BRITISH COUNCIL

EnglishScore

Linking EnglishScore to the Common European Framework of Reference for Languages (CEFR)

CRELLA (Centre of Research for English Language Learning and Assessment)

University of Bedfordshire

February 2022

Executive summary

Individuals and organisations around the world rely on EnglishScore to provide accurate and trusted assessments of English language proficiency across the A2 to C1 levels of the Common European Framework of Reference (CEFR).

This document reports on a project to link EnglishScore to the CEFR.

It should be read by score users, educators and others interested in the test.

To aid interpretation and to demonstrate compliance with global standards, it is based on:

- the Common European Framework of Reference for Languages (CEFR)[1] and draws on the related but more recent Companion Volume (CV)[2]
- the Council of Europe's Manual for Relating Language Examinations to the CEFR[3] (the CoE Manual).

According to the CoE Manual, defensible linking of any assessment to the CEFR must be grounded in a clear specification of the purpose and content of that assessment. The specification phase of the project to link EnglishScore to the CEFR is set out in a document titled 'EnglishScore: Test Purpose and Content' (Validity Report) available at www.englishscore.com.

Taken together, these documents facilitate comparisons with any other assessment that has been linked to the CEFR according to the Council of Europe's recommendations.

[1] Council of Europe, 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Strasbourg: Council of Europe. Available from: rm.coe.int/1680459f97 [Accessed 3 March 2022].

[2] Council of Europe, 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe. Available from: www.coe.int/lang-cefr [Accessed 3 March 2022].

[3] North, B., Figueras, N., Takala, S., Van Avermaet, P. and Verhelst, N., 2009. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.* Strasbourg: Council of Europe.

Table of contents

Introduction

This document reports on one aspect of the ongoing project to relate EnglishScore to the Common European Framework of Reference (CEFR).

It explains:

- the purpose of setting standards in relation to the CEFR
- the Bookmark method of standard setting
- the materials and procedures employed in advance of, and during the standard setting workshops
- the composition of the panel of experts involved
- the results, including the recommendations for CEFR cut scores and the evaluation of the workshops by the panellists.

Overview

EnglishScore

EnglishScore is an international assessment taken by young adult (16–17) and adult (18+) learners of English worldwide. Users may come from any language background and from any region of the world.

The CEFR

The CEFR is intended to ‘provide a common metalanguage for the language education profession in order to facilitate communication, networking, mobility and the recognition of courses taken and examinations passed’ (Council of Europe, 2020, p.26). It is used worldwide to help score users to interpret test results. It is therefore useful for test providers to offer guidance on how their tests relate to the framework. The CEFR has two key dimensions: a horizontal dimension of categories for describing different activities and aspects of competence and a vertical dimension representing progress in proficiency. The vertical dimension can be depicted as a progressive series of levels ranging from Pre-A1 to C2.

Linking tests to the CEFR

Relating performance on a test to the CEFR levels involves the accumulation of evidence in relation to both key dimensions of the framework. Following the recommendations of the Council of Europe Manual for Relating Language Examinations to the CEFR (Council of Europe, 2009), the purpose and content of the test should be described in relation to the horizontal dimension (activities and competences). The alignment of this aspect of EnglishScore to the CEFR (the specification stage of linking) is described in a document titled ‘EnglishScore: Test Purpose and Content’ (Validity Report) available at www.englishscore.com.

Standard setting

This document describes the complementary phase of linking referred to by the Council of Europe (2009) as *standard setting*. This involves relating test scores to the CEFR's vertical dimension: the reference levels. The document describes a series of standard-setting workshops designed to build understanding of how performance on EnglishScore relates to the test's target CEFR levels: A2 to C1. The workshops resulted in recommendations for cut scores for CEFR levels A2 to B2 on the Grammar and Vocabulary, and A2 to C1 for the Reading and Listening sections of the test.

The Centre for Research in English Language Learning and Assessment (CRELLA) conducted the online standard-setting workshops in January 2022 over a three-day period. Day one focused on the Grammar and Vocabulary section; day two on Reading; and day three on Listening. The Bookmark standard-setting method was applied to each section of the test.

Results

The standard-setting workshops were designed to elicit recommendations for the placement of cut scores representing four boundaries between CEFR levels for the purpose of reporting EnglishScore outcomes. The results can be summarised as follows:

	Score boundaries			
	A1/A2	**A2/B1**	**B1/B2**	**B2/C1**
Grammar and Vocabulary Section	8	16	26	
Reading Section	5	9	26	33
Listening Section	10	15	24	33

Method

Standard setting

Standard setting has been described as the process of specifying the level of performance on a test that is required for a test taker to be classified into a given performance category (such as grade *A*, *B* or *C*). It typically involves establishing one or more cut scores (scores that distinguish between the performance categories) (Cizek, 2012, p.4).

