CELPIP

CELPIP Research Series

Mapping the CELPIP Test Scores Onto the Common European Framework of Reference

Michelle Chen, Prometric

You-Min Lin, Prometric

How to cite this report

Chen, M. Y., & Lin, Y. (2025). *Mapping the CELPIP test scores onto the Common European Framework of Reference*. Prometric.

Available at: celpip.ca/cefr-report/

Abstract

This report documents a series of standard-setting studies to map CELPIP General test scores onto the Common European Framework of Reference (CEFR). The CELPIP test measures test takers' English language proficiency in four areas: reading, listening, speaking, and writing. This study focused on recommending the minimum scores needed to enter the B1, B2, C1, and C2 levels of the CEFR. The modified Angoff standard-setting approach was implemented for the Listening and Reading components of the CELPIP test, while the Bookmark method was utilized for the Writing and Speaking components. Fifteen English-language educators served on the standard-setting panels. The results of these studies provide score users with panel-recommended score equivalence between CELPIP levels and the CEFR levels.

Keywords: CEFR, CELPIP, standard-setting, score mapping, score equivalence

Table of Contents

1. Study Background and Purposes

The purpose of this study was to map CELPIP test scores onto the Common European Framework of Reference (CEFR). Standard-setting with expert panels was adopted as the methodology for this purpose. The outcomes of a standard-setting study are the minimum test scores needed to achieve defined performance levels. In this case, the performance levels are specified by the CEFR descriptors.

The CEFR is an internationally recognized standard for describing language proficiency. It provides a common basis for language learning, teaching, assessment, and performance benchmarking in educational and professional contexts. The framework is adopted for curriculum development, standardized testing, and comparison of language qualifications across countries. The CEFR describes language proficiency in three bands and six levels: Basic user (A1 and A2), Independent user (B1 and B2), Proficient user (C1 and C2) (Council of Europe, 2001).

CELPIP is a comprehensive English language testing program designed to measure the communicative language proficiency required for successful communication in general social, educational, and workplace contexts. Examples of such contexts include routine exchanges with service representatives or others in a public setting, day-to-day encounters with teachers or classmates, and common interactions with colleagues at work. The test tasks are designed to reflect the range of communicative situations that individuals may encounter in their daily lives.

The test is fully computer-delivered and consists of four components: Reading, Listening, Speaking, Writing. Test takers take the CELPIP test at one of Prometric's designated test centres at computer stations.

Reading: This component evaluates the ability to comprehend written texts from everyday and academic contexts. Passages include correspondence, expository articles, argumentative editorials, and informational texts, followed by questions assessing understanding of main ideas, details, and inferences.

Listening: This component assesses the ability to understand spoken English in real-life contexts. Test takers listen to conversations, discussions, and short talks, and answer related comprehension

questions targeting main ideas, supporting details, and inferred meanings.

Speaking: This component includes eight tasks covering a broad range of communicative scenarios. Test takers are asked to describe personal experiences, offer advice, explain ideas, make predictions, persuade, and justify opinions. Responses are evaluated using a rubric that considers coherence and meaning, lexical range, listenability, and task fulfilment.

Writing: This component includes two tasks: composing an email and responding to a survey question. Responses are evaluated according to a rubric assessing coherence and meaning, lexical range, readability, and task fulfilment.

The CELPIP test evaluates a wide range of proficiency in reading, listening, speaking, and writing. Scores for each component, as well as an overall score (calculated as the average across the four components), are reported on a 13-band scale (0–12). Table 1 presents the CELPIP levels and their corresponding descriptions.

Table 1. CELPIP Levels and Descriptors

CELPIP Level	CELPIP Descriptor
0	Insufficient information to assess
1	Insufficient information to assess
2	Limited ability in contexts related to immediate needs
3	Some proficiency in limited contexts
4	Adequate proficiency for daily life activities
5	Acquiring proficiency in everyday social, educational, or workplace contexts
6	Developing proficiency everyday social, educational, or workplace contexts
7	Adequate proficiency in somewhat demanding social, educational, or workplace contexts
8	Good proficiency in more demanding social, educational, or workplace contexts
9	Effective proficiency in some high-stakes social, educational, or workplace contexts
10	Highly effective proficiency in some high-stakes social, educational, or workplace contexts

11	Advanced proficiency in high-stakes social, educational, or workplace contexts
12	Expert proficiency in high-stakes social, educational, or workplace contexts

The purpose of this study is to identify, for each CELPIP test component, the minimum scores (cut scores) required to reach the B1, B2, C1, and C2 levels of the CEFR. The CEFR performance levels provide an internationally recognized framework for interpreting test results and support a range of institutional decisions. The standard-setting process is designed to produce cut scores that are defensible, evidence-based, and aligned with the language skills expected at each CEFR level. Linking CELPIP scores to CEFR levels supports meaningful interpretations of CELPIP test results and helps institutions and other stakeholders make informed decisions regarding instruction, placement, admissions, and other educational or professional applications.

