AVANT STAMP CEFR ENGLISH PROFICIENCY TEST: READING AND LISTENING CUT SCORE STUDY TECHNICAL REPORT

DR. CHARALAMBOS KOLLIAS POLYTOMOUS LIMITED

Technical Report

Dr Charalambos Kollias
Polytomous Limited

AVANT STAMP CEFR ENGLISH PROFICIENCY TEST: READING AND LISTENING CUT SCORE STUDY TECHNICAL REPORT

Dr. Charalambos Kollias
Polytomous Limited

To cite:

Kollias, C. (2025). Avant STAMP CEFR English proficiency test. Reading and Listening cut score study. Technical Report. Polytomous Limited, UK.

Contents

List of figures

List of tables

Acknowledgments

Polytomous Limited extends its sincere gratitude to Avant Assessment for their invaluable support throughout the cut score study. Their contributions, including organizing the workshops, recruiting judges, and providing essential information and data, were instrumental to the success of the study. Special appreciation is also extended to the judges who participated in the workshops. Their dedication and hard work were crucial in establishing the CEFR cut scores, and their commitment is deeply appreciated.

Executive Summary

This report outlines the outcomes of two virtual workshops conducted to establish Common European Framework of Reference (CEFR) cut scores for the Reading and Listening sections of the Avant STAMP for CEFR English proficiency test. Held between April 27 and May 26, 2024, these workshops employed the Item Descriptor (ID) Matching Method, a widely recommended approach for CEFR alignment, to ensure reliable and valid cut scores. This method emphasizes the direct alignment of test items with CEFR descriptors, enabling judges to systematically evaluate the linguistic demands of each item against proficiency standards.

The Avant STAMP for CEFR English proficiency test is a multistage, computer-adaptive assessment targeting English proficiency levels from A1 to C1. It evaluates Reading, Writing, Listening, and Speaking skills, making it a valuable tool for educational institutions and workplaces. Aligning the Reading and Listening sections with CEFR standards ensures the consistency and comparability of these sections in language proficiency evaluation.

Each workshop consisted of 14 trained judges representing diverse linguistic and educational expertise. The judges were situated across five continents: Oceania, South America, North America, Europe, and Asia.

The workshops involved five key stages:

1. **Orientation and Familiarization:** Judges were introduced to the CEFR framework and test content.
2. **Training:** Judges were trained in the ID Matching Method using practice items.
3. **Item Review and Matching (Round 1):** Judges assigned CEFR levels to test items and documented their reasoning.
4. **Consensus Building:** Discussions among judges facilitated shared understanding and alignment.
5. **Final Review (Round 2):** Judges revisited their decisions, refined their ratings, and completed final evaluations.

Key Findings

- **Judge Expertise and Agreement:** All judges demonstrated a deep understanding of the CEFR descriptors, with agreement levels exceeding 95% in both Reading and Listening workshops.

- **Rasch Analyses:** Person reliability indices were high (0.91 for Reading and 0.92 for Listening), indicating the item banks' robustness and their ability to distinguish multiple proficiency levels.

- **Classification Accuracy:** Cut scores achieved classification accuracy rates above 95% across all CEFR levels, validating the precision of the standard setting process.

- **Cut Score Consistency:** Conditional standard errors of measurement (cSEM) were minimized at the cut scores, supporting the reliability of learner classifications.

- **Item Bank Quality:** Item reliability indices for both sections reached 0.99, surpassing the minimum threshold of 0.90, confirming sufficient item bank sizes and construct validity.

The report concludes that the Reading and Listening cut scores are valid and reliable, aligning effectively with CEFR standards. These outcomes reinforce the test's utility for assessing English language proficiency and guiding educational and professional decisions.

1. Introduction

This report outlines the outcomes of two workshops conducted to establish the Common European Framework of Reference (CEFR) cut scores for the Reading and Listening sections of the Avant STAMP for CEFR English proficiency test. Both workshops were held in virtual environments between April 27 and May 26, 2024. The report details the methodology used and the validation process for establishing and reviewing the CEFR cut scores.

1.1 Background to the Avant STAMP for CEFR English proficiency test

The Avant STAMP for CEFR English proficiency test is a multistage, computer-adaptive assessment designed to measure English language proficiency across Reading, Writing, Listening, and Speaking skills, aligned with CEFR levels A1 to C1. This multistage design ensures that the difficulty of questions adjusts dynamically based on the test-takers' responses, providing a more accurate and personalized measure of proficiency. Targeted at learners aged 13 and older, the test comprises approximately 30 questions each in the Reading and Listening sections (35-40 minutes per section) and three prompts for each in the Writing and Speaking sections (20-25 minutes per section). It is intended to be used by educational institutions to monitor English language learning progress and to verify English proficiency standards for program enrolment, advancement, or graduation, as well as by workplaces as a means to confirm candidate's general English proficiency. The assessment requires proctoring at all times, which can be conducted onsite by an authorized proctor or through AvantProctor, a remote and secure proctoring service. For further information, see https://www.avantassessment.com/tests/stamp/cefr.

1.2 Aligning examinations to the CEFR

The Common European Framework of Reference for Languages (CEFR) offers a comprehensive framework for language learning, teaching, and assessment. It is widely used to ensure consistency and comparability in evaluating language proficiency across diverse contexts and languages. Aligning examinations with the CEFR requires a systematic process to confirm that test content, design, and scoring accurately reflect the CEFR's levels of language competence.

The Council of Europe's Manual for Relating Language Examinations to the CEFR (2009) outlines a structured approach for this alignment. The process involves five key stages (see Figure 1), summarized below:

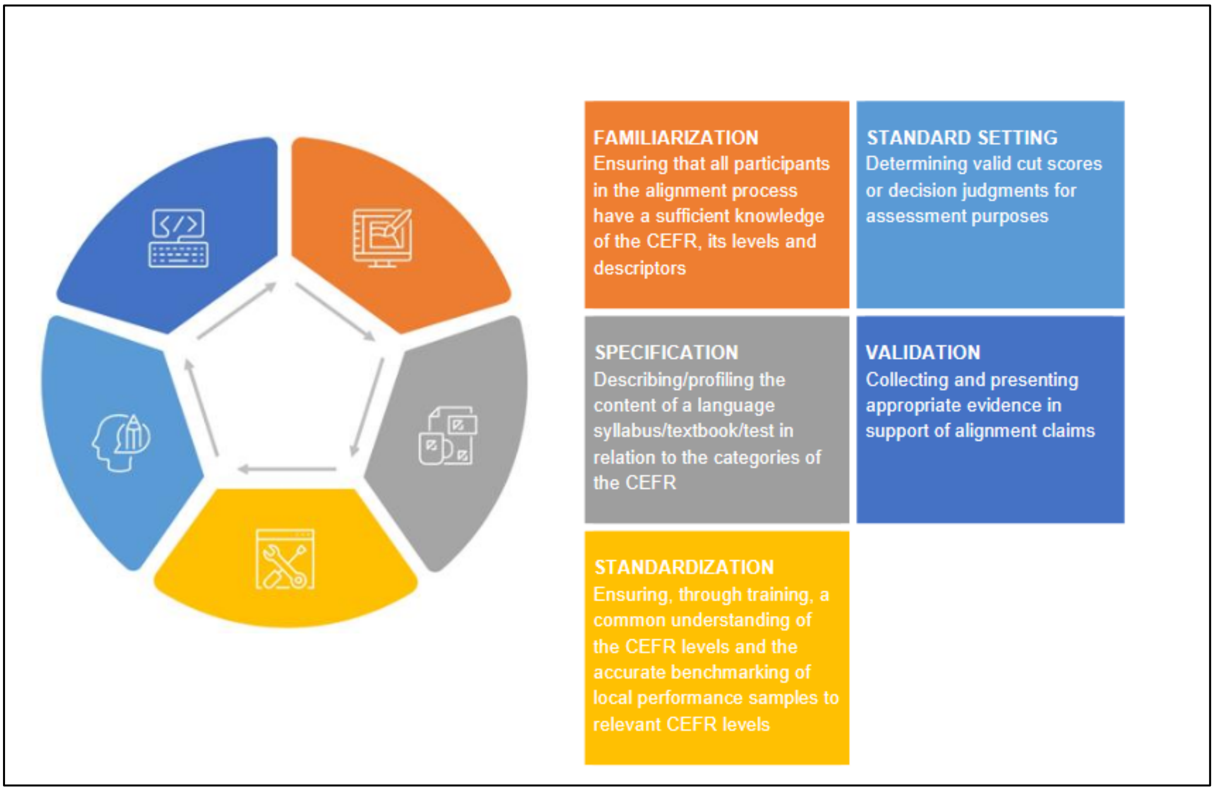
1. *Familiarisation*: In this stage, stakeholders (e.g., test developers, judges, etc.) become familiar with the CEFR descriptors, levels, and principles. This ensures that all participants understand the framework and its implications for examination development, standard setting, and validation.