There are many standard-setting methods, but the Bookmark method (Lewis et al., 1999; Mitzel et al., 2001), used for this project, is 'perhaps the most popular method currently used to set performance standards on large-scale educational achievement tests' (Cizek, 2012, p.10). The advantages claimed for the method include that it is relatively straightforward for panellists compared to other methods and that, because it uses empirical data, it connects the process of setting cut score to a measurement scale and that it helps judges to relate the test content to descriptors (Mitzel et al., 2001). EnglishScore makes use of a Rasch measurement scale which provides an estimate of the probability of each individual giving a correct response to each test item. This scale is independent of the specific items appearing on each test form.

The Bookmark method involves three rounds of activity in which groups of panellists (in the case of EnglishScore, experts in English language education) work through a test booklet, called an ordered item booklet (OIB), in which a representative set of test items have been reordered from the easiest item (the item that test takers answered correctly most frequently) on page 1 to the most difficult item (the item that test takers answered correctly least frequently) on the final page. Before the online workshops reported in this document, in a process described by the Council of Europe as *familiarisation*, panellists were given information and a series of preparatory tasks connected to the CEFR common reference levels, the test, EnglishScore, and the standard-setting process. This included reading the CEFR (especially the overview of the common reference levels in Section 3.6); reviewing sample material from Listening and Reading tests provided by the Council of Europe to exemplify the common reference levels; evaluating sample performances at identified CEFR levels; taking EnglishScore; and completing a questionnaire that asked them to assign descriptors to levels.

When applying the Bookmark method, panellists usually begin by agreeing on a shared definition of a ‘minimally competent candidate’ (MCC). In this case, the MCC is the test taker that *only just* matches the description of a CEFR level as presented in the tables of scales of ‘can-do’ descriptors. Meeting online via Microsoft Teams in in four small groups and in plenary sessions, the panellists discussed these descriptors and their interpretation. This was done initially for the test as a whole (in relation to CEFR Section 3.6) and then separately on each workshop day for the relevant section of the test. For the Grammar and Vocabulary section on day one of the workshops, the panellists were asked to refer to and discuss the CEFR tables for General Linguistic Range, Grammatical Accuracy and Vocabulary Range. For the Reading section on day two, they considered Overall Reading Comprehension. On day three, they considered Overall Listening Comprehension in preparation for applying bookmarks to the Listening section.
The Bookmark method involves consideration of the MCC’s chances of success on each item reviewed. For each item, following wording recommended by the Council of Europe (2009), panellists were asked, ‘Is it likely that a minimally competent A2/B1/B2/C1 English language learner will answer this item correctly?’. Panellists began by considering the A2 level MCC. When they reached the point in the OIB where their answer for the A2 level MCC changed from ‘Yes’ to ‘No’ – the point at which items became too difficult for that borderline, minimally competent A2 learner – they placed a bookmark in the OIB to represent the cut score for the A1:A2 threshold. They then answered the question for the B1 level MCC and so on, until they had placed a bookmark representing the B1/B2 threshold (in the OIBs for the Grammar and Vocabulary section) or the B2/C1 threshold (in the OIBs for the Reading and Listening sections).

The highest threshold for Reading and Listening (B2/C1) is not relevant to the Grammar and Vocabulary section because this opening section of the test serves to determine which of three pathways for Reading and Listening are presented to the test taker (see Figure 1). Test takers who fail to reach the A1/A2 threshold on Grammar and Vocabulary do not progress to the Reading and Listening Sections and are scored as ‘below A2’. Those who exceed the A1/A2 threshold but do not attain the B1/B2 threshold are directed to the lowest of the three pathways for Reading and Listening, which includes material written to target the A2 and B1 levels of the CEFR. Those exceeding the B1/B2 threshold are directed to the highest-level pathway, consisting of B2 and C1 material. Those who exceed the A2/B1 but not the B2/C1 threshold are directed to the intermediate pathway made up of B1 and B2 level material. Only those who follow the highest-level pathway for Reading and Listening and also surpass the B2/C1 threshold on those sections qualify for the C1 level.

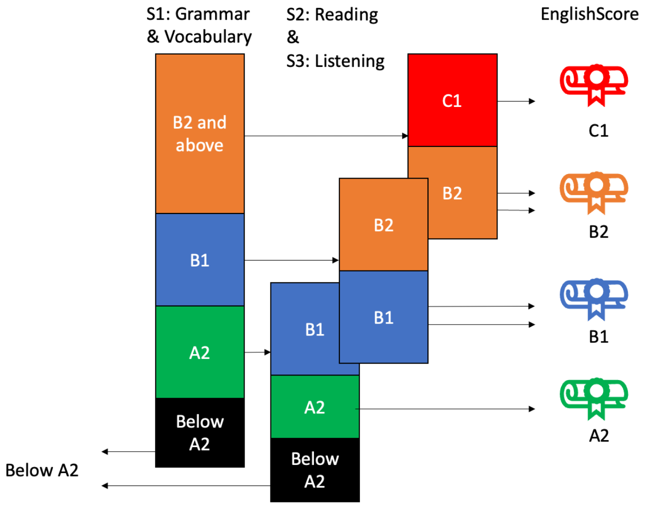

Figure 1. EnglishScore test outline.