The following sections describe the methods used to generate these cut scores, including the composition and training of the standard-setting panels, the procedures applied for each test component, and the approaches used to summarize and validate the results. The report then presents the outcomes for each CELPIP component, the resulting CELPIP–CEFR mappings, and panelists' reflections on the process. Finally, implications for score interpretation and use are discussed.

2. Method

The main task for panelists in these standard-setting studies was to recommend, for each of the four test components, the minimum scores needed to reach the targeted CEFR levels (B1, B2, C1, and C2). Two different standard-setting methods were applied. For Reading and Listening, a variation of the modified Angoff approach (Cizek & Bunch, 2007; Livingston & Zeiky, 1982) was used to establish cut scores for CEFR levels B1 through C2. For Speaking and Writing, the Bookmark method (Zieky, 2001) was applied, using ordered sets of real test-taker responses that allowed panelists to directly assess candidates' English-language skills against the MCC definition for each CEFR level. Both methods relied on the professional judgment of experienced panelists, who were trained, calibrated, and guided through multiple rounds

of discussion and rating to produce the final recommended cut scores. Implementation details for both methods are provided in the following subsections.

2.1 Panelists

Fifteen experts served on the standard-setting panels. Table 2 presents a summary of the panelists' self-reported highest educational qualifications, while Tables 3 and 4 present their years of experiences as language teachers and working with the CEFR, respectively. Fourteen panelists had experience teaching English as a second or foreign language, and the remaining panelist reported working as an assessment specialist, researcher, or teacher educator. Across all panelists, the average years of teaching experience is 16 (SD=8.62). All panelists indicated experience working with the CEFR, averaging 13 years (SD=5.95).

Table 2. Panellists' Highest Educational Background

Level of Education	Count
Bachelor's degree (e.g., BA, BEd, BSc)	2
Master's degree (e.g., MA, MSc, MEd)	10
Ph.D. candidate (ABD)	1
Doctoral degree (e.g., PhD, EdD)	2

Table 3. Panellists' Teaching Experience (in years)

Years of Experience Teaching English as a Second or Foreign Language	Count
0	1
1-4	1
5-9	1
10-14	2
15 and above	10

Table 4. Panellists' Experience with the CEFR (in years)

Years of Experience Working with the CEFR	Count
0	0
1-4	1
5-9	4
10-14	4
15 and above	6

2.2 Pre-workshop preparations

Prior to the standard-setting workshops, panelists received a training guide outlining the study objectives, schedules, and procedures. The guide also presented an overview of the CEFR framework as well as information on the construct, structure, and intended uses of the CELPIP test. Panelists were also provided with relevant CEFR descriptor scales and resources to familiarize themselves with the standards.

Each panelist was asked to review these materials in advance and draft preliminary definitions of a minimally competent candidate (MCC) - a candidate who just barely meets the minimum language ability required to be classified at a specific level - for CEFR levels B1, B2, C1, and C2. To help calibrate their expectations and ensure that their judgments were grounded in a realistic understanding of test taker performance, all panelists completed the CELPIP exam themselves, gaining direct experience with the content and format of the test.

2.3 Standard-setting process

Separate standard-setting panels were conducted for each of the four test components, starting with Reading, followed by Listening, Speaking, and Writing. Each panel took place over two consecutive days and followed a consistent three-step process to establish cut scores for each CEFR level.

Day 1:

- Step 1: Define the Minimally Competent Candidate (MCC) for the targeted CEFR level.
- Step 2: Complete the first round of judgment individually, either by estimating item difficulty (for Reading and Listening), or by evaluating ordered performance samples (for Speaking and Writing).

Day 2:

- Step 3: Complete the second round of judgments: Review Round 1 results, discuss items or performance samples with the greatest rating discrepancies, and submit a second set of judgments.

This process was repeated for all targeted CEFR levels (B1 to C2) across four workshops, each adhering to the same structure (see Appendix A for the workshop agenda). Because the same procedures were applied to both components that used the same method (modified Angoff: Reading and Listening; Bookmark: Speaking and Writing), the implementation steps are described by method rather than by individual component in the subsections that follow.