2. *Specification*: Examination content and tasks are explicitly linked to the CEFR levels following a 'self-audit' process. This involves defining the linguistic and pragmatic competencies expected at each level and ensuring the test items reflect these specifications.

3. *Standardisation Training and Benchmarking*: This stage involves extensive training for judges to create a common understanding of the CEFR levels and descriptors to ensure their consistent application. Accepted CEFR benchmarks serve as reference performances or items during this stage to standardize judges and align results.

4. *Standard Setting*: Thresholds for CEFR levels are established during this stage. Using methods like the Item Descriptor (ID) Matching method or benchmarking, cut scores for each level are set to reliably and validly distinguish between different proficiency levels.

5. *Validation*: The alignment of the examination to the CEFR is empirically validated. This includes collecting evidence to support claims about the reliability and validity of the test scores in relation to the CEFR levels, often involving statistical analyses, expert review, and performance comparisons.



FAMILIARIZATION
Ensuring that all participants in the alignment process have a sufficient knowledge of the CEFR, its levels and descriptors

STANDARD SETTING
Determining valid cut scores or decision judgments for assessment purposes

SPECIFICATION
Describing/profiling the content of a language syllabus/textbook/test in relation to the categories of the CEFR

VALIDATION
Collecting and presenting appropriate evidence in support of alignment claims

STANDARDIZATION
Ensuring, through training, a common understanding of the CEFR levels and the accurate benchmarking of local performance samples to relevant CEFR levels

Figure 1: Visual representation of CEFR alignment procedures.

Source: British Council et al., 2022.

2. Standard Setting Methodology

This section outlines the judge recruitment process, provides a profile of the judges, and describes the procedures followed before and during the workshop as well as the standard method chosen for setting cut scores. The workshop methodology is presented in a chronological sequence.

2.1 Judge recruitment

Judges were recruited directly or indirectly using a mixture of purposive and snowball sampling methods. They either received the project information sheet via email or found it posted on targeted web platforms. To participate in the workshops, judges needed to fulfill the following minimum requirements:

- a minimum of 5 years of experience teaching English as a Second Language (ESL)/ English as a Foreign EFL (EFL) and/or strong familiarity with English CEFR levels and descriptors,

- at least a graduate degree in Teaching English as a Foreign/Second Language or a related field (e.g. Applied Linguistics, Language Testing and Assessment, etc.),

- private access to a personal computer, and

- private access to a microphone and web camera.

The first criterion ensured that judges were familiar with (1) the test-taker population and/or (2) the CEFR descriptors and levels, as outlined by Raymond and Reid (2001). Judges were invited to complete an online background questionnaire to express their interest in participating in one or both workshops. Each workshop was scheduled to take place over four online synchronous sessions, totaling approximately 19.5 hours each (see Appendix A for the agenda). Those who met the minimum criteria were assigned three pre-workshop tasks (see section 2.3.1 for a description).

According to Brandon (2004), a minimum of 10 judges should be invited to a standard setting workshop. Seventeen judges were invited to participate in one or both of the virtual workshops. Each workshop had a panel of 14 judges. Table 1 presents the background information of the 17 judges.

Table 1: Judge demographic characteristics (N = 17)

Gender	Female (14) Male (3)
Highest Degree	Master's Degree (15) Ph.D. (2)
Country of residence	Australia (1) Argentina (1) Brazil (1) Canada (3) Colombia (3) Greece (2) Hong Kong (1) Romania (1) Qatar (1) UK (2) USA (1)
Years of ESL/EFL teaching	6 – 9 years (1) 10+ years (16)
CEFR levels taught	A1 – C1 only (3) A1 – C2 (14)
Current position	Postgraduate/ PhD student (2) Practicum/ Academic coordinator (3) State/Private English language teacher (5) Academic Director / Head of evaluation (2) College Professor/ (Senior) Lecturer/ Academic (3) Awarding body examiner/ Language testing consultant (2)
Standard setting workshop experience	Yes (11) No (6)

The panel of 17 judges brought a wealth of experience and diverse perspectives to the workshops. An overwhelming majority (94.11%) had over a decade of experience teaching ESL/EFL. Geographically, they were situated across five continents: Oceania, South America, North America, Europe, and Asia. Their current professional roles included postgraduate or PhD students, teaching and instructional positions such as state or private English language teachers, college professors, senior lecturers, or seasonal academics; coordination and administrative roles including practicum or academic coordinators and academic directors; and positions in assessment and evaluation, such as examiners for awarding bodies, language testing consultants, or heads of evaluation.

2.2 Selection of standard setting method

Standard setting is the process of establishing a cut score, a specific point on the test scale that categorizes test takers into two groups, each representing different levels of proficiency in the skill being assessed (Hambleton & Eignor,1979). This process involves gathering a panel of experts (judges) for a standard setting workshop, where they propose a cut score for a specific examination. The policy committee then assesses the workshop documentation and the judges' recommendations before finalizing the cut score (Kaftandjieva, 2004; Cizek, Bunch, & Koons, 2004).

The Item Descriptor (ID) Matching Method (Ferrara & Lewis, 2012) , recognized as one of the recommended approaches for CEFR alignment studies (Council of Europe, 2009), was employed in both the Reading and Listening workshops. This standard setting method involves judges reviewing items and aligning them with specific proficiency level descriptors (e.g., CEFR levels). Judges evaluate how well the content and difficulty of each item correspond to the language competencies outlined in the descriptors. The method focuses on establishing a direct connection between test items and clearly defined performance standards.