In asking panellists whether an MCC is likely to make a correct response to an item, it is important to define what is meant by the word 'likely' in terms of the test taker's response probability (RP): the minimum probability with which the MCC should be able to answer an item correctly. In this case, the threshold used was 67%, as originally recommended by Mitzel et al. (2001). Although RP thresholds have sometimes been set at a 50% or 80% chance of success (Karantonis and Sireci, 2006), it has been argued that it is relatively straightforward for panellists to conceptualise a two-thirds chance of success as representing a consistent level of mastery (Tiffin-Richards et al., 2013). Panellists' bookmarks were therefore placed at the point where they considered that the MCC at each level would have less than a two-thirds or 67% chance of responding correctly to an item, or where fewer than two-thirds of MCCs just at the relevant level would be expected to give a correct response.

Further familiarisation activities were provided during the online standard-setting workshops, including a review of responses to the familiarisation questionnaire and further discussion of the CEFR levels. These were followed by an explanation and discussion of the purpose of the Bookmark method and of the procedures to be followed. When the panellists indicated that they understood the procedures and were ready to begin, they were divided into three groups of three and one pair. The workshop leader moved between the groups to monitor their progress and to answer any questions that arose. The small group format offers greater opportunity for panellists to voice their rationales and promotes discussion. It also encourages independence in bookmark placement as each group of experts arrives at its own bookmark placements, which are then aggregated across groups.

A similar procedure was followed for each section of the test on successive days. In the first round, the panellists reviewed a first OIB before placing their first set of bookmarks. A plenary discussion was followed by a second round of group deliberation, with a second OIB (made up of previously unseen items) and bookmark placements. After plenary discussion of this second round, there was a third round of group deliberation, again with a new OIB, before panellists worked independently to place their final bookmarks.

Feedback information was provided to the panellists between rounds. After round 1, they were told how the placements of their bookmarks compared with those of their counterparts in other groups. After round 2, they were shown how their proposed cut scores would affect results for the test-taking population: the proportion of test takers that would be classified in each level. Presenting panellists with an OIB made up of previously unseen items for each round helped to ensure that they were exposed to numerous items representing a sample of the material that test takers might encounter.

Ordered item booklets

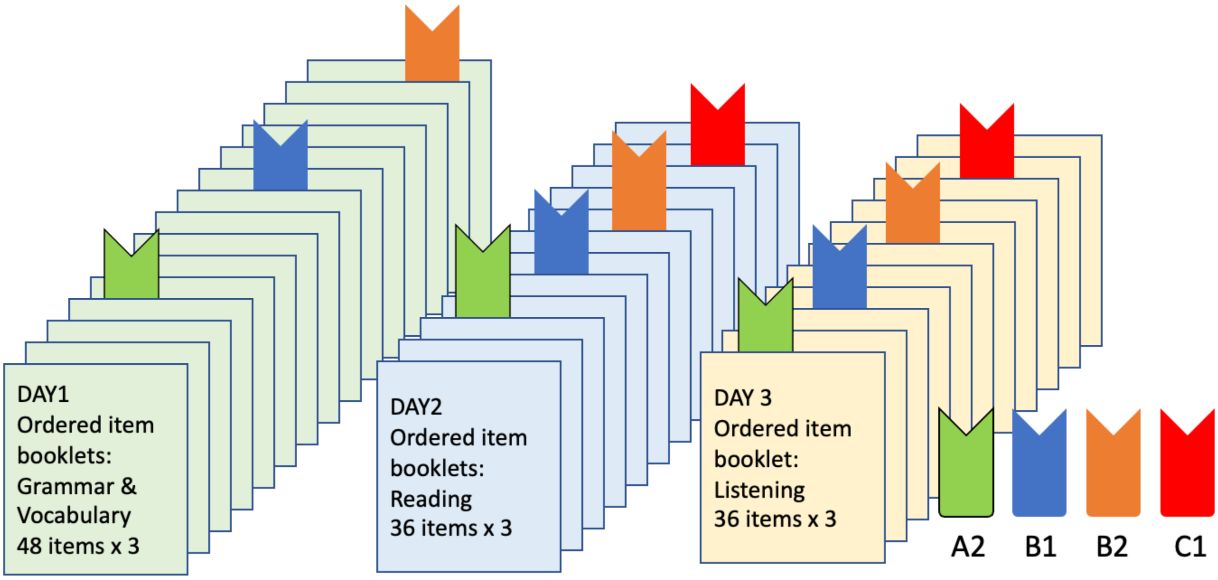
DAY1
Ordered item booklets: Grammar & Vocabulary 48 items x 3
DAY2
Ordered item booklets: Reading 36 items x 3
DAY 3
Ordered item booklet: Listening 36 items x 3
A2
B1
B2
C1

Figure 2. EnglishScore Bookmark standard setting.

