question whether a minimally-qualified person (at the border between the two levels) would show a correct response in 2 out of 3 cases (RP = 2/3). If the goal is to establish a cut-score between level A1 and A2, the question asked to the panel members is “Can a borderline student – between A1 and A2 – answer this item correctly in 2 out of 3 cases?” If the answer is yes on a single item, they proceed to the next item, if the answer is no, the panel members place a bookmark on the previous item, that is, the last item that the student could master. Note that in some procedures, as it was the case for this standard-setting, the bookmark is placed on the first item that the borderline student could not give a correct response to (and strictly speaking, the bookmark would be between the last item a student could master and the first item he could not master, Cizek & Bunch, 2007). Regardless of the instructions, it is important to make it clear to the participants of the standard-setting, that up to this bookmark-item a borderline student would answer the items correctly with the respective RP (or higher). After the bookmark is placed for a certain cut-point, the panelists move on to the next cut-score region.

Usually, cut-scores (or bookmarks) are set in two to three consecutive rounds, between the rounds, feedback is provided to the panel members. Feedback includes descriptive statistics of cut-score placements and frequency distributions of performance data (impact data).

An advantage of the bookmark method is the relatively simple task, on the other hand, the concept of the RP is sometimes not easily understood and requires some time for training (Karantonis & Sireci, 2006).

### 1.2.3 The Item-descriptor Matching method (IDM)

The IDM is very similar to the bookmark method. But the full understanding of the concept of a response probability is not a necessary requirement for panel members to perform the task (Ferrara, Perie & Johnson, 2002). It has to be noted that the concept of the RP is relevant during calibration of items, but it is not for undergoing the task. The task in the IDM is to rate every item by matching the cognitive demands of an item that are relevant for a correct solution to the descriptions of the performance levels. Hence, it is not relevant to imagine a group of students handling the item demands in 2 out of 3 cases.

The items are also presented in OIBs, ordered by increasing item difficulty. The task for the participants is usually to match items to the performance levels. The IDM is easy to implement and the cognitive load for a panel member to undergo this task is manageable. Otherwise, the evaluation of the ratings is not straightforward and bears some technical as well as practical problems (e.g., the concept of the threshold regions, Cizek & Bunch, 2007). A modified version of the IDM was also implemented for the Mathematics standard-setting (Freunberger, 2013).

### 1.2.4 The Basket method

The question asked to the panel members in the Basket method is “At what CEFR level can a test taker already answer the following item correctly?” The task is quite similar to the task in the IDM, whereas in the Basket method no difficulty-ordering of the items takes place (Kaftandjieva, 2010). Each item is placed in a basket that relates to a specific CEFR level. For example, items in basket A2 should be answered correctly by students at (or above) performance level A2.
Given that the difficulty-ordering of items (along with the numerical item-difficulties) in the IDM method can easily confuse participants of standard-setting workshops, it is advisable to not present items in an ordered-item-booklet or at least to omit the numerical values of item-difficulties (Council of Europe, 2009). Thus, the basket-method seems a very easy-to-implement method, but in its application, it is also one of the most lenient methods. For example, the cut-scores between each performance level are calculated by summing up the number of items put in the respective baskets. If there are four items in the level-1-basket and six in the level-2-basket, the cut-score between level 2 and level 3 would be the sum of basket one and two (4 + 6 = 10 items). A student on level 3 should at least answer 11 (10 + 1) items correctly. Hence, this method provides minimal-standards for each level. But, a student on level 3 is probably also able to answer some questions on higher levels. So, this method considers minimal requirements only. Another problem arises if no items are placed in some of the baskets. In such a case, no cut-scores can be set between certain levels.

According to Kaftandjieva (2010), some of the main advantages of the method are: Its economic use because of usually only one round of judgment; its simplicity; it does not require complex statistical analyses for setting cut-scores; its wide range of applicability; its clear interpretability; and it can be applied prior to test administration. Because it does not incorporate empirical test data, this can be equally regarded as disadvantage. This means that there exists no link between empirical item-difficulty and judgements made by the panelists (which usually should be taken into account as validity argument).

Alternatives to the Basket method are the Compound Cumulative method and the Cumulative Cluster method. Both methods are similar in their judgment processes but the calculations of the cut-scores are more complex and they both take empirical data into account, which increases the validity of these methods as compared to the original Basket method.

### 1.3 Familiarisation and Training of the Panelists

Familiarisation refers to the process where judges become familiarised with the CEFR, training is more related to the rating process of the standard-setting. This phase will generally be similar for productive and receptive skills, although several modifications are specific for certain skills (e.g., rating of videos for speaking, rating of scripts for writing).

The familiarisation activities, suggested in the manual, are presented in four main categories from which we chose the following, with some modifications tailored to our test-purposes (see the Appendix for exemplary activities and tasks performed during the familiarisation phase):

1. **Preparatory activities before the seminar:**
   Material was sent to participants five weeks prior to the workshops. This included section 3.6 of the CEFR, a link to the CEFTrain website[^5] and some instruction for familiarising with the CEFR. The training on the CEFTrain homepage included *scale training* (e.g., ‘ball game’) and *assessment training* (rating of texts and videos). With respect to section 3.6 of the CEFR manual, participants were asked to elaborate the differences among the CEFR-descriptors.

[^5]: [http://www.helsinki./project/ceftrain](http://www.helsinki./project/ceftrain)
2. **Introductory activities at the seminar:**
   A brief introduction was given on the CEFR in general, presenting its main features, and most activities mentioned in point 1 above were repeated in this phase. A brief discussion on salient features of CEFR levels was included and participants were asked to assign a set of different PLDs selected from several scales to the CEFR levels. The participants also received aggregated feedback on rating performances as basis for further discussions.

3. **Qualitative activities of the CEFR scales:**
   The relevant CEFR descriptors were discussed for the test under consideration and for the respective dimensions.