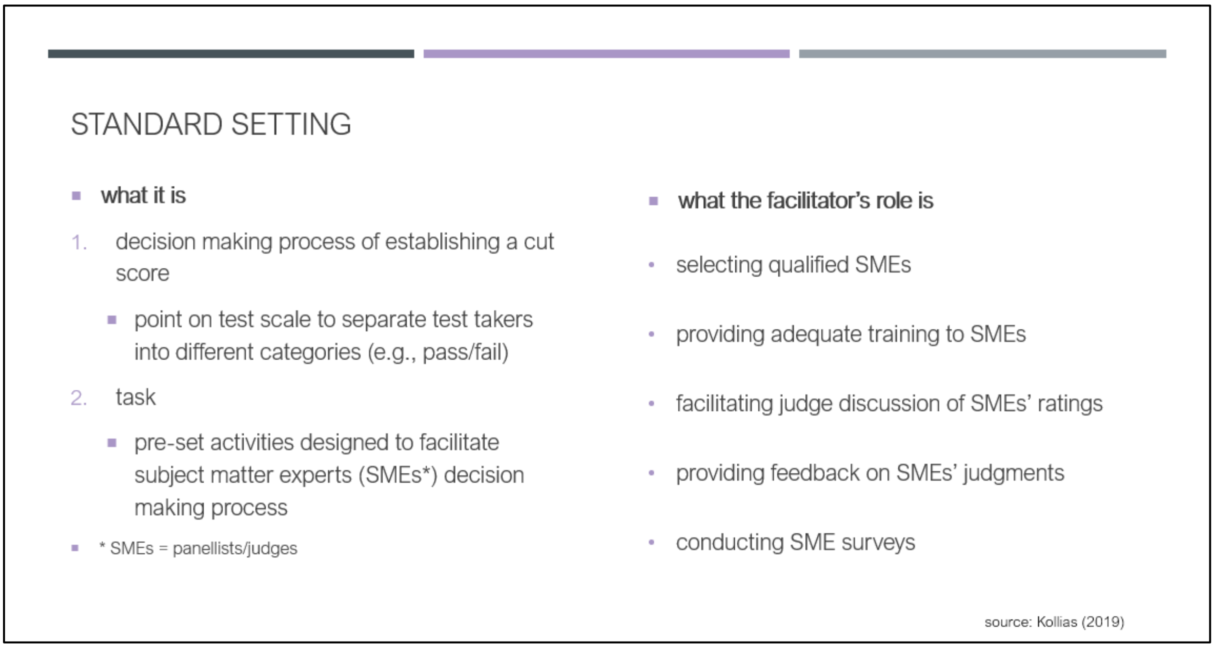
This method offers several advantages in CEFR standard setting, such as ensuring that test items are directly aligned with CEFR level descriptors and enhancing validity and construct representation by focusing on the linguistic skills and knowledge required at each level. The method also promotes transparency by providing a clear rationale for assigning items to specific proficiency levels, which increases the credibility of the process and adds transparency to the Specification stage of the alignment process. Its adaptability allows for the inclusion of various item formats, making it suitable for diverse types of language assessments. Additionally, it supports judge calibration by anchoring decisions to well-defined descriptors, fostering consistency among judges. These strengths make the method particularly effective for aligning language assessments with CEFR standards and ensuring that test outcomes accurately reflect the intended proficiency levels (Kanistra, 2025, forthcoming).

2.3 Virtual workshop stages

Each virtual workshop consisted of six main stages: the orientation stage, the familiarization stage, the training in the method stage, the item review and matching stage (Round 1), the consensus-building discussion stage, and the final review stage (Round 2). At the end of each stage, judges conducted an evaluation of the process. These evaluations (refer to Section 3.1 for details) were reviewed and addressed as necessary. Below is a comprehensive description of each stage.

2.3.1 The orientation stage

The orientation stage included providing judges with (i) an overview of the virtual workshop, (ii) their role within the workshop, and (iii) the netiquette to observe during the virtual sessions (see Figures 2 and 3).



STANDARD SETTING

- what it is

1. decision making process of establishing a cut score
   - point on test scale to separate test takers into different categories (e.g., pass/fail)
2. task
   - pre-set activities designed to facilitate subject matter experts (SMEs*) decision making process

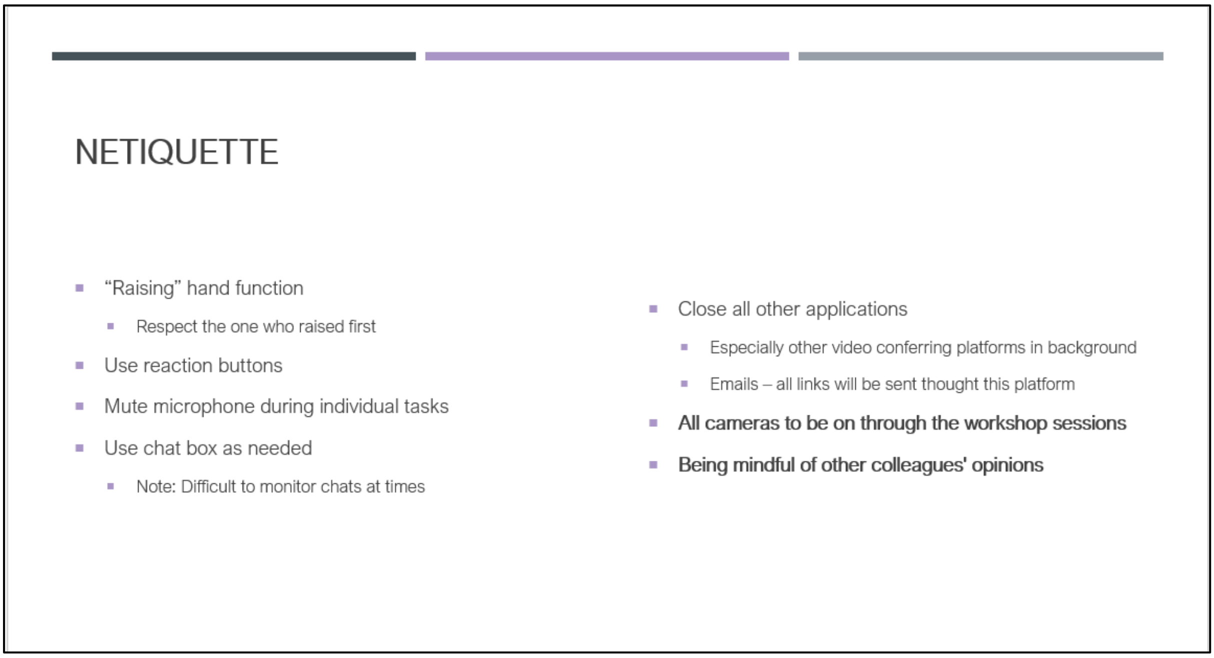
- * SMEs = panellists/judges

- what the facilitator's role is

- selecting qualified SMEs
- providing adequate training to SMEs
- facilitating judge discussion of SMEs' ratings
- providing feedback on SMEs' judgments
- conducting SME surveys

source: Kollias (2019)

Figure 2: Slide from of standard setting workshop overview



NETIQUETTE

- "Raising" hand function
  - Respect the one who raised first
- Use reaction buttons
- Mute microphone during individual tasks
- Use chat box as needed
  - Note: Difficult to monitor chats at times

- Close all other applications
  - Especially other video conferring platforms in background
  - Emails – all links will be sent thought this platform
- All cameras to be on through the workshop sessions
- Being mindful of other colleagues' opinions

Figure 3: Slide describing the netiquette to be used throughout the workshop

2.3.2 The familiarization stage

The familiarization stage aimed to ensure that judges became familiar with the CEFR levels and descriptors, reviewed the test items to understand their content, format, and intent, and were trained to align test items with the CEFR descriptors by focusing on specific skills and competencies outlined.