OIBs were prepared for the Grammar and Vocabulary, Reading and Listening sections of the test. Each booklet represented an assessment that had been administered to test takers, but with items reordered according to their difficulty on the EnglishScore Rasch measurement scale. For the Grammar and Vocabulary section, the booklets included 48 items; for Reading and Listening, 36 items (see Figure 2). On the Grammar and Vocabulary section, all test takers receive the same 48 items, but for Reading and Listening, there are three overlapping pathways, and each test taker responds to 12 Reading and 12 Listening items. For each skill, pathway 1 has six items that target the A2 level and six that target B1. Pathway 2 has six B1 and six B2 items; pathway 3, six B2 and six C1 items. Each OIB contained a full set of 36 items from these overlapping pathways (6 targeting A2, 12 B1, 12 B2, 6 C1) for the relevant skill, ordered according to their scaled Rasch difficulty.

Each item appeared on one page of the booklet accompanied by the following statistics: the Rasch scaled difficulty, to indicate how difficult each item was in relation to those that preceded or followed it; the mean score (the proportion of test takers encountering the item who gave a correct response); infit and outfit mean square fit statistics (indicating how closely patterns of responses to each item matched the predictions of the Rasch model); and point biserial correlations between the item scores and Rasch measures (higher figures indicate greater discrimination between higher and lower ability test takers).

Panel

Any standard-setting study depends on the panel of experts that deliberate on the most appropriate cut score. The panel should be expert in the test, the population of test takers and the performance standards: the common reference levels of the CEFR (Council of Europe, 2009). In this case, the panel, which included the same participants over the three workshop days, was made up of 11 individuals. Of these, four (hereafter referred to as the item writers) worked as item writers for EnglishScore and were therefore familiar with the test. Four (the educators) were educators who worked with learners at teaching centres in China, Qatar, Thailand and Vietnam: representing countries with large numbers of EnglishScore test takers. The remaining three (the language testers) were language testing specialists with knowledge and experience of using the CEFR, but with limited knowledge of EnglishScore.

Although chosen for their specific expertise, the three groups also had skills and experiences in common. All panellists were qualified teachers of English with four having more than twenty years, four eleven to twenty years, two six to ten years and one less than five years of experience. They reported a wide range of qualifications. Five held a Masters degree in English Language Teaching or allied discipline, including one with a PhD in this area and one with a Postgraduate Certificate in Education (PGCE). Seven reported a Cambridge Certificate in English Language Teaching for Adults (CELTA) and one a Trinity College London Certificate in TESOL. Three held a Diploma (Cambridge DELTA, or Trinity College London DipTESOL), with a fourth studying for the DELTA.

Four panellists (all the language testers and one educator) had previously participated in CEFR linking panels and all described themselves on a five-point scale ranging from ‘No knowledge’ to ‘Extensive knowledge’ as having at least a ‘Moderate knowledge’ of the CEFR: none described themselves as having ‘No knowledge’ or ‘Limited knowledge’ of the framework. Those claiming a moderate level of knowledge reported using the framework for purposes such as syllabus design or grading learners. Five claimed a ‘Substantial’ level of knowledge, using the framework for purposes such as test design, standard setting and rating. One in this group (a language tester) reported 18 years’ experience in using the CEFR for a range of purposes. Only one panellist (an item writer) reported ‘Extensive knowledge’ and had been involved in training others to apply the CEFR in evaluating test items.

Results

Familiarisation questionnaire

A questionnaire shared with the panellists before the workshops provides an indication of how familiar they were with the CEFR common reference levels. The panellists were asked to judge the CEFR levels (from Pre-A1 to C2) of 32 descriptors: 12 drawn from the CEFR General Linguistic Range, Grammatical Accuracy or Vocabulary Range scales and 10 each from the CEFR scales for Listening comprehension and Reading comprehension. Of the 352 judgements made by panellists, 224 (64%) were accurate and a further 112 (32%) were within one CEFR level. Of the 128 incorrect judgements, 70 were below and 58 above the CEFR level of the descriptor, suggesting that as a group, the panellists were consistently neither harsh nor lenient in their interpretation of the levels.

In all, 3 of the 32 descriptors were misclassified by a majority of panellists. These included ‘Can employ very simple principles of word order in short statements’, a Pre-A1 statement placed at A1 by six panellists; ‘Can understand the main points and important details in stories and other narratives (e.g. a description of a holiday), provided the delivery is slow and clear’, a B1 descriptor placed at A2 by six panellists and at A1 by one. ‘Can make appropriate inferences when links or implications are not made explicit’ was judged to be C1 by seven panellists and B1 by one, although it is a C2 descriptor. Of the three groups of panellists, the language testers were the most accurate, correctly identifying an average of 26.7 (83%) of the 32 descriptors. The item writers averaged 20.5 (64%) and the educators 14.25 (45%). Correlations between panellist judgements and CEFR levels, based on conversion of CEFR levels to numeric values from 1 (Pre-A1) to 7 (C2), and the percentage of descriptors accurately placed by each panellist are displayed in Table 1.

Table 1. Panellist judgements of descriptor level; correlations and per cent correct identification of CEFR levels.