4. **Preparation for rating: productive skills.**
   Ratings on exemplar scripts or videos had to be performed. For example, for writing, several scripts (from the COE or DIALANG homepages) had to be judged. Participants made use of can-do-statements (CDS) for A2 and B1. They rated a set of five scripts by using A2 CDS and set of three scripts by using the B1 CDS. The 6 CDS for writing and speaking were developed by our consultancy team and should best cover the Austrian standards test. For speaking, four videos (two for each level, downloaded from the CIEP – Centre international d’études pédagogiques – homepage[^6]) had to be rated using either A2 or B1 CDS developed for the speaking standard-setting.

5. **Preparation for rating: receptive skills.**
   Participants had to rate items according to the descriptors. The standard-setting method was explained and ratings on exemplar items from the national standards test had to be performed.

2 Productive Skills:  
CEFR-Linking for Writing

2.1 Workshop Schedule and Participants

The CEFR-linking workshop for writing took place at the BIFIE Salzburg from 5th–7th of November 2012. One half-day was spent with a familiarisation phase, the rest of the workshop time was spent on judging of students’ performances and discussions. The group of expert judges consisted of 14 people with professional background in the fields of language testing, teaching and psychometrics. The most strongly represented expert’s group (7 persons) was teacher trainers (e.g. lecturer at college of education). 4 out of 7 participants were teachers at schools, 2 participants had a professional background solely as teachers, and 2 participants were employees from the BIFIE. More than half of the respondents had an additional occupation within the framework of educational standards in Austria or were involved in test development. More details on participants’ background can be found in (Zuber, in press). The workshop language was English.

2.2 Familiarisation

The familiarisation phase was supervised and conducted by our external consultants, Otmar Gassner and Rainer Brock, all materials and analyses were provided by the local organizers. Prior to the workshop meeting, participants received training activities that included working on the CEFtrain homepage. During the workshop the first half-day was spent on familiarising with the CEFR. This included a presentation to the CEFR with specific emphasis on the scales for writing, training activities and group discussions. Five different familiarisation activities helped to maximise the degree of familiarity with the CEFR and prepared participants for the task of the standard-setting.

To evaluate the degree of familiarity with the CEFR-scales, two training activities asked for assigning CEFR-levels to performance descriptors selected from different writing scales. The ratings were then evaluated by local organizers and the results presented to the expert-panel. Two additional activities focused on the rating of 8 sample scripts, and specific emphasis was put on the distinction between levels A1, A2 and B1. Then, the standard-setting rating task was introduced and trained with the sample scripts. The results from the second training activity (assignment of CEFR-levels to descriptors) are shown in Figure 3 and Figure 4.

Figure 3 shows two plots that were also presented for feedback within the familiarisation phase. Here, the group performance is shown, whereas in Figure 3 (left) the correct assignment per level is depicted, Figure 3 (right) provides information about discrepant rating behaviour (i.e., when an A2 descriptor is rated as A1 or B1, this results in a discrepancy of 1). Both figures show that the group performance was high, especially for A1 descriptors (85.9 % correct assignment). Figure 4 was not presented as feedback to the panel members, it provides information of rating behaviour for each judge with average correct assignments of 71.8 %. Figure 4 (inlet) shows mean Kappa-coefficients for each judge contrasted with the standard deviations. A low mean value indicates low agreement of one judge with all the other judges. The scatter-plot contrasts mean-Kappas with their standard-deviations (SDs). A low mean-Kappa value and a low SD would indicate anomalous rating behaviour, where a respective judge would consistently disagree with the group.
2.3 Script Selection and Rating Process

We used the contrasting-groups method (described above) to determine cut-scores. The experts’ task was to judge 120 scripts produced by students in a pre-pilot trial in two consecutive rating rounds. Each student received two writing tasks, one aiming at level A2 (short prompt) and one aiming at level B1 (long prompt). The scripts were separately scaled by using many-facet Rasch measurement (Eckes, 2011; Linacre, 1994) and scores were transformed onto the PISA score metric with a mean of 500 and a standard-deviation of 100. We chose to use the PISA score metric as participants were more familiar with this metric than with the traditional Rasch logit metric.

For the cut-score A1/A2, judges rated 60 scripts (from 3 short prompts and 1 long prompt) by using 6 can-do-statements (CDS, cf. Table 1). Each statement had to be answered with Yes or No. The ratings of the videos (in speaking) and scripts (writing) were performed online using a web-based computer software (see the Appendix for an example). The script performances were selected in a way that lower performances were over-represented. In a second round, judges were asked to rate another set of 60 scripts (from 3 long-prompts) using CDS for B1 (Table 1). Here, we selected scripts with performances that spread evenly across the proficiency scale, with a slight over-representation of higher performances.
Figure 4: Results from training activity 2 (not presented to the panel) for writing. A) Number of correct assignments per judge (red line represents the group mean, grey lines the standard-deviation). B) Mean-Kappa (top) and mean-Kappa contrasted with standard-deviations (bottom) per judge.

<table>
<thead>
<tr>
<th></th>
<th>A2</th>
<th>B1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Can write a series of simple phrases and sentences about their family, living conditions, educational background, events, past activities and personal experiences.</td>
<td>Can write straightforward, detailed texts on a range of familiar subjects.</td>
</tr>
<tr>
<td>2</td>
<td>Can expand learned phrases through simple recombinations of their elements and communicate simple information.</td>
<td>Can convey simple, straightforward information and exploit a wide range of simple language flexibly to express much of what he/she wants.</td>
</tr>
<tr>
<td>3</td>
<td>Can link groups of words and simple sentences with simple connectors like „and“, „but“ and „because“ and possibly other frequently occurring connectors.</td>
<td>Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.</td>
</tr>
<tr>
<td>4</td>
<td>Has a (limited) repertoire of basic language and sentence patterns, which enables him/her to deal with simple everyday situations with predictable content, though he/she will generally have to compromise the message.</td>
<td>Has enough language to get by, with some restrictions, but lexical and structural limitations may cause repetition and even difficulty with formulation at times.</td>
</tr>
<tr>
<td>5</td>
<td>Uses some simple structures correctly, but (may) still systematically make basic mistakes; nevertheless, it is usually clear what he/she is trying to say.</td>
<td>Uses a generally sufficient range of structures correctly and communicates with reasonable accuracy. Errors may occur, but it is mostly clear what he/she is trying to express.</td>
</tr>
<tr>
<td>6</td>
<td>Can control a sufficient vocabulary for the expression of basic communicative needs and can cope with simple survival needs.</td>
<td>Can control a sufficient vocabulary to express him/herself on most topics pertinent to his/her everyday life. Spelling is accurate enough to be followed most of the time.</td>
</tr>
</tbody>
</table>