2.3.1 Reading and Listening

For the Reading and Listening components of the CELPIP test, the modified Angoff method was adopted to set the cut scores for each CEFR level. The expert panel first discussed and agreed on the definition of the MCC for each CEFR level. Once consensus was achieved, panelists independently reviewed each test item and estimated the probability that a MCC would answer the item correctly. These initial ratings were analyzed and presented to the panel for discussion, after which panelists were given the opportunity to affirm or revise their first ratings.

Orientation

On the first day of the workshop, panelists participated in a general orientation session introducing the purpose and procedures of the modified Angoff standard-setting study. The session outlined the steps the panel would follow throughout the process and emphasized the importance of active participation and group discussion. Questions and comments about the procedure were addressed to ensure clarity and alignment among panelists before rating activities began.

Development of the Standard of Minimum Competence

Following the orientation, the panel developed a shared definition of the MCC by identifying what such a candidate would be able to do successfully and what they would find challenging at each CEFR level in listening or reading. This discussion was grounded in the panelists' interpretations of the relevant CEFR scales, with descriptors adapted to reflect the group's understanding of a borderline candidate just entering each CEFR level.

For the Listening component, the MCC definition was based primarily on CEFR descriptors for oral comprehension; for the Reading component, the focus was placed on descriptors for reading comprehension. However, the panelists were encouraged to refer to additional CEFR descriptor scales and resources to inform their development of the MCC. This process ensured that the standard of minimal competence was clearly articulated for use in subsequent judgments.

Calibration Exercises

After agreeing on the MCC definition for CEFR level B1, panelists completed practice ratings for a small set of test items using the modified Angoff procedure. For each item, they were asked "*How many out of 100 minimally competent candidates do you think will answer this question correctly?*" Individual ratings were aggregated and discussed to ensure consistent application of the MCC standard when judging item difficulty.

First Modified Angoff Rating

Following the calibration exercises, panelists independently reviewed all items in the exam using a web-based tool. For each item, they provided an answer and estimated the proportion of MCCs expected to answer the item correctly. Panelists were instructed to make their judgment within the context of CELPIP, considering its structure, format, and timing constraints, as well as relevant item characteristics, such as passage complexity (e.g., vocabulary, syntax, length, topic or subject, and information density) and question requirements.

Second Modified Angoff Rating

A second round of rating process was conducted on Day 2 to encourage reflection and achieve greater consensus among panelists. Prior to this session, each panelist received an individualized feedback report showing their responses to the exam questions, obtained CELPIP score for the

component, item-level ratings, and recommended cut scores for each CEFR level from Round 1. During the workshop, the distribution of panel ratings was reviewed, with particular attention to items showing the greatest discrepancies in Round 1. Panelists discussed their rationales, re-examined their ratings in light of the MCC standard, group discussion, and their individualized feedback reports, and then submitted revised ratings. This process was repeated for CEFR levels B1, B2, C1, C2.

2.3.2 Speaking and Writing

For the Speaking and Writing components of the CELPIP test, the Bookmark method (Zieky, 2001) was used to establish the concordance between CELPIP and CEFR levels. The Bookmark method is an item-centered standard-setting procedure widely applied in performance-based language assessments. In this context, actual test taker responses (i.e., writing and speaking samples) were first scored using the standard CELPIP rubrics and then arranged in ordered performance booklets from weakest to strongest performance.

After discussing and agreeing on the characteristics of MCC for each CEFR level, panelists independently reviewed the ordered responses and placed a "bookmark" at the point where performance first met the MCC definition. The procedure was conducted in two rounds, with group discussion to review results, resolve discrepancies, and enhance agreement before final bookmark placements were determined.

Orientation

On the first day of the workshop, panelists participated in an orientation session introducing the Bookmark standard-setting method. The session outlined the steps the panel would follow, explained the CELPIP scoring system, and emphasized the importance of active participation and discussion. Questions and comments were addressed to ensure panelists had a clear understanding of the process before engaging with the performance samples.

Development of the Standard of Minimal Competence

Following the orientation, the panel collaboratively developed a shared definition of the MCC for each CEFR level by identifying what such candidates would be able to do successfully and what they would find challenging in speaking or writing. This discussion was grounded in the panelists' interpretations of the relevant CEFR scales, with descriptors adapted to reflect

the group's understanding of a borderline candidate just entering each CEFR level.