Educators			*Item writers*			*Applied linguists*		
	r	% correct		*r*	% correct		*r*	% correct
P1	0.853	37.5%	P5	0.885	59.4%	P9	0.928	87.5%
P2	0.876	34.4%	P6	0.965	84.4%	P10	0.939	71.9%
P3	0.894	46.9%	P7	0.951	71.9%	P11	0.979	90.6%
P4	0.956	59.4%	P8	0.864	40.6%			
Average	0.895	44.53%	Average	0.916	64.06%	Average	0.949	83.33%

Although the Council of Europe (2009) does not provide guidance on how accurately judges should be able to classify CEFR descriptors when they participate in a CEFR standard-setting workshop, the results summarised in Table 1 suggest that each panellist had a good understanding of the common reference levels with correlations between their ratings for the descriptors and the official CEFR levels ranging from 0.853 to 0.979 (see Table 1).

A further indication of the extent to which the group shared a coherent understanding of the common reference levels can be obtained through an intraclass correlation coefficient. This offers a measure of the degree to which judges are consistent with each other. It is a number that can range between 0 and 1, with higher figures indicating greater consistency. In this case, the intraclass correlation coefficient for the 11 panellists of 0.87 suggested that there was a high level of agreement among them as a group in their interpretation of the CEFR scales.

Bookmark procedures

Following a review of the familiarisation activities and further discussion of the levels at the beginning of the first workshop, the panellists agreed verbally that they were sufficiently familiar with the CEFR and consistent in their judgements to proceed.

Throughout the judgement process, panellists had access to the OIBs and to the CEFR scales that had been discussed during the familiarisation activities.

The results of the three rounds of judgements for each Section of the test are set out in Tables 4 to 6 in the Appendix. The tables display the panellists' individual recommendations for cut scores and summary statistics for each round.

The median of the recommendations from the third round are usually taken as the recommended cut scores for the test (Mitzel et al., 2001). These cut score recommendations can be viewed from two perspectives. First, they can be understood simply in terms of the number of correct responses required for a test taker to be classified as A2, B1, B2 or C1 on the test form reviewed. However, this interpretation is limited to the specific set of test material seen in each round. A second perspective is provided by the Rasch measurement scale, which takes account of variation in the difficulty of obtaining a given score on each test form. For example, a score of 20 on Grammar and Vocabulary was associated with a scaled difficulty of 457 in the OIB in round 1, 441 in round 2 and 476 in round 3. To provide the most appropriate indication of the number of items a test taker needs to answer correctly to be classified at each level, the final panel recommendations are therefore converted to a scale based on the average difficulty (on the Rasch measurement scale) across forms of obtaining each score point on the test.

In relation to observed scores, the panellists' cut score recommendations were generally consistent across the three rounds, except in the case of the Grammar and Vocabulary section where, following feedback on the proportion of test takers that would progress to follow each pathway, the third round of judgement involved reductions in the cut scores for A2/B1 and B1/B2 of 8 points and 15.5 points, respectively. While these differences were substantial in relation to the number of items, they were much less dramatic in relation to the Rasch measurement scale. At these thresholds, to reach the same point on the Rasch measurement scale, test takers would have needed a higher score in rounds 1 and 2 than in round 3. The A2/B1 cut score for round 3 of 438 was only 3 points lower than the recommendation for round 2 and 25 points lower than that for round 1. These figures should be interpreted in relation to the average length of the scale across the 3 sets of material of 425 points and the conditional standard error of measurement (CSEM: an estimate of the error involved in measuring test performance) of the test of 10 points at these levels. The differences between rounds at the B1/B2 threshold (104 points between rounds 1 and 3; 83 points between rounds 2 and 3) were greater. As these differences are several times larger than the CSEM of the test, they would seem to reflect a meaningful change in how the panel interpreted this threshold between rounds 2 and 3.

Viewed both from the perspective of observed scores and from the perspective of the Rasch measurement scales, the cut scores for Reading and Listening remained relatively stable across judgement rounds. For Reading, the scaled cut scores fell between round 1 and round 3 by 23 points at A1/A2, increased by 28 points at A2/B1, by 6 points at B1/B2 and by 1 point at B2/C1. For Listening, the changes between rounds were larger. The greatest occurred at the A1/A2 threshold, which increased by 85 points between round 1 and round 3. The higher-level thresholds all also increased: by 50 points at A2/B1, by 19 at B1/B2 and by 2 at B2/C1. Again, these figures should be evaluated in relation to an average scale length of 425 points across assessment forms and CSEM of between 9 and 13 points. For both the Reading and the Listening sections, the standard-setting process tended to lead to more challenging thresholds in later rounds, more notably for Listening than for Reading and at the A1/A2 and A2/B1 levels, with the higher-level thresholds (B1/B2 and B2/C1) remaining relatively unchanged on both sections across rounds.

Combining the information from the recommendations made by the panellists based on numbers of correct responses and from the conversion of these to points on the Rasch measurement scale suggests the cut score recommendations shown in Table 3. The scaled (Rasch) values and conditional standard errors of the test are also displayed.