Table 1: Can-do-statements for writing used in the contrasting-groups method.
2.4 Definition of Cut-Scores

For each student’s performance (script) we yielded 6 ratings (yes-no-statements for each CDS) per judge (cf. Table 1). The next step was to define a criterion that allowed us to classify a student’s performance – on basis of the CDS ratings – as either being an A2 performance or not (in case of the first cut-score). From our ratings, we calculated the sum across ratings for each judge; a certain script can thus have a maximum score of 6 (all CDS were endorsed) and a minimum score of 0 (no CDS was endorsed). These steps were performed after the workshop due to time constraints, and therefore, cut-scores and impact data were not reported to the participants.

The sum scores were then categorized into 2 categories: scripts with sum scores between 0 and 3 were defined as being below the respective CEFR-level (e.g., below A2, below B1), scripts with sum scores between 4 and 6 were defined as being performances at a respective level (e.g., at A2 or at B1). The categorization was performed on a single-judge-level.

To define cut-scores on the proficiency scale, we applied three different procedures: (i) midpoint analysis, (ii) logistic regression, and (iii) receiver operator characteristics. The midpoint-procedure (Cizek & Bunch, 2007) is easy to implement and the cut-score is defined as the mean of the two group means from the two proficiency distributions (e.g. A2 vs. not-A2).

Logistic regression can be used to determine the point of intersection of two distributions (Cizek & Bunch, 2007, p.109). The common definition of logistic regression enables to find a score point where the probability of belonging to a certain category is $p = .50$ (Gelman & Hill, 2007; Sireci & Clauser, 2001). This is given as

$$\ln\left(\frac{p_k}{1 - p_k}\right) = \alpha + \beta x,$$  \hspace{1cm} (1)

where $p$ is the probability of belonging to category $k$ and $x$ is the cut-score. After estimating the regression coefficients $\alpha$ and $\beta$ the equation can be solved for $x$ which yields the cut-score, i.e., the intersection point of the two group distributions. If we assume a probability of $p = 0.50$, $\ln\left(\frac{0.5}{0.5}\right) = \ln(1) = 0$, so $x = \frac{50}{5}$ which is the point where above that score the likelihood of belonging to the category (e.g. A2 vs. A1) is higher than 50 %.

Receiver operating characteristics (ROC) graphs are mainly used in medical research for optimal decision making in diagnostic testing (Fawcett, 2006; Kaftandjieva, 2010). A ROC graph usually contrasts the true positive rate with the false positive rate and provides information about the optimal score that best controls for false positives and true positives. We adopted this method for every single judge in a way that every latent score could reflect a potential cut-score. As we used 60 scripts, we could gain 60 different cut-scores. For example, the method selects a potential cut-score and calculates the true positive rate (e.g., scripts rated as being at A2 and also classified by that cut-score as A2) and false positive rate (e.g., scripts rated as being below A2 and classified by that cut-score as A2). The true positive rate is usually termed sensitivity and the false positive rate 1-specificity. The area under an ROC curve (AUC) gives an index of the performance of a certain classifier (i.e, the quality of the judgements). To find the optimal cut-score we used the Youden index implemented in the R-package pROC (R Core Team, 2012; Robin et al., 2011). Figure 5 shows ROC curves for one rater and for both cut-scores.
Figure 5: ROC (Receiver Operating Characteristics) curves for one rater. The cut-scores are represented as the optimal coordinates reflecting maximal sensitivity and specificity. The AUC provides information about the rating quality.

Figure 6 depicts the overall cut-scores (i.e., averages over all judges with standard-errors) comparing the three different methods. A repeated measures ANOVA with factors Method (log. regression, midpoint analysis, ROC) and Cut-score (two cut-scores) was applied to compare the different methods. Results showed significant effects for factors Method \( F(2,24) = 13.30; p < .01 \) and Cut-score \( F(1,12) = 409.16; p < .001 \) and a significant Method x Cut-score interaction \( F(2,24) = 26.79; p < .001 \).

Figure 6: Average cut-scores with standard errors for three different analysis approaches.
2.5 Conclusion and Final Cut-Scores

A major issue in discussing the most adequate procedure for cut-score determination is the concept of misclassifications (cf. section 1.2.1). Especially in large-scale assessments where stakeholders would like to initiate washback-effects, the standards should be more conservative and in consequence more stringent. Also, the main focus of the Austrian national educational standards is system monitoring and school development, and, hence not single student competency diagnostics. Thus, we preferentially apply a method that equally weights types of misclassifications. The logistic regression approach seems here the most adequate method, as well as the midpoint-analysis. Logistic regression is frequently used in the contrasting groups method and highly accepted, therefore the cut-scores obtained from this procedure define the cut-scores for English writing. But, it has to be noted that since the use of different methods yielded highly different cut-scores (see section 2.4), the outcomes of standard-settings must be treated very cautiously.

By using the cut-scores from the logistic regression approach, the cut-score between A1 or below and A2 is 464.46 and the cut-score between A2 and B1 or above is 559.67. Cut-scores (and also impact data) could not be presented to the participants because cut-scores were calculated after the workshop meeting.

2.6 Evaluation

For every standard-setting process, evaluation is important to check whether participants understood the task, had comments for improving the process or were unsure about the results (Morgan & Perie, 2004; Raymond & Reid, 2001). For this, participants received several questionnaires during the process; we report only the most important and notable results. A detailed evaluation report about the standard-setting process can be found in Zuber (in press).