For the Speaking component, the MCC definition was based primarily on CEFR descriptors for spoken production; for the Writing component, the focus was placed on descriptors for written production. However, the panelists were encouraged to refer to additional CEFR descriptor scales and resources to inform their development of the MCC. This process ensured that the standard of minimal competence was clearly articulated for use in subsequent judgments.

Calibration Exercise

To ensure consistent application of the MCC definition, a calibration exercise was conducted for the B1 level. Panelists independently reviewed the ordered booklet of CELPIP performance samples and identified the first performance sample that met the B1 MCC standard. Their selections were collected, summarized, and presented to the group for review. Panelists discussed their choices, justified their reasoning considering the MCC definition, and worked toward a more unified interpretation of the B1 threshold before proceeding to formal ratings.

Ordered Performance Booklets

Panelists were provided with ordered booklets of authentic CELPIP performance samples, organized from the lowest to the highest quality of performance based on previously assigned CELPIP scores.

Speaking Booklet: 24 responses to three Speaking tasks: Giving Advice, Describing a Scene, and Expressing Opinions. Samples were drawn from actual test-taker performances scored through the standard CELPIP rating process and ranged from levels 2 to 12. Each response was 1-1.5 minutes in length.

Writing Booklet: 22 responses to two Writing tasks: Writing an Email and Responding to a Survey Question. Samples were drawn from actual test-taker responses scored through the standard CELPIP rating process and ranged from levels 2 to 12. Responses were required to be 150-200 words, with the selected samples ranging from 101 to 215 words.

Panelists reviewed the samples sequentially, deciding for each one whether it met the MCC standard. Once panelists identified the first sample that met the threshold, they recorded its number as the cut point and

stopped reviewing the remaining samples. The CELPIP score associated with that selected sample became their recommended score for the CEFR level under consideration.

First Bookmark Rating

After completing the MCC definition for a given CEFR level, panelists applied the standard to conduct the first round of Bookmark ratings. Independently, they placed a "bookmark" at first performance sample that met the MCC definition for the targeted CEFR level. This same process of defining the MCC and conducting the first Bookmark placement was consistently followed for each of the CEFR levels (B1, B2, C1, and C2) in succession.

Second Bookmark Rating

The initial Bookmark placements were compiled and reviewed. On the second day, these results were presented to the panel for discussion, with particular focus on samples where judgments showed the greatest discrepancies. For each CEFR level, the panel reviewed the overall distribution of the selected MCC samples (i.e., the count of panelists choosing each sample as the threshold), discussed the rationale behind their choices, and reconsidered their judgments in light of peer feedback and the agreed MCC standard. Each panelist also received their own CELPIP level for the component to help contextualize their decisions. Following the discussion, panelists independently submitted a second round of Bookmark placements (i.e., the sample selected as the MCC threshold) via a web-based tool. This process was repeated for CEFR levels B1, B2, C1 and C2. The final Bookmark placements were recorded and used to determine the CELPIP score levels corresponding to each CEFR level.

2.4 Standard-setting evaluation survey

After each standard-setting workshop, panelists completed a survey designed to evaluate their experience and the overall standard-setting process. This survey collected feedback on the clarity of instructions, the appropriateness of materials provided, and the adequacy of support throughout the workshop. Panelists were also encouraged to comment on the structure of the workshop, the effectiveness of group discussions, any challenges they encountered, and their confidence in the judgments they had made for each CEFR level.

3. Results

This section presents the outcomes of the standard-setting studies for all four CELPIP components (Reading, Listening, Speaking, and Writing). Results are organized by component and reported in three parts for each: the panelists' cut score judgments, the correspondence between CELPIP and CEFR levels, and panelists' confidence in the judgments made for that component.

First, the panel's judgments are summarized by round for each component. Round 2 ratings are considered the panel's final recommendations, reflecting insights and refinements from group discussion after the initial round.

Second, the correspondence between CELPIP scores and CEFR levels is presented. For Reading and Listening, the CEFR cut scores are based on Round 2 averages. Because a cut score represents the minimum score required to meet a CEFR level, the next higher CELPIP level (rounded up) is considered to correspond to that level. For example, if the recommended C1 cut score for Listening is 28.25, and, hypothetically, the adjacent CELPIP level cut scores are 27.95 (level 8) and 30.33 (level 9), then CELPIP level 9 corresponds to CEFR C1.