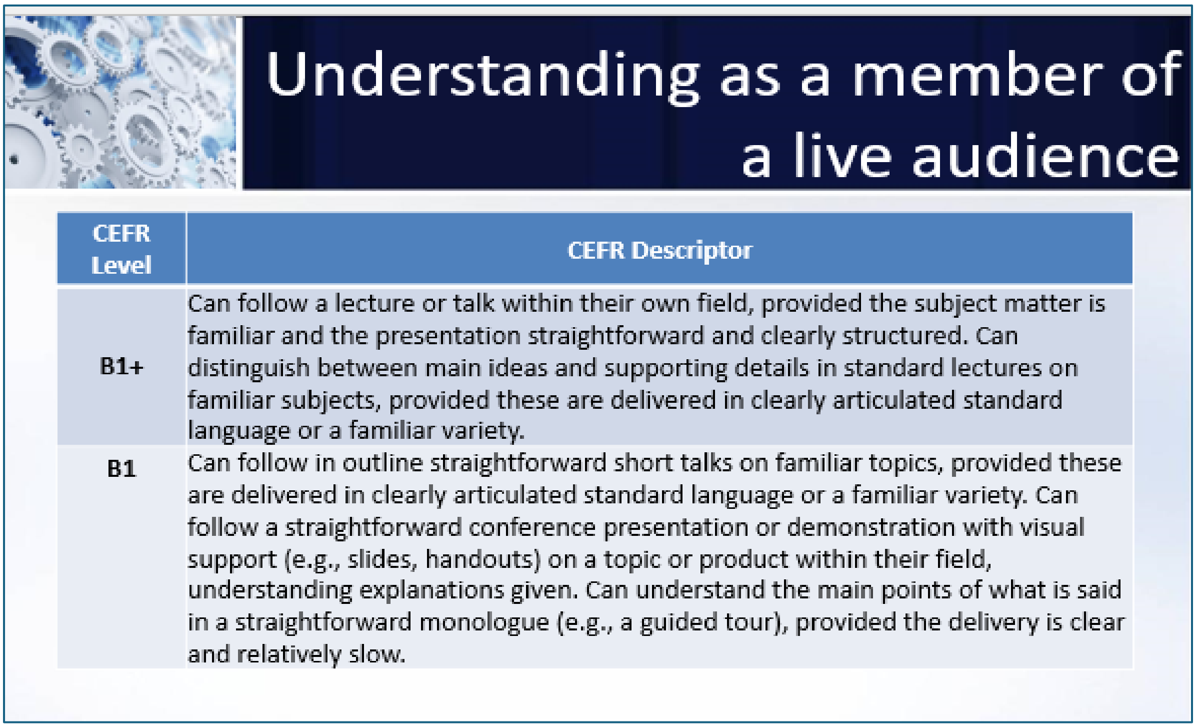
Judges participated in both pre-workshop and during-workshop activities to become familiar with the CEFR descriptors. Before the workshop, judges completed three tasks: (i) matching CEFR descriptors to their corresponding levels, (ii) identifying key features in each descriptor that signify transitions between CEFR levels, and (iii) completing a timed multistage adaptive test. Judges received PDF versions of their responses to the first two activities prior to attending the first session of the workshop, allowing them to reference their work throughout the sessions (see Appendix B for examples of pre-workshop activities). In addition, judges were given a set of coded CEFR scales for easy reference to individual descriptor codes during the workshop (see Table 2 for scales used during the workshop and Appendix C for example of a coded CEFR scale).

Table 2: CEFR scales

Reading scales	**Listening scales**
Overall reading comprehension	Overall oral comprehension
Reading correspondence	Understanding conversation between other people
Reading for orientation	Understanding as a member of a live audience
Reading for information and argument	Understanding announcements and instructions
Reading for instruction	Understanding audio media and recordings
	Watching TV, film and video

(Council of Europe, 2020)

During the familiarization stage, judges received feedback on the first pre-workshop activity, followed by a discussion on all the CEFR scales that would be used in the workshop. Based on their responses to the second pre-workshop activity, judges were asked to discuss the salient features in each descriptor that helped them distinguish the CEFR levels (see Figure 4).

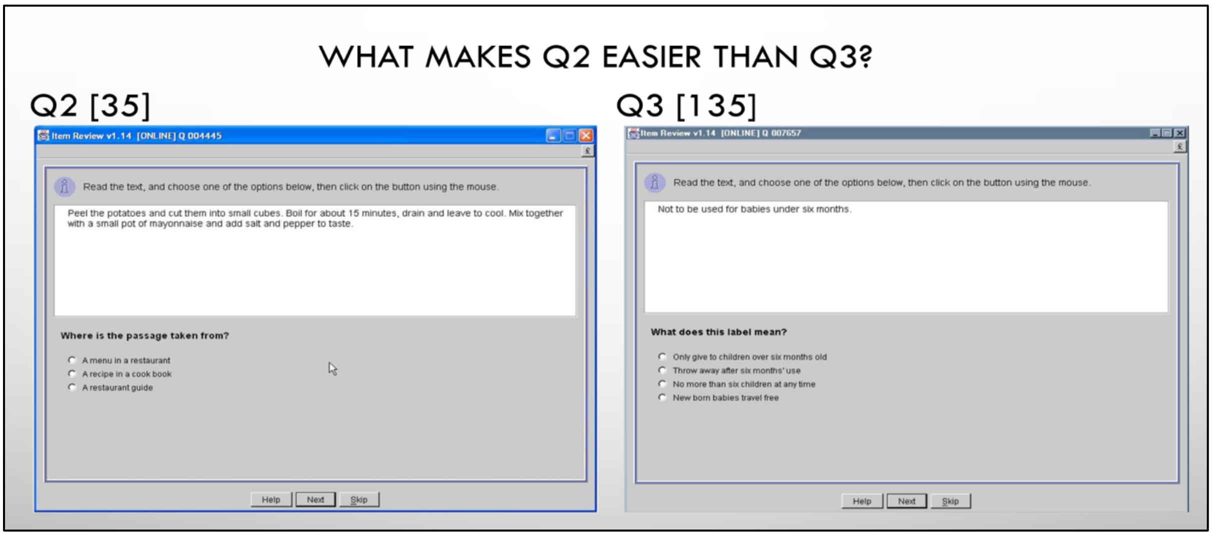


Understanding as a member of a live audience

CEFR Level	CEFR Descriptor
B1+	Can follow a lecture or talk within their own field, provided the subject matter is familiar and the presentation straightforward and clearly structured. Can distinguish between main ideas and supporting details in standard lectures on familiar subjects, provided these are delivered in clearly articulated standard language or a familiar variety.
B1	Can follow in outline straightforward short talks on familiar topics, provided these are delivered in clearly articulated standard language or a familiar variety. Can follow a straightforward conference presentation or demonstration with visual support (e.g., slides, handouts) on a topic or product within their field, understanding explanations given. Can understand the main points of what is said in a straightforward monologue (e.g., a guided tour), provided the delivery is clear and relatively slow.

Figure 4: Slide used during the familiarization stage

A CEFR-related activity was then conducted, in which judges were tasked with assigning a CEFR level to either a reading passage and item (in the Reading workshop) or a listening passage and item (in the Listening workshop). Judges were encouraged to compare each passage with the one that preceded or followed it, discussing the factors that made one passage or item more or less challenging. This activity was followed by a group discussion to reach a shared consensus on the CEFR level assigned to each passage and item (see Figure 5 for an example of a comparative passage and item activity). At the end of this stage, judges were asked to complete evaluation 1.