Regarding necessary prerequisites before the actual rating-rounds, most participants were familiar with the CEFR, and most of them were confronted with it in their jobs, which is a good prerequisite for the standard-setting. All participants thought that the familiarisation phase was helpful for the method as well as for the process. The feedback after the workshop was favorable in general, and some participants made suggestions for future workshops: One person remarked that more discussion phases during the rating process would have been helpful, another one asked for impact data and feedback.
3 Productive Skills: CEFR-Linking for Speaking

3.1 Workshop Schedule and Participants

The CEFR-linking workshop for speaking took place at the BIFIE Salzburg from 4th–6th of December 2012. Regarding the background of the participants, 7 persons were teacher trainers (e.g., lecturer at college of education), 3 were teachers at schools, 2 participants had a professional background solely as teachers. 4 participants worked at research institutions. 3 persons also participated in the writing workshop. The workshop language was English.

3.2 Familiarisation

The familiarisation phase was supervised and conducted by external consultants, Claudia Mewald and Fiona Lackenbauer. All materials and analyses were provided by the local organizers. Prior to the workshop meeting, participants received training activities that included working on the CEFtrain homepage. During the workshop the first half-day was spent on familiarising with the CEFR. This included a presentation to the CEFR with special emphasis on the scales for speaking, training activities and group discussions. As for writing, there were five different familiarisation activities in order to establish the degree of familiarity with the CEFR and to prepare participants for the procedure of the standard-setting. Three training activities asked for assigning CEFR-levels to performance descriptors selected from different speaking scales. The ratings were then evaluated by local organizers and the results presented to the participants. An additional activity focused on the rating of 4 sample videos, specific emphasis was put on the distinction between levels A1, A2 and B1. These videos included non-native students speaking about various topics in English, and were therefore similar to those in the standard-setting itself. Discussion among the participants concentrated on the distinction of the CEFR-levels. Then, the standard-setting rating task was introduced here and trained with these videos. The results from the second training activity (assignment of CEFR-levels to descriptors) are shown in Figure 7 and Figure 8.

Figure 7 shows two plots that were also presented for feedback within the familiarisation phase, similarly to the feedback in writing (section 2.2). In Figure 7, the group performance is shown, as well as the correct assignment per level is depicted, and in addition, it provides information about discrepant rating behaviour (i.e., when an A2 descriptor is rated as A1 or B1, this results in a discrepancy of 1). Both figures show that the group performance is generally high. Figure 8 was not presented as feedback to the panel members. Figure 8 provides information of rating behaviour for each judge with average correct assignments of 80.8 %. Figure 8 (inlet) shows mean Kappa-coefficients contrasted with the standard-deviations.

3.3 Video Selection and Rating Process

We used the contrasting-groups method (described above) to determine cut-scores. The experts’ task was to judge 44 videos of students’ performances, recorded in a pre-pilot trial, in two consecutive rating rounds. The selection of the videos ensured to enable an overview of the proficiency scale. The performances were scaled by using...
many-facet Rasch measurement (Eckes, 2011) and scores were transformed onto the PISA score metric with a mean of 500 and a standard-deviation of 100.

For the cut-score A2/A2, judges rated 22 videos by using 6 can-do-statements (CDS, cf. Table 2). Each statement had to be answered with Yes or No. Similarly to writing, the ratings of the videos were performed online using a web-based computer software. The video performances were selected in a way that lower performances were over-represented. In a second round, judges were asked to rate another set of 22 videos using CDS for B1 (Table 2). Here, we selected videos with performances that were evenly spread across the proficiency scale, with a slight over-representation of higher performances.

Figure 7: Results from training activity 2 (presented to the panel) for speaking. Number of correct assignments per CEFR-level (left) and level-discrepancies (right).

Figure 8: Results from training activity 2 (not presented to the panel) for speaking. Number of correct assignments per judge (red line represents the group mean, grey lines the standard-deviation). The inlet shows mean-Kappas contrasted with standard-deviations per judge.
Table 2: Can-do-statements for speaking used in the contrasting-groups method.

<table>
<thead>
<tr>
<th></th>
<th>A2</th>
<th>B1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Can communicate limited information in a simple list of points.</td>
<td>Can convey simple, straightforward information as a linear sequence, getting across main points comprehensibly.</td>
</tr>
<tr>
<td>2</td>
<td>Can use simple questions to start, maintain, or end conversations.</td>
<td>Can initiate, maintain and close simple face-to-face conversations.</td>
</tr>
<tr>
<td>3</td>
<td>Can make himself understood in short contributions despite very noticeable hesitation and false starts.</td>
<td>Can keep going in longer stretches with some pausing for grammatical and lexical planning and repair.</td>
</tr>
<tr>
<td>4</td>
<td>Pronunciation is generally clear enough to be understood despite a possible foreign accent.</td>
<td>Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.</td>
</tr>
<tr>
<td>5</td>
<td>Can use basic sentence patterns and a repertoire of basic phrases, groups of words and formulae. He/she may have to compromise the message and search for words.</td>
<td>Has sufficient language to express himself/herself with some hesitations and circumlocutions, but lexical or structural limitations may cause repetition and even difficulty with formulation at times.</td>
</tr>
<tr>
<td>6</td>
<td>Can control a narrow repertoire of simple structures and vocabulary correctly, but still systematically makes basic mistakes. The message is usually clear.</td>
<td>Can control a repertoire of simple structures and vocabulary with reasonable accuracy. Errors may occur, but the message is clear.</td>
</tr>
</tbody>
</table>

### 3.4 Definition of Cut-Scores

For each of the 44 videos (22 for each cut-score) we yielded 6 ratings (yes-no-statements for each CDS) per judge (cf. Table 2). We followed a similar procedure as we did in writing for establishing whether a video would be classified at a certain level: we calculated the sum across ratings for each judge – a video could have a score ranging from 0 to 6, then the sum scores were categorized into 2 categories: videos with sum scores between 0 and 3 were defined as being below the respective CEFR-level (e.g., below A2, below B1), videos with sum scores between 4 and 6 were defined as being performances at a respective level (e.g., at A2 or at B1). The categorization was performed on a single-judge-level. As in writing, these steps took place after the workshop, therefore, cut-scores and impact data were not reported to the participants.