For Speaking and Writing, the CEFR correspondences are determined using the mode of Round 2 Bookmark placements. Although the booklets presented to the panel did not include scoring information, all performance samples had been previously scored following standard operational procedures and reviewed by internal teams, ensuring each sample was associated with a proper CELPIP level. Panelists identified the first performance sample meeting the minimal CEFR requirements. Using the mode as the summary statistic reflects the group consensus, reduces the impact of outlier judgments, and eliminates the need to round mean scores, since CELPIP levels are reported as integers.

Third, the panelists' confidence ratings for each component are summarized to show how confident they felt in their own and the group's judgments for that specific component and CEFR level.

Following the component-specific results, we summarize panelists' reflections on the standard-setting process and their workshop experience. Finally, we present the aggregated results, combining the mapping results

from all four components to report the recommended overall CELPIP level correspondences to each CEFR level.

3.1 Reading

Table 5 presents the results for the Reading component, for which the maximum raw score is 38 points. Between Rounds 1 and 2, the panel's mean cut score recommendation remained unchanged for B1, showed minimal change for C1 and C2, and decreased slightly for B2. For all levels, standard deviations decreased from Round 1 to Round 2, indicating greater consensus among panelists following inter-round discussion and feedback.

Table 5. Average Panel Ratings: Reading

	Round 1				Round 2			
CEFR Levels	Mean	SD	Min	Max	Mean	SD	Min	Max
B1	14.51	3.00	7.75	18.90	14.51	1.26	12.55	17.30
B2	24.22	3.79	16.55	29.85	23.36	2.78	18.05	28.85
C1	30.93	2.74	25.35	35.30	30.99	1.73	27.75	34.35
C2	34.56	1.81	31.95	37.40	34.67	1.28	32.25	36.65

Table 6 summarizes the recommended mapping of CELPIP Reading levels to CEFR levels B1 to C2.

Table 6. Mapping CELPIP Levels onto the CEFR Scale: Reading

CEFR Levels	Reading CELPIP Levels
B1	5-7
B2	8-9
C1	10-11
C2	12

Table 7 presents the panel's reported confidence. Following the standard-setting workshop for the Reading component, panelists rated their confidence in clearly understanding the minimally competent candidate, their individual judgments, and the group decisions at each CEFR level. Ratings were provided on a four-point scale, ranging from strongly disagree (1) to strongly agree (4). Their reflections serve as a source of validity evidence for the outcomes of the standard-setting process.

Table 7. Panellists' Confidence in the Recommended Cut Scores: Reading

CEFR Levels	Clear Understanding of the MCC	Confidence in Personal Judgment	Confidence in Group Decision
B1	3.67 (0.47)	3.67 (0.47)	3.67 (0.47)
B2	3.83 (0.37)	3.60 (0.49)	3.67 (0.47)
C1	3.75 (0.43)	3.67 (0.47)	3.60 (0.49)
C2	3.75 (0.43)	3.80 (0.40)	3.67 (0.47)

Note: Responses were coded as follows: strongly disagree = 1, disagree = 2, agree = 3, and strongly agree = 4. For each CEFR level, both the mean confidence score and the standard deviation (SD, shown in parentheses) are presented.

Across all statements, the mean ratings were above 3.60 (with 4 as the maximum), and the standard deviations were relatively small (< 0.49), indicating that most panelists were very confident in their understanding, judgment, and the group decisions across all CEFR levels. Indeed, all panelists reported agreement or strong agreement with the statements at every CEFR level.

3.2 Listening

Table 8 presents the results for the Listening component, which has a maximum raw score of 38 points. Between Rounds 1 and 2, the mean cut score recommendations showed minimal change for all CEFR levels. The standard deviations decreased from Round 1 to Round 2 for all levels, indicating greater panelist consensus following inter-round discussion and feedback.

Table 8. Average Panel Ratings: Listening

	Round 1				**Round 2**			
CEFR Levels	Mean	SD	Min	Max	Mean	SD	Min	Max
B1	13.69	2.34	9.00	16.75	13.61	1.85	9.70	16.35
B2	21.34	4.08	13.45	27.20	21.77	2.99	17.85	26.55
C1	29.36	2.23	25.60	32.70	29.55	1.84	26.15	32.85
C2	34.35	1.23	30.95	36.05	34.28	1.10	31.20	35.60

Table 9 summarizes the recommended mapping of CELPIP Listening levels to CEFR levels B1 to C2.