WHAT MAKES Q2 EASIER THAN Q3?

Q2 [35]

Read the text, and choose one of the options below, then click on the button using the mouse.

Peel the potatoes and cut them into small cubes. Boil for about 15 minutes, drain and leave to cool. Mix together with a small pot of mayonnaise and add salt and pepper to taste.

Where is the passage taken from?

○ A menu in a restaurant
○ A recipe in a cook book
○ A restaurant guide

Q3 [135]

Read the text, and choose one of the options below, then click on the button using the mouse.

Not to be used for babies under six months.

What does this label mean?

○ Only give to children over six months old
○ Throw away after six months' use
○ No more than six children at any time
○ New born babies travel free

Figure 5. Comparing reading passages and items (*adapted from DIALANG 2005*)

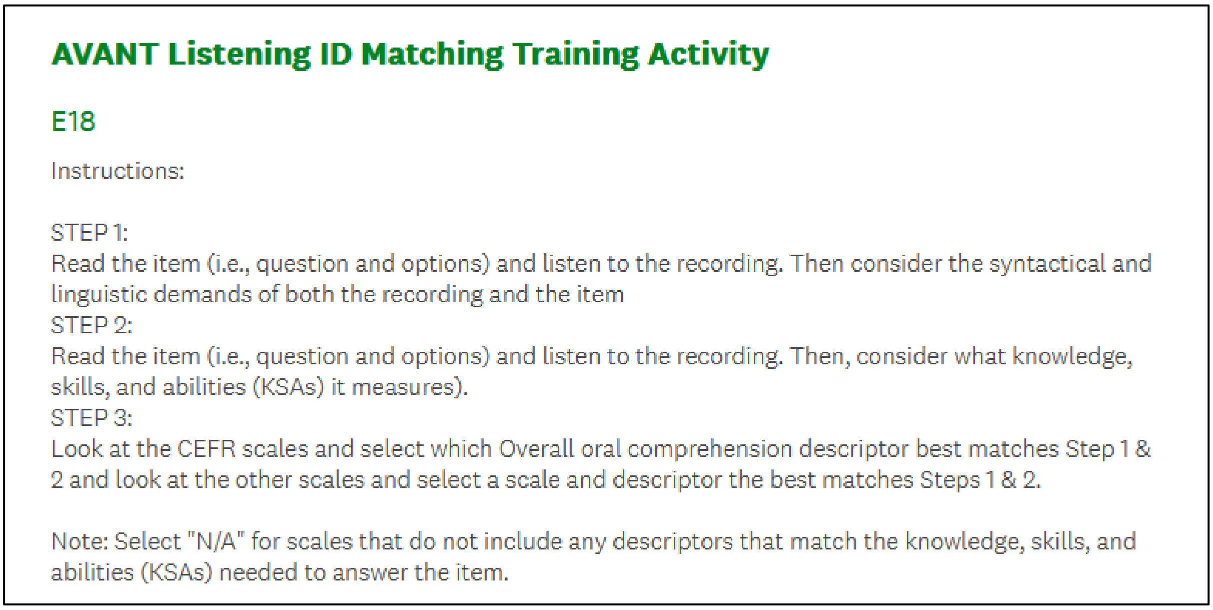
2.3.3 The training in the method stage

At this stage, judges received training in the ID Matching method by applying it to practice items. Following Kanistra (2025, forthcoming), judges were instructed to consider two key questions for each item:

(i) *Which CEFR descriptor(s) best align with the knowledge, skills, and abilities necessary for test-takers to successfully answer the item?*

(ii) *What makes this item more challenging than the one preceding it?*

Judges were also instructed to read or listen to the passage along with its corresponding item, analyzing the syntactical and linguistic demands of both. They then were asked to review the item to determine the specific knowledge, skills, and abilities (KSAs) it assessed. Finally, they were asked to refer to the CEFR scales to select the scale and descriptor that most accurately reflected the demands of the item (see Figure 6 the standard setting instructions and Figure 7 for an example of training in the method activity). A discussion ensued to ensure that judges were accurately interpreting the method, the CEFR descriptors, and the corresponding CEFR levels. Judges were also trained to interpret the *reality* information in the form of item difficulty measures assigned to each group of items and/or each item. Additional instructions were provided to help judges understand and evaluate the item difficulty measures effectively. At the end of this stage, judges were asked to complete evaluation 2.



AVANT Listening ID Matching Training Activity

E18

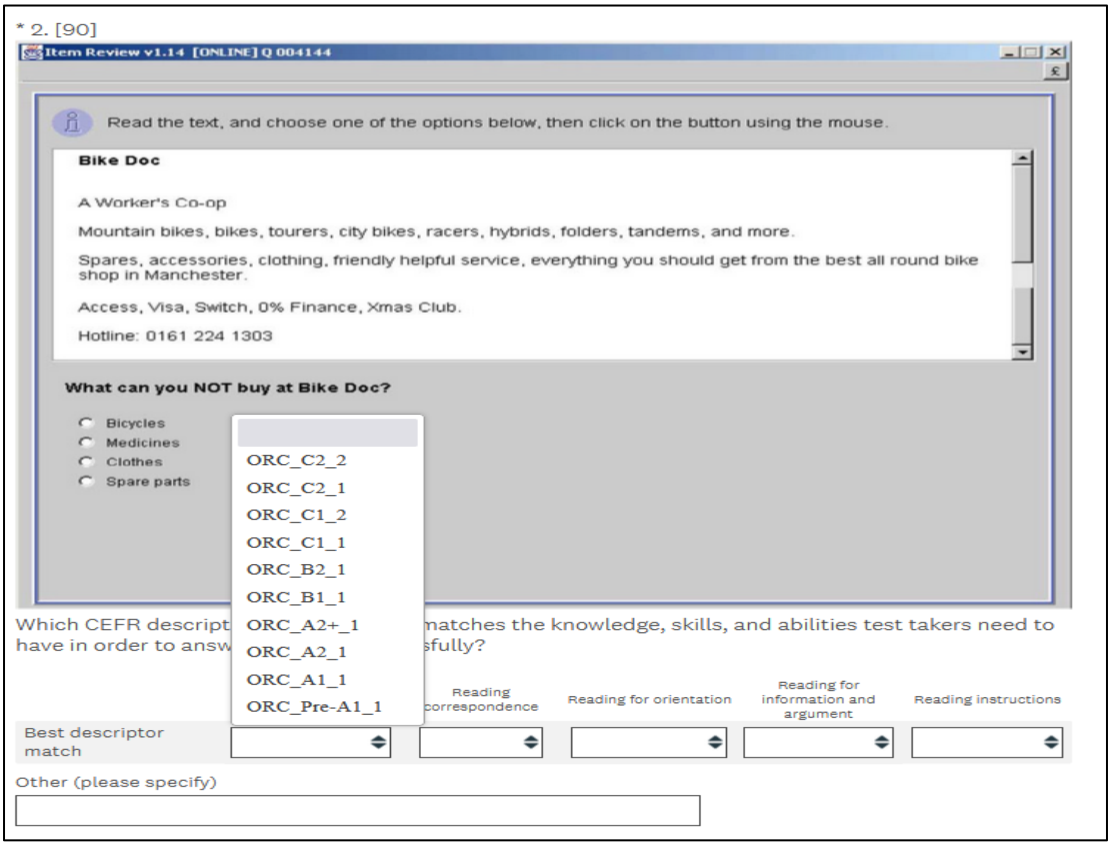
Instructions:

STEP 1:
Read the item (i.e., question and options) and listen to the recording. Then consider the syntactical and linguistic demands of both the recording and the item
STEP 2:
Read the item (i.e., question and options) and listen to the recording. Then, consider what knowledge, skills, and abilities (KSAs) it measures).
STEP 3:
Look at the CEFR scales and select which Overall oral comprehension descriptor best matches Step 1 & 2 and look at the other scales and select a scale and descriptor the best matches Steps 1 & 2.