Table 2. Cut score recommendations based on standard-setting workshops.

	A1/A2	Meas.	CSEM	**A2/B1**	Meas.	CSEM	**B1/B2**	Meas.	CSEM	**B2/C1**	Meas.	CSEM
Gr. & Vo.	8	404	10	16	454	10	26	490	9			
Read.	5	424	10	9	514	9	26	610	11	33	682	13
List.	10	448	10	15	496	9	24	552	10	33	640	11

Validation

Following the online workshops, the participants were invited to complete an online survey, based on a model provided by the Council of Europe (2009) asking for their views on the workshops and the success of the standard-setting exercise (see Table 2). This serves to support the validity of the workshops because it allows participants to express how satisfied they have been with the procedures followed and decisions arrived at.

Table 3. Responses to the post-workshop survey.

	1 Strongly disagree	*2 Disagree*	*3 Agree*	*4 Strongly agree*
I understood the purpose of the seminar.			2	5
I understood how to answer the pre-seminar online questionnaire.			1	6
The resources provided in preparation were helpful.		1		6
The training provided helped me to understand the judgement process.			6	1
I understood the instructions for the activities.			5	2
I felt I had sufficient understanding of the CEFR.			2	5
There was adequate time for reflection and discussion before making the judgements.			4	3
I was able to make my viewpoints known when we were in breakout groups.			1	6
Use of the videoconferencing facility was effective.			3	4
I am confident in the decisions I have made.			5	2

Of the seven participants who completed the survey, all but one expressed agreement with all ten statements (Table 3). This indicates that the panellists were generally satisfied with the workshops and confident in the standards they set. The only negative evaluation concerned preparation resources, although the other six panellists responding to this question strongly agreed that these had been helpful.

The difficulty for this one panellist seems to have stemmed from the concept of the MCC. This has been a recurrent issue affecting several standard-setting approaches, particularly the challenge of distinguishing between the minimally competent and the typical candidate at a level (Skorupski, 2012). Other panellists also referred to having some difficulty in defining the MCC, one commenting that it was 'conceptually slippery' with another expressing concern that the groups might have interpreted the term differently. A third panellist commented that a longer initial discussion of the concept might have been helpful. The provision of additional preparatory activities to introduce the concept and more time for discussion of the MCC on day one might both be helpful for future workshops. In practice, although there was extensive discussion of the MCC over the three days, only the highest cut score on the Grammar and Vocabulary section changed very markedly between consecutive rounds, suggesting that the question had been largely resolved between the second and third judgement rounds for this section.

A further potential source of evidence for the cut scores arrived at is convergence between different sources. In the case of EnglishScore, all material is written to target a specific CEFR level. In effect, this represents a content-based claim for the performance standards embodied by the test. Making the same assumption that mastery is represented by a two-thirds chance of success on items at the target level, cut scores for each CEFR level would be A1/A2: 5, A2/B1: 18 and B1/B2: 34 for Grammar and Vocabulary; and for both Reading and Listening: A1/A2: 4, A2/B1: 10, B1/B2: 22 and B2/C1: 34. These latter figures are close to the score recommendations arrived at by the panels of 5, 9, 26 and 33 for Reading and 10, 15, 24 and 33 for Listening. To this extent, the evidence from the test design and the panel workshops was mutually corroborating.

In terms of numbers of correct responses, the conclusions for Grammar and Vocabulary appeared less consistent with the test design. The suggested cut score of 5 points for A1/A2 matched the test designers' intentions, but the cut score of 12 as the threshold for B1 was below the 18 suggested by the designers and 22 for B2 was out of line with the designers' 34. However, on the Rasch measurement scale, the differences were less marked. When seen in relation to the average difficulty (on the Rasch scale) of observed score points across test forms, the A1/A2 cut score was 23 points higher, the A2/B1 6 points lower and the B1/B2 48 points higher on the measurement scale than planned for in the test design.

The standard-setting workshops represent an important step forward in understanding the relationship between EnglishScore and the CEFR, providing a sound basis for revisions to the EnglishScore CEFR cut scores. The recommendations for cut scores will be implemented and their impact will be monitored. As the test continues to develop, we will follow the advice of the Council of Europe (2009) and keep the relationship to the CEFR under constant review.

References

Cizek, G.J., ed., 2012. *Setting Performance Standards: Foundations, Methods, and Innovations.* Abingdon: Routledge.

Council of Europe, 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Strasbourg: Council of Europe. Available from: rm.coe.int/1680459f97 [Accessed 3 March 2022].

Council of Europe, 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Strasbourg: Council of Europe. Available from: www.coe.int/lang-cefr [Accessed 3 March 2022].

Karantonis, A. and Sireci, S.G., 2006. The Bookmark Standard-Setting Method: A Literature Review. *Educational Measurement: Issues and Practice*, *25*(1), pp.4–12.