Table 9. Mapping CELPIP Levels onto the CEFR Scale: Listening

CEFR Levels	Listening CELPIP Levels
B1	6
B2	7-9
C1	10-11
C2	12

Table 10 presents the panel's reported confidence following the standard-setting workshop for the Listening component. Panelists rated their confidence in three areas at each CEFR level: their understanding of the minimally competent candidate, their individual judgments, and the group decisions reached. Ratings were provided on a four-point scale, ranging from strongly disagree (1) to strongly agree (4), and these reflections serve as a source of validity evidence for the outcomes of the standard-setting process.

Table 10. Panellists' Confidence in the Standard-Setting Results: Listening

CEFR Levels	Clear Understanding of the MCC	Confidence in Personal Judgment	Confidence in Group Decision
B1	3.87 (0.34)	3.80 (0.40)	3.73 (0.44)
B2	3.80 (0.40)	3.73 (0.44)	3.67 (0.47)
C1	3.80 (0.40)	3.73 (0.44)	3.67 (0.47)
C2	3.87 (0.34)	3.80 (0.40)	3.67 (0.47)

Note: Responses were coded as follows: strongly disagree = 1, disagree = 2, agree = 3, and strongly agree = 4. For each CEFR level, both the mean confidence score and the standard deviation (SD, shown in parentheses) are presented.

Across all statements, the mean ratings exceeded 3.67 (with 4 as the maximum), and the standard deviations were relatively small (< 0.47), indicating that panelists were highly confident in their understanding, judgment, and the group decisions across all CEFR levels. Indeed, all panelists reported agreement or strongly agreement with the statements at every CEFR level.

3.3 Speaking

Table 11 presents the results for the Speaking component, for which score levels range from 0 to 12 in integer values. Between Rounds 1 and 2, the mode of the panel's CELPIP score recommendation remained unchanged for B1, B2, C1, and decreased by one level for C2. For all levels, the range of the panel selected performance levels decreased from Round 1 to Round 2, reflecting greater consensus among panelists following inter-round discussion and feedback. At Round 2, the median matched the mode for all levels, indicating that the mapping would be consistent whether using the mode or median to summarize group judgments.

Table 11. Summary of Panel Ratings: Speaking

	Round 1				Round 2			
CEFR Levels	Mode	Median	Min	Max	Mode	Median	Min	Max
B1	5	5	4	5	5	5	5	5
B2	7	7	6	8	7	7	6	7
C1	9	9	9	10	9	9	9	9
C2	12	11	9	12	11	11	11	12

Table 12 summarizes the recommended mapping of CELPIP Speaking levels to CEFR levels B1 to C2.

Table 12. Mapping CELPIP Levels onto the CEFR Scale: Speaking

CEFR Levels	Speaking CELPIP Levels
B1	5-6
B2	7-8
C1	9-10
C2	11-12

Table 13 summarizes panelists' confidence following the standard-setting workshop for the Speaking component. At each CEFR level, panelists rated their confidence in understanding the minimally competent candidate, their individual judgments, and the group decisions on a four-point scale (1 = strongly disagree, 4 = strongly agree). These ratings provide validity evidence for the standard-setting outcomes.

Table 13. Panellists' Confidence in the Standard-Setting Results: Speaking

CEFR Levels	Clear Understanding of the MCC	Confidence in Personal Judgment	Confidence in Group Decision
B1	3.80 (0.40)	3.80 (0.40)	3.80 (0.40)
B2	3.87 (0.34)	3.67 (0.47)	3.67 (0.47)
C1	3.87 (0.34)	3.73 (0.44)	3.67 (0.47)
C2	3.80 (0.40)	3.73 (0.44)	3.53 (0.62)

Note: Responses were coded as follows: strongly disagree = 1, disagree = 2, agree = 3, and strongly agree = 4. For each CEFR level, both the mean confidence score and the standard deviation (SD, shown in parentheses) are presented.

The mean ratings for all statements exceeded 3.53, with standard deviations below 0.62, indicating most panelists showing high confidence across all areas. All but one panelist reported agreement or strong agreement at every CEFR level. One panelist reported disagreement with the group decision at C2; although prompted to explain their rating, no reasons were provided for this response.

3.4 Writing

Table 14 presents the results for the Writing component, for which score levels range from 0 to 12 in integer values. Between Rounds 1 and 2, the mode of the panel's CELPIP score recommendations remained unchanged for B1, increased by one level for B2 and C2, and increased by two levels for C1. For all levels except B1, the range of panel selected performance levels decreased from Round 1 to Round 2, reflecting greater consensus among panelists following inter-round discussion and feedback. The range for B1 was small at Round 1 (min = 4, max = 5) and remained unchanged. At Round 2, the median matched the mode for all levels, indicating that the mapping would be consistent whether using the mode or median to summarize group judgments.