Note: Select "N/A" for scales that do not include any descriptors that match the knowledge, skills, and abilities (KSAs) needed to answer the item.

Figure 6. Example of ID Matching method instructions

(*adapted from Kanistra, 2025, forthcoming)*



* 2. [90]

Item Review v1.14 [ONLINE] Q 004144

Read the text, and choose one of the options below, then click on the button using the mouse.

Bike Doc

A Worker's Co-op

Mountain bikes, bikes, tourers, city bikes, racers, hybrids, folders, tandems, and more.

Spares, accessories, clothing, friendly helpful service, everything you should get from the best all round bike shop in Manchester.

Access, Visa, Switch, 0% Finance, Xmas Club.

Hotline: 0161 224 1303

What can you NOT buy at Bike Doc?

○ Bicycles
○ Medicines
○ Clothes
○ Spare parts

ORC_C2_2
ORC_C2_1
ORC_C1_2
ORC_C1_1
ORC_B2_1
ORC_B1_1
ORC_A2+_1
ORC_A2_1
ORC_A1_1
ORC_Pre-A1_1

Which CEFR descript matches the knowledge, skills, and abilities test takers need to have in order to answ sfully?

		Reading correspondence	Reading for orientation	Reading for information and argument	Reading instructions
Best descriptor match					

Other (please specify)

Figure 7: Example of training task (*adapted from DIALANG, 2005)*

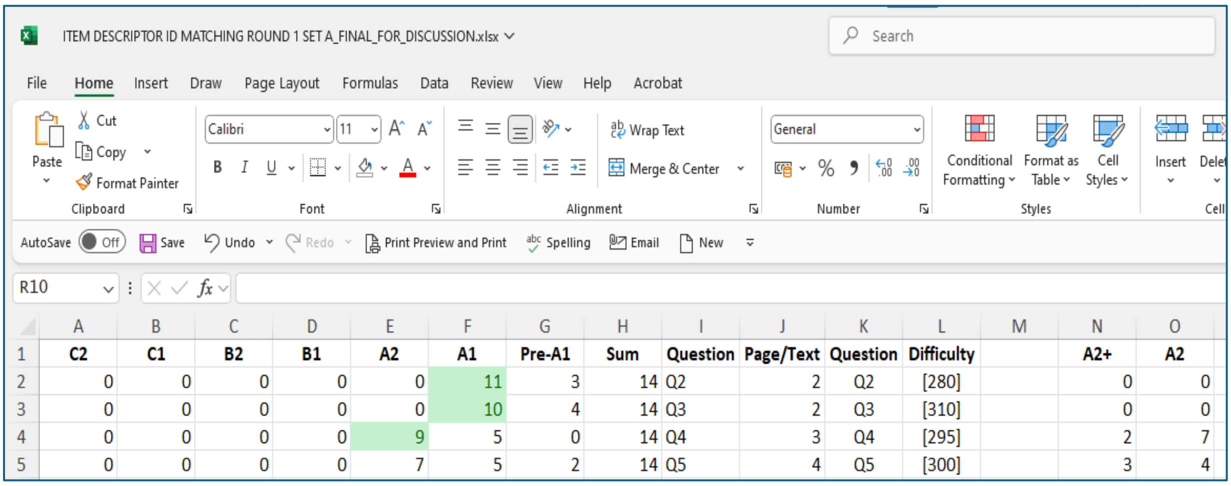
2.3.4 The item review and matching stage (Round 1)

In this stage (Round 1), judges analyzed test items, assigned appropriate CEFR levels and descriptors, and documented their reasoning for their selections. With 86 items to review per workshop, the items were divided into two sets (Set A and Set B) to ensure each set could be completed within a designated session. The items were arranged in order of difficulty in an Ordered Item Booklet (OIB), but for texts with multiple items, the average difficulty of those items determined their placement. This guaranteed that texts and their associated items were presented together. Each item included a difficulty measure, and each set of items began with an average difficulty measure for the text.

Judges reviewed the Reading or Listening items in a virtual environment, assessing each item based on its content, difficulty, and alignment with the CEFR scales and descriptors. They were tasked with matching each item to the CEFR scale and descriptor that best represented the skills needed to answer it. Judges were required to select one descriptor from the overall Reading or Listening scale and, where applicable, at least one descriptor from another relevant scale. This approach encouraged judges to record a rationale for assigning items to specific levels. After each Round 1 set (SET A and SET B), judges were sent their individual ratings for each item so that they could use them in the next stages of the workshop. At the end of this stage, judges were asked to complete evaluation 3.

2.3.5 The consensus-building discussion stage

In this stage, judges received *normative* information about their individual ratings through visual feedback (see Figure 2.8). This allowed them to compare their item alignment with that of the group (Kollias, 2023; Maurer & Alexander, 1992) before proceeding to the next stage. An in-depth discussion followed, during which judges were encouraged to articulate the reasoning behind their decisions based on the descriptors and scales chosen in Round 1. The discussion provided the judges with a deeper understanding of the test items, CEFR scales, and their corresponding descriptors. They were also prompted to reflect on the discussion and note any adjustments they planned to implement in the following stage (Round 2).


ITEM DESCRIPTOR ID MATCHING ROUND 1 SET A_FINAL_FOR_DISCUSSION.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	C2	C1	B2	B1	A2	A1	Pre-A1	Sum	Question	Page/Text	Question	Difficulty		A2+	A2
2	0	0	0	0	0	11	3	14	Q2	2	Q2	[280]		0	0
3	0	0	0	0	0	10	4	14	Q3	2	Q3	[310]		0	0
4	0	0	0	0	9	5	0	14	Q4	3	Q4	[295]		2	7
5	0	0	0	0	7	5	2	14	Q5	4	Q5	[300]		3	4

Figure 8: Example of Round 1 judge feedback

2.3.6 The final review stage (Round2)

The final stage (Round 2) required judges to review the items again and either confirm their initial alignment decisions or make final revisions based on the discussion in the previous stage. If they felt no adjustments were needed, judges could choose to keep their Round 1 ratings unchanged. At the end of this stage, they completed evaluations 4 and 5.

3. Validating the cut scores

This section outlines the evaluation process for the standard setting workshop, which was assessed based on three types of validity evidence: (1) procedural, (2) internal, and (3) external (Cizek & Earnest, 2016; Hambleton & Pitoniak 2006).

3.1 Procedural validity

Procedural validation assesses how accurately and consistently the standard setting procedures were described, applied, and adhered to. It also considers the level of confidence judges had in the process. Table 3 provides an overview of the evaluation elements related to procedural validation.