To define cut-scores on the proficiency scale, we used the aforementioned midpoint analysis (i.e., calculating the mean of the two groups mean, see Figure 2). For this, we applied the procedure for every judge, and calculated a mean cut-score from these individual scores. For determining the cut-scores for speaking, we proposed two other methods (i.e., logistic regression and receiver operator characteristics). While it is important to consider alternative approaches, the number of videos in this standard-setting was rather small due to time-constraints (22 videos per cut-score), leaving logistic regression and receiver operator characteristics as rather unjustified choices due to the small number of ratings per judge. As the midpoint analysis is presented in the COE manual (Council of Europe, 2009), is easy to apply and also easily communicated to the public, this method was selected.
3.5 Conclusion and Final Cut-Scores

As already mentioned (see section 2.5), the focus of the Austrian national educational standards test is on system monitoring and school development rather than diagnostics of single student competencies, therefore, we used a method that equally weighs different types of misclassifications. The midpoint-analysis (unweighted for misclassifications) seems an adequate method here, therefore, the cut-scores obtained from this procedure defined the cut-scores for speaking. However, it has to be noted that the alternative methods are possible, that could lead to different cut-scores (see section 2.4).

By using the cut-scores from the midpoint analysis the cut-score between A1 or below and A2 is 468.60 and the cut-score between A2 and B1 or above is 550.70.

3.6 Evaluation

Similarly to writing, participants received several questionnaires during the process; we report only the most important and notable results, more details can be found in the evaluation report (Zuber, in press).

All participants reported to be familiar with the CEFR in general as well as with the specific levels A1-B2. All participants reported that their job required a general or even more a specific knowledge of the CEFR. Moreover, all but one participant were familiar with the CEFR linking process prior to the standard-setting workshop. Again, most participants thought that the familiarisation phase was helpful for the method as well as for the process. The feedback after the workshop was favorable, with only minor suggestions for improvement.
4 Receptive Skills: CEFR-Linking for Reading and Listening

4.1 Workshop Schedule and Participants

For the receptive skills reading and listening a four-day workshop meeting (two days were assigned for each skill) was organised at the BIFIE Salzburg from 20th–23rd of November 2012. One half-day was spent with familiarisation, the rest of the workshop time was spent on rating and discussing items. The first half of the workshop was spent on reading, the second half on listening. Both skills required the rating of items and for both skills, we conducted the standard-setting in 4 rating rounds. Between each round, different feedbacks were given and group discussions were engaged. The group of expert judges consisted of 16 people with professional background in the fields of language testing, teaching and psychometrics, 3 persons also participated in the standard-setting for writing. The most strongly represented group (7 persons) is teacher trainers, 2 participants also work as teachers at school. 4 participants are employees at the BIFIE. More than half of the respondents have an additional occupation within the framework of educational standards in Austria or are involved in test development (see Zuber, in press). One person had to leave early and is thus not included in the data for listening. The workshop language was English.

4.2 Item Selection

The reading and listening items were calibrated using the Rasch-model (Rasch, 1960), and values were subsequently transformed onto the PISA-metric with a mean of 500 and a standard-deviation of 100.

For both skills 80 items were selected, descriptive statistics of item quality are presented in Table 3. For item-selection, the following criteria were defined (see also Harsch, Pant & Köller, 2010):

- Only items with good psychometric quality (e.g., discrimination indices, difficulty indices) were considered.
- Empirical item difficulty should be spread evenly across the latent scale.
- All item formats should be represented (one item format was not included due to incompatible characteristics with the standard-setting method).
- Good coverage of the test-construct.

<table>
<thead>
<tr>
<th>Item Statistics</th>
<th>Reading</th>
<th>Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>Min</td>
<td>24.80 %</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>67.71 %</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>89.30 %</td>
</tr>
<tr>
<td>Discrimination</td>
<td>Min</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 3: Item quality of the selected item-pool.
4.3 Methodological Considerations

For the standard-setting for the Mathematics standards-test by the BIFIE Salzburg, a modified version of the IDM was implemented (Freunberger, 2013; Freunberger & Yanagida, 2012). The modified version of the IDM was in parts also used for the English standard-setting.

The appropriate methods should be economic in use and easy to handle (i.e., feedback data, impact data, etc., should be quickly available and not dependent on complex analysis procedures). Also, the cognitive load for the participants should be held as low as possible. Additionally, we also wanted to implement two slightly different methods to be able to present diverse feedback to the participants. Therefore, a rating process similar as in the IDM or basked method was implemented and used to facilitate ratings and discussions on a single-item level. For determination of cut-scores a bookmark method was used. The bookmark as well as the IDM technique are methodologically related to item-mapping procedures, thus, prior determination of item-parameters (i.e., item-difficulties) is inevitable (Ferrara et al., 2002; Zwick, Senturk, Wang & Loomis, 2001). Interestingly, previous studies have shown that inter-judge consistency and also inter-rater-agreement is higher in item-mapping methods when compared to traditional Angoff procedures (Wang, 2003).

We implemented a rating-process that is similar to judgments performed in the IDM, where participants are asked to relate each item to the specific performance-level. Here, we asked: “What is the lowest level at which a test taker can answer the following item correctly?”, in round one and two. Subsequently, in round three and four, panelists were asked to place bookmarks between the different CEFR-levels. This blended procedure bears several advantages. First of all, for validating a standard-setting process, it is important to evaluate the rating behaviour of judges (Congdon & McQueen, 2000; Eckes, 2005; Engelhard & Stone, 1998; Engelhard, 2011; Kozaki, 2004; Lynch & McNamara, 1998; Nichols, Twing, Mueller & O’Malley, 2010). Using a method like the IDM allows to derive clear assumptions about the rater decisions or the rating behaviour. For example, as every item is rated, we can provide bar-charts as feedback to the panelists showing for each item the category frequency of the whole panel (e.g., item 1 was rated by 70 % of the participants with A2, and by 30 % as B1 item). This information supports the discussions after the first rating round, the panel can easily spot homogeneously rated items and heterogeneously rated ones. Also, the post-standard-setting process allows more statistical analyses about rater-agreement than this would be the case when using the original bookmark procedure solely.