Table 14. Summary of Panel Ratings: Writing

	Round 1				Round 2			
CEFR Levels	Mode	Median	Min	Max	Mode	Median	Min	Max
B1	4	4	4	5	4	4	4	5
B2	6	6	5	7	7	7	7	7
C1	8	8	7	10	10	10	9	10
C2	11	11	9	12	12	12	11	12

Table 15 summarizes the recommended mapping of CELPIP Writing levels to CEFR levels B1 to C2.

Table 15. Mapping CELPIP Levels onto the CEFR Scale: Writing

CEFR Levels	Writing CELPIP Levels
B1	4-6
B2	7-9
C1	10-11
C2	12

Table 16 summarizes panelists' confidence following the standard-setting workshop for the Writing component. At each CEFR level, panelists rated their confidence in understanding the minimally competent candidate, their individual judgments, and the group decisions on a four-point scale (1 = strongly disagree, 4 = strongly agree). The mean ratings for all statements exceed 3.67, with standard deviations below 0.47, indicating consistently high confidence across all areas. Indeed, all panelists reported agreement or strong agreement at every CEFR level. These ratings provide validity evidence for the standard-setting outcomes.

Table 16. Panellists' Confidence in the Standard-Setting Results: Writing

CEFR Levels	Clear Understanding of the MCC	Confidence in Personal Judgment	Confidence in Group Decision
B1	3.93 (0.25)	3.87 (0.34)	3.87 (0.34)
B2	3.80 (0.40)	3.73 (0.44)	3.87 (0.34)
C1	3.73 (0.44)	3.80 (0.40)	3.87 (0.34)
C2	3.87 (0.34)	3.67 (0.47)	3.67 (0.47)

Note: Responses were coded as follows: strongly disagree = 1, disagree = 2, agree = 3, and strongly agree = 4. For each CEFR level, both the mean confidence score and the standard deviation (SD, shown in parentheses) are presented.

3.5 Evaluation of the standard-setting process

Table 17 summarizes panelists' evaluations of the standard-setting process across the four CELPIP components. Ratings were provided on a four-point scale (1 = strongly disagree, 4 = strongly agree) and addressed two broad areas: training and preparation (e.g., helpfulness of pre-workshop materials, clarity of procedures, and usefulness of Round 1 feedback) and participation and process (e.g., sufficiency of discussion opportunities, clarity of rating instructions, and respect shown by panelists and facilitators).

Table 17. Panellists' Ratings of the Standard-Setting Process

	Reading		**Listening**		**Speaking**		**Writing**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Training & Preparation								
Helpful pre-workshop materials	3.67	0.47	3.80	0.40	3.87	0.34	3.93	0.25
Informative Day 1 orientation	3.73	0.44	3.87	0.34	3.87	0.34	3.93	0.25
Clear method & procedures	3.40	0.49	3.93	0.25	3.87	0.34	4.00	0.00
Helpful Round 1 feedback	3.67	0.47	3.80	0.40	3.73	0.44	4.00	0.00
Participation & Process								
Enough discussion opportunities	3.87	0.34	3.93	0.25	3.93	0.25	3.93	0.25
Clear rating instructions	3.53	0.50	3.93	0.25	3.87	0.34	3.93	0.25
Enough time for tasks	3.87	0.34	4.00	0.00	4.00	0.00	4.00	0.00
Respected by panelists	3.80	0.40	3.93	0.25	3.87	0.34	3.93	0.25
Respected by facilitators	3.80	0.40	4.00	0.00	3.93	0.25	4.00	0.00

Across all components, the mean ratings were consistently high (all ≥ 3.40) with relatively small standard deviations (all ≤ 0.50), indicating strong consensus among panelists and generally positive perceptions of the standard-setting process. The highest ratings were observed for "Enough time for tasks" (mean = 4.00 for Listening, Speaking, and Writing) and "Clear method & procedures" for Writing (mean = 4.00), suggesting that panelists found the pacing and procedural clarity highly satisfactory.

These results provide important validity evidence supporting both the process and outcomes of the study. High levels of agreement across components indicate that panelists felt well-prepared, had sufficient opportunity to participate in discussions, and were confident in the fairness and professionalism of the process. This strengthens the interpretability of the resulting cut scores and the credibility of the CELPIP–CEFR mapping recommendations.