Lewis, D.M., Mitzel, H.C., Green, D.R. and Patz, R.J., 1999. *The Bookmark Standard Setting Procedure.* Monterey, CA: McGraw-Hill.

Mitzel, H.C., Lewis, D.M., Patz, R.J. and Green, D.R., 2001. The bookmark procedure: Psychological perspectives. In: G.J. Cizek, ed., *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, pp.249–281.

North, B., Figueras, N., Takala, S., Van Avermaet, P. and Verhelst, N., 2009. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.* Strasbourg: Council of Europe.

Skorupski, W.P., 2012. Understanding the Cognitive Processes of Standard Setting Panelists. In: G.J. Cizek, ed. *Setting Performance Standards: Foundations, Methods, and Innovations*. Abingdon: Routledge, pp.135–148.

Tiffin-Richards, S.P., Anand Pant, H. and Köller, O., 2013. Setting Standards for English Foreign Language Assessment: Methodology, Validation, and a Degree of Arbitrariness. *Educational Measurement: Issues and Practice*, *32*(2), pp.15–25.

Appendix

Table 4. EnglishScore Grammar and Vocabulary section: panel recommendations for cut scores.

	ROUND 1						***ROUND 2***						***ROUND 3***					
Panellists	*A1/A2*	*Rasch*	*A2/B1*	*Rasch*	*B1/B2*	*Rasch*	*A1/A2*	*Rasch*	*A2/B1*	*Rasch*	*B1/B2*	*Rasch*	*A1/A2*	*Rasch*	*A2/B1*	*Rasch*	*B1/B2*	*Rasch*
1	7	369	30	527	43	594	10	403	24	473	36	548	4	384	10	430	20	476
2	6	360	18	438	33	540	4	360	14	425	42	573	6	413	15	454	20	476
3	6	360	15	428	44	611	8	392	19	438	42	573	5	404	12	438	22	490
4	6	360	27	492	43	594	8	392	21	460	36	548	8	428	21	480	36	497
5	6	360	16	431	46	624	8	392	18	437	42	573	8	428	18	464	42	590
6	6	360	17	436	42	588	6	380	17	436	42	573	6	413	17	458	42	590
7	6	360	30	527	45	613	10	403	24	473	37	551	4	384	10	430	20	476
8	5	357	16	435	23	465	5	371	16	431	23	472	5	404	16	454	23	497
9	13	410	27	492	34	541	8	392	21	460	36	548	6	413	12	438	23	497
10	6	360	16	431	46	624	8	392	18	437	42	573	4	384	12	438	21	480
11	10	384	28	499	43	594	10	403	25	474	42	573	5	404	10	430	20	476
Mean	7.00	367.27	21.82	466.91	40.18	580.73	7.73	389.09	19.73	449.47	38.18	555.00	5.55	405.23	13.91	446.73	26.27	504.09
Median	6	360	18	438	43	594	8	392	19	438	42	573	**5**	**404**	**12**	**438**	**22**	**490**
SD	2.37	16.08	6.42	40.51	7.19	47.88	2.00	13.71	3.58	18.72	5.78	29.99	1.44	14.62	3.73	16.45	9.02	43.41

Table 5. EnglishScore Reading section: panel recommendations for cut scores.

	ROUND 1								***ROUND 2***								***ROUND 3***							
Panellists	*A1/A2*	*Rasch*	*A2/B1*	*Rasch*	*B1/B2*	*Rasch*	*B2/C1*	*Rasch*	*A1/A2*	*Rasch*	*A2/B1*	*Rasch*	*B1/B2*	*Rasch*	*B2/C1*	*Rasch*	*A1/A2*	*Rasch*	*A2/B1*	*Rasch*	*B1/B2*	*Rasch*	*B2/C1*	*Rasch*
1	7	465	8	470	29	622	36	738	6	500	9	539	29	651	35	731	1	217	8	480	17	562	30	650
2	6	440	8	470	29	622	36	738	3	403	6	500	28	645	34	707	4	361	8	480	17	562	30	650
3	6	440	8	470	29	622	36	738	3	403	6	500	22	604	34	707	4	361	8	480	28	631	34	701
4	5	433	8	470	21	579	31	635	5	447	11	553	18	592	29	651	4	361	9	514	24	610	29	645
5	1	231	8	470	26	602	36	738	1	304	8	525	26	624	36	736	6	424	20	576	30	650	34	701
6	6	440	9	486	22	582	34	681	3	403	6	500	21	597	27	627	8	480	20	576	28	631	33	682
7	5	433	15	544	29	622	34	681	6	500	15	579	26	624	34	707	6	424	9	514	34	701	36	790
8	6	440	11	521	18	550	33	671	6	500	9	539	29	651	35	731	6	424	14	540	24	610	32	666
9	7	465	21	579	27	604	34	681	5	447	11	553	19	595	30	661	4	361	11	520	25	610	36	790
10	7	465	9	486	16	546	33	671	6	500	9	539	29	651	35	731	6	424	9	514	29	645	35	783
11	6	440	16	546	33	671	36	738	6	500	11	553	18	592	33	691	6	424	11	520	18	568	33	682
Mean	5.64	426.55	11.00	501.14	25.36	602.04	34.45	700.84	4.55	446.02	9.18	534.58	24.09	620.74	32.91	698.27	5.00	387.65	11.55	519.62	24.91	616.39	32.91	703.52
Median	6	440	9	486	27	604	34	681	5	447	9	539	26	624	34	707	6	424	9	514	25	610	33	682
St.Dev.	1.69	66.08	4.40	39.47	5.35	36.19	1.69	37.75	1.75	63.37	2.75	26.05	4.57	25.33	2.91	36.93	1.84	68.73	4.55	33.95	5.65	42.33	2.43	57.31