4.4 Familiarisation and Training

For the receptive skills, judges received material and instructions prior to the workshop meeting (e.g., CEFTrain, section 3.6 of the CEFR, etc.). During the workshop, a brief presentation introduced the goal of the standard-setting and general information on the national standards tests was given. Additional presentations from our external consultancy group NILE (Norwich Institute for Language Education; Dave Allan, Keith Morrow, Thom Kiddle) introduced the CEFR and the reading and listening test. The participants were then asked to assign CEF-levels (from A1 to B2) to descriptors taken from the respective reading and listening scales. The results of these tasks are presented in Figure 9. Based on the feedback critical aspects of the descriptors and features, characteristics and keywords specific for the different levels were discussed. For the training of the rating process, six sample items from the item-
pool had to be judged according to the instructions from the IDM method before the first round started.

In general, group performance for the first familiarisation activity was acceptable (cf. Figure 9). As can be seen in Figure 10, for reading the group performance is 75 %, two raters (rater R01 and R11) showed slightly poorer performances when compared with the group mean (below 1 SD). Also, for listening R05 and R11 were below average (mean = 71.1 %). We considered the exclusion of these raters but also looked at their rating behaviour during the standard-setting rating task. Here, no abnormalities could be observed and thus, we decided to have all available raters included for the calculation of the final cut-scores.

![Discrepancies of Judgements](image1)

![Discrepancies of Judgements](image2)

*Figure 9: Results from the familiarisation activities (presented to the panel) for the receptive skills. Number of correct assignments per CEF-level is shown on the left the level-discrepancies on the right, respectively.*
4.5 Standard-Setting Rounds

The standard-setting started with reading and we used two rounds for both the IDM and the bookmark method. As previously outlined, no major changes of cut-score placements are usually found between round 2 and 3 (Karantonis & Sireci, 2006), thus, two rounds seemed adequate.
Round 1: Judges were asked to work through the OIB and to respond to each item (item keys were handed out after the first round). They should determine, for each item, the cognitive demands required to answer the item correctly (Ferrara et al., 2002). They then answered the following question: “What is the lowest level at which a test taker can answer the following item correctly?” We also instructed them to keep in mind that below A2 means not at level A2, this can either be A1 or even below. Furthermore, saying that an A2 person can answer the item does not mean an A1 person could not give the correct response, although it is unlikely. It just means that a correct response should not reasonably be required at lower levels. This implies that if an item is judged as A2, we implicitly say that a person at higher levels (e.g., B1) can also give a correct answer to this item.

The judges placed their judgements of competency levels in a web-based computer software. For every item we got ratings ranging from 1-3, for levels A1 or below, A2 and B1 or above, respectively. The ratings were evaluated and rating behaviour was presented to the judges. We showed simple bar graphs plotting the percentage agreement on each item with respect to CEF-levels. On basis of that feedback judges discussed homogeneous and heterogeneous rating-behaviour.

Round 2: Judges were asked to adjust their ratings from round 1, keeping in mind important discussion points and the feedback data. Data were analysed and the same type of feedback as in round 1 was given. The discussion more strongly focussed on the specific CEF-levels, and particulary on typical A1, A2, or B1 items.

Round 3: A detailed presentation and a short discussion introduced the bookmark method. The primary focus was on a common understanding of the concept of a borderline student or a minimally-competent person. The question asked was: “Can a borderline student – between level X and level Y – answer this item correctly in 2 out of 3 cases?” When they answered with no, the judges had to place a bookmark at that page in their OIB. Data were evaluated and a plot showing bookmark placements for the whole group was presented and discussed. Subsequently, also the impact data were presented showing the percentage distribution of students being at a certain level.

Round 4: In the final round, ratings and cut-score placements were being repeated. Similar feedback was presented as in the previous round. In a final discussion, judges should reach consensus about defensible cut-scores that are plausible for all experts.

4.6 Rater Analyses

Using the ratings from the first two rounds, we were able to calculate rater analyses (aside from those stemming from the familiarisation phase, see section 4.4).

Fleiss’ Kappa (Fleiss, 1971) was calculated for every round of the IDM for reading as well as for listening (i.e., round 1 and round 2, respectively). Fleiss’ Kappa, a variation of Cohen’s kappa coefficient (Cohen, 1960) for more than two raters, is a measure of agreement between raters, in which $\kappa = 1$ would indicate perfect agreement. Results can be found in Table 4. Interpreting the obtained values according to Landis and Koch (1977), there is a fair ($0.21 < \kappa < 0.40$) amount of agreement in both first rounds, and a moderate ($0.41 < \kappa < 0.60$) amount of agreement in the second rounds, which is satisfactory.

7 The person parameters were estimated with the Rasch-model from the pilot data, for calculating and plotting impact data, plausible values were used (see Embretson & Reise, 2000).
In addition, the *intraclass correlation coefficient* (ICC) can be used for assessing the agreement or consistency between raters (Bartko, 1966; McGraw & Wong, 1996). ICCs describes the ratio of variance that is caused by the different items to the total variance. In an ideal situation, where the variance in the ratings is only due to the different items and not due to the raters, the ICC would reach a value of 1. Measures for agreement (i.e., whether raters gave the same ratings to items) and consistency (i.e., whether raters brought items into the same order) can be found in Table 4.

### Table 4: Measures for rater analyses. ICC = Intraclass Correlation Coefficient.

<table>
<thead>
<tr>
<th></th>
<th>Fleiss' Kappa</th>
<th>Agreement</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 1</td>
<td>.34</td>
<td>.50</td>
<td>.55</td>
</tr>
<tr>
<td>Round 2</td>
<td>.51</td>
<td>.67</td>
<td>.69</td>
</tr>
<tr>
<td><strong>Listening</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 1</td>
<td>.33</td>
<td>.35</td>
<td>.39</td>
</tr>
<tr>
<td>Round 2</td>
<td>.42</td>
<td>.43</td>
<td>.47</td>
</tr>
</tbody>
</table>

4.7 Final Cut-Scores for the Receptive Skills

The latent ability values for each cut-score-item were averaged over judges to derive the group cut-scores. The results are shown in Table 5. The cut-scores for reading are generally higher than for listening. This corresponds to other national standards test, for example the test in Germany (Harsch et al., 2010) – participants familiar with the standards test in Germany reported during our workshop that this phenomenon is likely to occur and also well-known. In the final discussion, we asked for the general acceptability and defensibility of these cut-scores. The participants raised no critical points about the standard-setting process or the cut-scores, although it was mentioned that the first cut-score for listening was quite low. Other issues concerned weaknesses of the CEFR in general, a common opinion was that the scales represented by their can-do-statements vary in their amount of severity or leniency. This could also account for variability between the cut-scores for reading and listening.