3.6 Overall

Table 18 summarizes the recommended mapping of CELPIP overall score levels to CEFR levels B1 through C2. In addition to individual component scores, the CELPIP test reports an overall performance level (labelled 'Average' on the score report), calculated as the mean of the four component levels and rounded to the nearest integer. Applying the same procedure, the minimum CELPIP levels corresponding to each CEFR level were averaged across components and rounded to the nearest integer to derive the minimum overall CELPIP level required for each CEFR level.

Table 18. Mapping CELPIP Levels onto the CEFR Scale: Overall

CEFR Levels	CELPIP Levels
B1	5-6
B2	7-9
C1	10-11
C2	12

4. Conclusions

The purpose of this study was to map the Reading, Listening, Speaking, and Writing components of the CELPIP test to the CEFR levels B1 through C2. Two well-established standard-setting methods, the modified Angoff (for Reading and Listening) and the Bookmark method (for Speaking and Writing), were used to identify the minimum test scores to reach each targeted CEFR level.

The standard-setting process followed best practices recommended in the literature (AERA APA, & NCME, 2014; Cizek, 2012; Cizek & Bunch, 2007; Katz, 2019; Hambleton & Pitoniak, 2006). This included careful selection of qualified panelists and ensuring a sufficiently large panel to capture diverse perspectives; providing adequate time for panelists to develop a shared understanding; comprehensive training; defining MCC; conducting multiple

rounds of judgments, and using data to inform their judgments. Panelists' responses to the post-workshop evaluation survey provide important validity evidence in support of the standard-setting outcomes. Across all components, panelists demonstrated relatively strong agreement, particularly following the second round of ratings, where consensus improved after structured discussion and feedback. Panelists also reported high confidence in their understanding of minimally competent candidates and in their own and the final group decisions, further supporting the credibility of the recommended cut scores.

The resulting CELPIP–CEFR mappings provide clear guidance for policymakers, institutions, and other stakeholders who rely on CELPIP scores to inform admissions, hiring, and other decisions. These mappings facilitate meaningful interpretation of CELPIP levels and promote consistency with international language proficiency standards.

While the findings are well-supported, they should be interpreted with consideration of certain limitations. Standard-setting judgments, by design, are based on expert consensus, which may reflect the perspectives of the specific panel convened for this study. Additionally, the CEFR provides a broad descriptive framework of language competence, whereas standardized assessments, such as CELPIP, capture a snapshot of test takers' proficiency at a specific point in time, based on performance on a sample of tasks in a testing environment. This fundamental difference between a broad framework and a specific test means that the resulting correspondences represent informed approximations rather than exact one-to-one matches. Finally, future updates to the CELPIP test, CEFR descriptors, or test-taking populations may warrant re-evaluation of these cut scores to ensure ongoing alignment.

In conclusion, this report has documented the standard-setting studies conducted to establish cut scores for the Reading, Listening, Speaking, and Writing components of the CELPIP test, mapping CELPIP performance levels to the CEFR. The recommended cut scores and resulting CELPIP–CEFR correspondences are supported by validity evidence and contributes to strengthening the meaningful use of CELPIP scores. These results provide defensible and reliable guidance for stakeholders, enabling informed and consistent decisions in educational, professional, and policy contexts.

References

AERA, APA, & NCME (2014). Standards for educational and psychological tests. Washington DC: American Educational Research Association.

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. Routledge.

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

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Katz, I. R. (2019). *Standard Setting Panelist Cognition: A Framework and Implications for Practice*. ETS Research Report Series.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. *Educational measurement, 4*(1), 433-470.

Livingston, S. & Zeiky, M. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting Performance Standards* (pp. 19–52). Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix A: Standard-Setting Workshop Agenda

Day 1

Introduction to the Standard-Setting Methodology and Procedures

CEFR B1 Discussion and First Rating

- Define MCC at CEFR B1
- Calibration Exercise for CEFR B1
- First Round of Judgements for CEFR B1

CEFR B2 Discussion and First Rating

- Definition of MCC at CEFR B2
- First Round of Judgements for CEFR B2

CEFR C1 Discussion and First Rating

- Definition of MCC at CEFR C1
- First Round of Judgements for CEFR C1

CEFR C2 Discussion and First Rating

- Definition of MCC at CEFR C2
- First Round of Judgements for CEFR C2

Day 2

Review and Discussion of First Round Results

Second Round of Judgements for CEFR B1

Second Round of Judgements for CEFR B2

Second Round of Judgements for CEFR C1

Second Round of Judgements for CEFR C2