Table 3: Procedural evaluation evidence

Evaluation element	Description
Explicitness	The degree to which the standard setting purposes and processes were clearly and explicitly articulated a priori.
Practicability	The ease of implementation of the procedures and data analysis; the degree to which procedures are credible and interpretable to relevant audiences.
Implementation	The degree to which the following procedures were reasonable, and systematically and rigorously conducted; selection and training of participants, definition of the performance standard and data collection.
Feedback	The extent to which participants have confidence in the process and in the resulting cut score(s).
Documentation	The extent to which features of the study are reviewed and documented for evaluation and communication purposes.

Source: Cizek & Earnest (2016).

The initial three components of procedural validity were discussed in the previous section (Section 2: Standard Setting Methodology) of this report, while the fourth component will be examined through an analysis of the judge evaluation surveys.

Evaluation surveys (Feedback)

During each workshop, judges were given opportunities to provide feedback by completing a series of surveys. A total of five surveys were conducted per workshop at specific intervals (Cizek, 2012). In each survey, judges rated their level of agreement with specific statements using a five-point scale (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree). They were also invited to share comments or ask questions. After each survey, responses were reviewed, and comments or questions were addressed accordingly. Table 4 displays a summary of the Reading and Listening evaluation surveys (see Appendix D for individual survey evaluations).

Table 4: Summary of reading and listening evaluation surveys

Evaluation	Surveys administered	No. of statements	Section	Min. Scale score	Max. Scale score	Average scale score	N
1.	End of orientation session	9	Reading	2	5	4.47	14
			Listening	2	5	4.49	14
2.	End of training session	8	Reading	3	5	4.49	14
			Listening	3	5	4.59	14
3.	End of Round 1	7	Reading	2	5	4.46	14
			Listening	1	5	4.41	14
4.	End of Round 2	10	Reading	3	5	4.51	14
			Listening	3	5	4.65	14
5.	Final	6	Reading	2	5	4.48	14
			Listening	3	5	4.70	14

During the Reading section workshop, a total of 560 responses were collected (14 judges evaluating 40 statements each). Among these, 306 responses (54.64%) were categorized as ‘Strongly Agree’, 226 (40.36%) as ‘Agree’, 24 (4.29%) as ‘Neutral’, 4 (0.71%) as ‘Disagree’, and 0 (0.00%) as ‘Strongly Disagree’. The same judge expressed ‘Disagree’ four times regarding the pacing and timing of the orientation session and usefulness and functionality of the technologies in evaluations 3 and 5.

For the Listening section workshop, 560 responses were recorded (14 judges evaluating 40 statements each). Among these, 345 responses (61.61%) were ‘Strongly Agree’, 191 (34.11%) were ‘Agree’, 22 (3.93%) were ‘Neutral’, while 1 response each (0.18%) was recorded as ‘Disagree’ and ‘Strongly Disagree’. The ‘Strongly Disagree’ response concerned the opportunity to ask questions during Round 1, likely due to the judge forgetting the instruction to use the chat feature for questions. The single ‘Disagree’ response referred to the pacing of the orientation session.

Overall, the survey evaluations were positive, as 95.00% of the Reading workshop statements and 95.71% of the Listening workshop statements were recorded as ‘Agree’ or ‘Strongly Agree’, thus supporting procedural validity.

3.2 Internal validity

Internal validation pertains to the consistency and accuracy of the results, as well as the reliability of the recommended cut scores, ensuring they are not influenced by chance. This section focuses on evaluating the internal validity of these cut scores. The evaluation considers factors such as consistency within the method, intraparticipant (intra-judge) consistency, interparticipant (inter-judge) consistency, and the consistency and accuracy of decisions. Table 5 outlines the evaluation elements for internal validation.

Table 5: Internal evaluation elements

Evaluation element	**Description**
Consistency within method	The precision of the estimate of the cut score(s).
Intraparticipant consistency	The degree to which a judge can provide ratings that are consistent with the empirical item difficulties, and the degree to which ratings change across rounds.
Interparticipant consistency	The consistency of item ratings and cut scores across judges.
Decision consistency	The extent to which repeated application of the identified cut scores (s) would yield consistent classifications of examinees.
Other measures	The consistency of cut scores across item types, content areas and cognitive processes.

Source: Cizek & Earnest (2016).

3.2.1 Overview of analysis framework

Building on the works of Kanistra and Kollias (2024) as well as Kollias (2023), the internal evaluation of the recommended cut scores, utilizing classical test theory (CTT) and Rasch measurement theory, involved analyzing various indices associated with internal cut score assessment. Figure 9 summarizes the indices reviewed and the analyses conducted for each element of the internal evaluation.

In high-stakes settings, multiple indices are recommended to assess both intra- and inter-judge consistency, agreement, and reliability (Kaftandjieva, 2010). Therefore, this study evaluates intra- and inter-judge consistency using several indices.

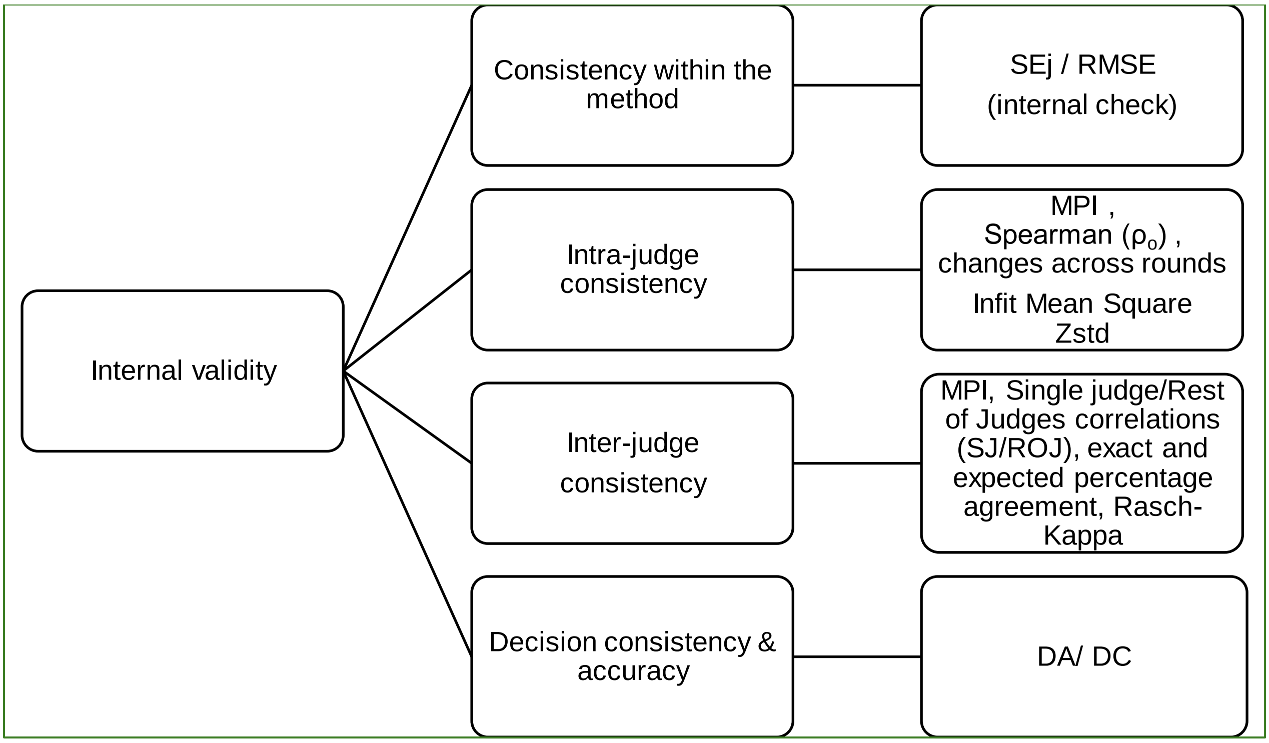

Internal validity
Consistency within the method
SEj / RMSE
(internal check)
Intra-judge consistency
MPI ,
Spearman (ρ_o) ,
changes across rounds
Infit Mean Square Zstd
Inter-judge consistency
MPI, Single judge/Rest of Judges correlations (SJ/ROJ), exact and expected percentage agreement, Rasch-Kappa
Decision consistency & accuracy
DA/ DC