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cut-Score</strong></td>
<td>SD</td>
<td><strong>Cut-Score</strong></td>
</tr>
<tr>
<td>A2 / B1 or above</td>
<td>575.16</td>
<td>10.43</td>
</tr>
<tr>
<td>below A2 / A2</td>
<td>406.21</td>
<td>18.10</td>
</tr>
</tbody>
</table>

*Table 5: Final bookmark cut-scores for the receptive skills.*
4.8 Evaluation

Similarly to writing and speaking, participants received several questionnaires during the process, here we only report the most important and notable results, more details can be found in the evaluation report (Zuber, in press).

Regarding necessary prerequisites before the actual rating-rounds, all participants were familiar with the CEFR, and most of them were confronted with it in their jobs. Again, most participants thought that the familiarisation phase was helpful for the method as well as for the process.

For reading as well as for listening, the instructions were not always clear. Specifically, some participants struggled with the concepts that were introduced in the bookmark method, the borderline student and the response probability of 2/3, which are needed to examine the items. This shows that for future workshops, the bookmark method needs to be explained in more detail, a few more test items could be administered in order to practice the rating process, or another, easier to understand method could be chosen.

Regarding the comments, there were a few remarks about items in the OIB, other participants commented favorably about the standard-setting process. While in general the feedback was positive, it is important to keep issues of the bookmark method in mind for future standard-settings.
5 Summary and Conclusion

For the productive skills cut-scores were determined using the contrasting-groups method, for the receptive skills we used a bookmark procedure. The contrasting-groups method was easy to implement and the rating task was clear. One drawback of this method was that data informing about rating behaviour is not quickly available. For future standard-settings analysis scripts should be written in advance of the workshops to keep participants informed about the ratings performed by the whole group. Another critical issue this method bears is the calculation of the final cut-scores that are derived from the ratings. Several statistical approaches can be chosen which can lead to diverse cut-scores, highly affecting the impact of the test results. Although this problem is inherent in most of the commonly used standard-settings methods, it needs to be discussed during the implementation of the method.

The bookmark method, used for the receptive skills, can easily be implemented and results are quickly available. In our standard-setting, the understanding of the minimally competent person was not easily achieved and it is uncertain in how far misunderstandings have led to diverse cut-score placements. This issue needs to be clarified for future standard-settings. The acceptability of the cut-scores for the receptive skills was generally high. A point raised by the participants concerning future item-development was the need for developing a higher amount of easier, A1-items.

When comparing our cut-scores with already available cut-scores from other countries (e.g., Harsch et al., 2010), the ranking of cut-scores across the sub-skills seems somehow comparable (although a scientific investigation is lacking at the moment). Listening shows lowest cut-scores, followed by reading, writing and speaking (at least for the first cut-score). As discussed with our consultants, this might correspond to general learning models of second language acquisition. Future work might shed light on this hypothetical consideration.
Linking the Austrian Standards-Based Test for English to the CEFR: The Standard-Setting Process

List of Figures

1. CEFR - Linking Framework.
2. The contrasting groups method. A) Judges are asked to rate student's work as either B1 or B2 performances. Two performance groups, plotted as frequency distributions, result from this rating process. To define a cut-score the midpoint of the two distributions can be calculated reflected by the mean of both group-means. B) The cut-score defined in A) lies at a testscore value of 24, for example. This cut-score leads to 19 false negatives (FNs) and 18 false positives (FP) and a total of 9.3% misclassifications.
3. Results from training activity 2 (presented to the panel) for writing. Number of correct assignments per CEFR-level (left) and level-discrepancies (right).
4. Results from training activity 2 (not presented to the panel) for writing. A) Number of correct assignments per judge (red line represents the group mean, grey lines the standard-deviation). B) Mean-Kappa (top) and mean-Kappa contrasted with standard-deviations (bottom) per judge.
5. ROC (Receiver Operating Characteristics) curves for one rater. The cut-scores are represented as the optimal coordinates reflecting maximal sensitivity and specificity. The AUC provides information about the rating quality.
6. Average cut-scores with standard errors for three different analysis approaches.
7. Results from training activity 2 (presented to the panel) for speaking. Number of correct assignments per CEFR-level (left) and level-discrepancies (right).
8. Results from training activity 2 (not presented to the panel) for speaking. Number of correct assignments per judge (red line represents the group mean, grey lines the standard-deviation). The inlet shows mean-Kappas contrasted with standard-deviations per judge.
9. Results from the familiarisation activities (presented to the panel) for the receptive skills. Number of correct assignments per CEFR-level is shown on the left the level-discrepancies on the right, respectively.
10. Results from familiarisation activities (not presented to the panel) for the receptive skills. Number of correct assignments per judge (red line represents the group mean, grey lines the standard-deviation). The inlets show mean-Kappa contrasted with standard-deviations.

List of Tables

1. Can-do-statements for writing used in the contrasting-groups method.
2. Can-do-statements for speaking used in the contrasting-groups method.
3. Item quality of the selected item-pool.
5. Final bookmark cut-scores for the receptive skills.
References


6 Annexes

1. Example: Pre-Workshop Activities for Writing

2. Example: Workshop Activities for Reading: Assignment of Levels to PLDs

3. Example: Online Rating for Writing
Pre-workshop activity 1: Working with CEFTrain

You can use CEFTrain in various ways. As pre-workshop activities we would recommend working with the following parts from CEFTrain:

1. Feel free to use this link to start clicking yourself through the maze of the CEFR scales. This should give you a good orientation of the way the scales are organized in the CEFR.
   http://www.kriterien.friesenrath/ceftool/index.php?l=0
   • Optional: Assist your own foreign language proficiency in a language other than English.
   http://www.kriterien.friesenrath/ceftool/index.php?l=0.1

2. Do the WRITING ball game. Do it until you get it right. Do it again three days after.
   • Optional: Do the other ball games.
   http://www.kriterien.friesenrath/ceftool/index.php?l=0.5

3. Do this rating exercises. Very quick and most certainly a success.
   http://www.kriterien.friesenrath/ceftool/index.php?l=0.6
   Do not expect too much, there is only one test.