Table 6. EnglishScore Listening section: panel recommendations for cut scores.

	ROUND 1								ROUND 2								ROUND 3							
Panellists	A1/A2	Rasch	A2/B1	Rasch	B1/B2	Rasch	B2/C1	Rasch	A1/A2	Rasch	A2/B1	Rasch	B1/B2	Rasch	B2/C1	Rasch	A1/A2	Rasch	A2/B1	Rasch	B1/B2	Rasch	B2/C1	Rasch
1	1	306	8	414	17	526	32	633	1	338	8	406	18	494	34	666	1	373	7	452	15	518	36	700
2	1	306	10	446	15	490	33	638	5	381	9	410	21	520	34	666	2	415	7	452	21	547	34	640
3	3	363	8	414	18	532	33	638	5	381	11	440	21	520	34	666	4	434	14	496	26	591	34	640
4	4	372	10	446	15	490	25	573	6	390	11	440	18	494	28	565	4	434	11	484	16	518	29	618
5	3	363	8	414	18	532	33	638	5	381	11	440	21	520	34	666	6	448	10	482	21	547	33	636
6	1	306	10	446	15	490	35	659	5	381	9	410	21	520	34	666	7	452	21	547	29	618	36	700
7	1	306	8	414	17	526	32	633	1	338	8	406	18	494	34	666	10	482	17	520	28	614	36	700
8	6	397	14	471	20	540	31	628	6	390	14	474	20	518	31	618	6	448	14	496	20	544	31	632
9	5	376	11	450	18	532	25	573	6	390	11	440	18	494	28	565	6	448	13	490	23	552	34	640
10	3	363	8	414	18	532	33	638	5	381	9	410	21	520	34	666	1	373	18	533	29	618	36	700
11	8	414	17	526	32	633	36	670	8	406	17	489	32	619	36	680	8	462	17	520	32	635	36	700
Mean	3.27	351.99	10.18	441.24	18.45	529.58	31.64	629.18	4.82	377.86	10.73	433.36	20.82	519.20	32.82	644.88	5.00	433.63	13.55	497.48	23.64	572.86	34.09	663.86
Median	3	363	10	446	18	532	33	638	5	381	11	440	21	520	34	666	6	448	14	496	23	552	34	640
St. Dev.	2.33	39.52	2.93	34.72	4.76	39.36	3.56	30.33	2.09	21.16	2.72	28.49	3.97	35.57	2.64	42.34	2.90	34.40	4.52	30.41	5.59	43.15	2.34	34.85

Contact Information

About the British Council
The British Council builds connections, understanding and trust between people in the UK and other countries through arts and culture, education and the English language.

We work in two ways – directly with individuals to transform their lives, and with governments and partners to make a bigger difference for the longer term, creating benefit for millions of people all over the world.

We help young people to gain the skills, confidence and connections they are looking for to realise their potential and to participate in strong and inclusive communities. We support them to learn English, to get a high-quality education and to gain internationally recognised qualifications. Our work in arts and culture stimulates creative expression and exchange and nurtures creative enterprise.

We connect the best of the UK with the world and the best of the world with the UK. These connections lead to an understanding of each other's strengths and of the challenges and values that we share. This builds trust between people in the UK and other nations which endures even when official relations may be strained.

We work on the ground in more than 100 countries. In 2019–20, we connected with 80 million people directly and with 791 million overall, including online and through our broadcasts and publications.

About EnglishScore
EnglishScore is a global test and certificate of English for employment and education from the British Council and has more than 2 million new users per year across 150 countries.

Designed to help millions of people to unlock the potential that the English language gives them, the EnglishScore mobile test can be taken from anywhere, at any time, with results available immediately.

The free Core Skills test assesses proficiency in grammar, vocabulary, reading and listening and can take up to 40 minutes to complete. A speaking test is also available to assess pronunciation, fluency and communication skills.

Test results are reported using the Common European Framework of Reference for Languages, the global standard used by many other international tests, such as TOEFL ITP, TOEIC and IELTS. Test-takers also have the opportunity to purchase a certificate to prove their level to employers and organisations.

Contact EnglishScore
For questions about the test, including content development, test scoring, security or certification, please contact:

EnglishScore

Scale Space
58 Wood Lane
London W12 7RZ
United Kingdom
contact@englishscore.com