4. This link gets you to a set of 11 writing tasks.
   http://www.kriterien.friesenrath/ceftool/index.php?l=0.7
   Rate the texts as suggested in the instructions and make a note of judgements you would want to disagree with. This should take you a little longer.

Pre-workshop activity 2: Scale Features

If you want to use the CEFR scales in practice, you must have internalised some of their inherent characteristics. To put it simply, you should recognize an A2 descriptor when you see one.

Have a good look at all six writing scales and collect typical formulations or proverbs that are strongly connected with a particular level, e.g. “short” or “simple” in A1, A2, and up to A2+.

Do this for the levels A1 to B2.

Look at the following scales:

1. Overall written production
   a. Creative writing
   b. Reports and analyses

2. Overall written interaction
   a. Correspondence
   b. Notes, messages and forms

For your convenience, the scales have been added here:

<table>
<thead>
<tr>
<th>OVERALL WRITTEN PRODUCTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
</tr>
<tr>
<td>A2</td>
</tr>
<tr>
<td>B1</td>
</tr>
<tr>
<td>B2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REPORTS &amp; ESSAYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
</tr>
<tr>
<td>A2</td>
</tr>
</tbody>
</table>
Pre-workshop activity 3:
Level Differentiation and Content Coherence in CEFR Levels

On the basis of what you know already and CEF, pp.33-36,

1. **Summarize the content features that differentiate between the levels:***
   - A1 and A2/A2+ and
   - A2/A2+ and B1;
   - (In what way are the higher levels different from the lower levels, e.g. A2/A2+ from A1?)
   - and

2. **on the other hand, establish coherence between them:***
   - (In what way are the higher levels connected with the lower levels, e.g. A2/A2+ and A1?)
### Reading

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Can understand specialised articles outside his/her field, provided he/she can use a dictionary occasionally to confirm his/her interpretation of terminology.</td>
</tr>
<tr>
<td>A2</td>
<td>Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus, reference lists and timetables.</td>
</tr>
<tr>
<td>B1</td>
<td>Can get an idea of the content of simple informational material and short simple descriptions, especially if there is visual support.</td>
</tr>
<tr>
<td>B2</td>
<td>Can identify the main concerns in clearly signalled argumentative texts.</td>
</tr>
<tr>
<td>C1</td>
<td>Can understand short, simple messages on postcards.</td>
</tr>
<tr>
<td>C2</td>
<td>Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary terms.</td>
</tr>
<tr>
<td>C3</td>
<td>Can locate specific information in lists and isolate the information required (e.g. use the &quot;Yellow Pages&quot; to find a service or tradesman).</td>
</tr>
<tr>
<td>C4</td>
<td>Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency items.</td>
</tr>
<tr>
<td>C5</td>
<td>Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency items.</td>
</tr>
<tr>
<td>C6</td>
<td>Can read straightforward instruction books on a subject related to his/her field and shared with a satisfactory level of comprehension.</td>
</tr>
<tr>
<td>C7</td>
<td>Can scan clearly through long and complex texts, locating relevant details.</td>
</tr>
<tr>
<td>C8</td>
<td>Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.</td>
</tr>
<tr>
<td>C9</td>
<td>Can understand complex instructions in his field, including details on conditions and warnings, provided he/she can reread difficult sections.</td>
</tr>
<tr>
<td>C10</td>
<td>Can receive familiar names, words and very basic phrases on simple topics in the most common everyday situations.</td>
</tr>
<tr>
<td>C11</td>
<td>Can understand lengthy, complex instructions in his field, including details on conditions and warnings, provided he/she can reread difficult sections.</td>
</tr>
<tr>
<td>C12</td>
<td>Can understand clearly written, straightforward instructions for a piece of equipment.</td>
</tr>
<tr>
<td>C13</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C14</td>
<td>Can read straightforward instruction books on a subject related to his/her field and shared with a satisfactory level of comprehension.</td>
</tr>
<tr>
<td>C15</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C16</td>
<td>Can identify specific information in simple materials in the context of normal conversation, such as letters, brochures and short official documents.</td>
</tr>
<tr>
<td>C17</td>
<td>Can understand long and complex texts on a topic of interest, including details on conditions and warnings, provided he/she can reread difficult sections.</td>
</tr>
<tr>
<td>C18</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C19</td>
<td>Can understand long and complex texts on a topic of interest, including details on conditions and warnings, provided he/she can reread difficult sections.</td>
</tr>
<tr>
<td>C20</td>
<td>Can obtain information, ideas and opinions from highly specialised sources within his/her field.</td>
</tr>
<tr>
<td>C21</td>
<td>Can understand longer texts in order to locate desired information, and gather information from different parts of a text, or from different texts in order to fulfil a specific task.</td>
</tr>
<tr>
<td>C22</td>
<td>Can identify specific information in simple materials in the context of normal conversation, such as letters, brochures and short official documents.</td>
</tr>
<tr>
<td>C23</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C24</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C25</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C26</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C27</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C28</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C29</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C30</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
<tr>
<td>C31</td>
<td>Can understand everyday signs and notices in public places, such as entrance announcements, telephone directions, workplace rules, such as directions, instructions, hazard warnings.</td>
</tr>
</tbody>
</table>

### Task 6

**Tom couldn’t believe his eyes. He had really won 100,000 euros.**

Tom was excited. He had never won so much money before. He was going to buy a new car. He was going to travel to Europe. He was going to do all the things he had been dreaming about. He was going to enjoy his money. He was going to travel to Europe. He was going to do all the things he had been dreaming about. He was going to enjoy his money. He was going to travel to Europe. He was going to do all the things he had been dreaming about.