Figure 9: Analysis framework

Source: adapted from Kanistra and Kollias (2024) and Kollias (2023)

Judge CEFR ratings were coded from 0.5 (Pre-A1) to 6 (C2) (see Table 6) to facilitate quantitative analyses. The plus levels (i.e., A1+, A2+, B1+, B2+) assigned by judges were quantified as an average of the two adjacent scores. For example, an A2 judgment was coded as 2, and a B1 as 3; thus, A2+ was coded as 2.5, and so on.

Table 6: Coding CEFR level judgments to numeric values

CEFR Level judgment	**Assigned numeric Value**	**CEFR Level judgment**	**Assigned numeric Value**
Pre-A1	0.5	B1+	3.5
A1	1	B2	4
A1+	1.5	B2+	4.5
A2	2	C1	5
A2+	2.5	C2	6
B1	3		

3.2.2 Classical test theory

This section describes the indices used to assess the internal consistency of the standard setting study within the classical test theory (CTT) framework.

Consistency within the method

Consistency within the method indicates how closely the cut score would align if the standard setting process were repeated (Cizek & Earnest, 2016; Hambleton & Pitoniak, 2016). To evaluate this consistency, an internal check can be performed by comparing the standard error of measurement (SEM) of the test instrument with the standard error of judgment (SEj). Guidelines for this comparison suggest that the SEj should not exceed half of the SEM (Cohen, Kane, & Crooks, 1999). Following Kollias (2023), the root mean square error (RMSE) was used instead of the SEM as cut scores were expressed in logits. The SEj is calculated using the following formula from Equation 1:

$$SEj \ = \frac{SDj}{\sqrt{(n-1)}} \qquad (1)$$

where,

SDj = standard deviation (i.e., population) of the judges' recommended cut cores

n = number of judges

The cut scores were further evaluated using the conditional standard error of measurement (cSEM), which represents the SEM at the specific cut score point on the logit scale (Sireci et al., 2008). This metric provides an estimate of the precision of the cut score location within the assessment's scoring framework.

Additionally, the accuracy of the cut score location was assessed using the conditional reliability (cReliability) of the recommended cut score. The cReliability was calculated using Equation 2 (Nicewander, 2019), where I(X,θ) is the score information function, derived from the test characteristic curve file (TCCFILE) generated by the software program Winsteps (version 5.8.3, Linacre, 2024c):

$$\rho(X, X'|\theta = \frac{I(X,\theta)}{1+I(X,\theta)} \qquad (2)$$

For foreign language proficiency tests, scores are generally considered acceptable when the cReliability values fall within the range of 0.80 to 0.90 (Nicewander, 2018, 2019). Accordingly, a cut score is deemed appropriate if its cReliability value lies within this recommended range. This evaluation ensures that the cut score is both accurate and reliable, reinforcing the validity of the classification decisions derived from the assessment.

The Misplacement Index (MPI)

The Misplacement Index (MPI) was developed for ordinal scales. It provides both an overall consistency measure and individual consistency scores for each judge across items. This analysis thoroughly examines factors influencing judges' agreement and consistency, including possible item-specific idiosyncrasies (Kaftandjieva, 2010). Equation 3 illustrates how the MPI is calculated for a specific judge.

$$MPI = 1 - \frac{\sum_{i=1}^{N} w_i}{\sum_{j=1}^{k} nj(N-nj)} \qquad (3)$$

Where:
N is the total number of items
k is the number of levels of competence
nj is the number of items at level j
w is the number of discrepancies for an item

The MPI ranges from 0 to 1, with a maximum value of 1 indicating perfect agreement between a judge's ranking and the descriptors' CEFR levels. In other words, when a judge consistently assigns higher CEFR values to descriptors of higher CEFR levels, the MPI approaches 1. Conversely, inconsistent rankings result in values closer to 0 (Kaftandjieva, 2010). Kaftandjieva (ibid.) further suggested that in CEFR benchmarking and alignment studies, an MPI value exceeding 0.70 should be expected for each judge.

3.2.3 Rasch measurement theory (RMT)

The reliability, consistency, and agreement among judges in this standard setting study were also evaluated using the Rasch measurement theory (RMT) framework. Linacre (1989, 1994) emphasized that the Many-Facet Rasch Measurement (MFRM) model was developed to address subjectivity in rater-mediated assessments and account for the impact of lenient or strict raters in high-stakes testing contexts. Building on the original Rasch model (Rasch, 1960/1980) for binary-scored items, the MFRM model expands its application by accommodating varying levels of achievement and awarding partial credit for intermediate performance levels. Specifically designed for rater-mediated assessments, the MFRM model calculates the probability of success on items by comparing item difficulty and test-taker ability on a unified logit scale, effectively isolating test-taker ability from biases introduced by individual raters.

Engelhard (2009) defined the MFRM model that operationalizes the conceptual model of standard setting and benchmarking studies as follows (see Equation 4):

$$ln\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = \theta_n - \delta_i - \omega_j - \tau_k \qquad (4)$$

where:

P_{nijk} is the probability of judge n giving a rating of k on an item i for performance standard j,

P_{nijk-1} is the probability of judge n giving a rating of $k-1$ on an item i for performance standard j,

β_n judgment of minimal competence required to pass for judge n ,

δ_i judgment of difficulty for an item i,

ω_j judgment of performance standard for round j, and

τ_k judged threshold of rating category k relative to category $k-1$

(Engelhard, 2009, p. 314)

The MFRM model has been widely used in the field of language testing to support tasks such as item bank calibration (Wolfe, 2004), research on rater behavior and rating scale (Bond et al., 2021; Eckes, 2015; Engelhard, 2013; Engelhard & Wind, 2018; Lestari & Brunfaut, 2023; Myford & Wolfe, 2004a, 2004b), and validation studies (Harsh et al, 2024; Wolfe & Everett V. Smith, 2007a; 2007b). In benchmarking and alignment studies, Rasch models are employed to evaluate the consistency of ratings across rounds, assess the alignment of judges' CEFR judgments (Eckes, 2009; Harsch & Kanistra, 2020; Kanistra & Kollias, 2024; Kollias, 2023), and investigate the relationship between judges' ratings and item difficulties (Harsch & Hartig, 2015).

The Reference Supplement H (Eckes, 2009) to the Council of Europe's Manual (2009) identifies MFRM as one of the most effective models for analyzing inter-judge and intra-judge consistency (Kaftandjieva & Takala, 2000, as cited in Kaftandjieva, Standard Setting, 2004). This effectiveness can be attributed to RMT, which facilitates the assessment of both intra-judge and inter-judge consistency at individual and group levels (Eckes, 2015; Kanistra & Kollias, 2024; Kollias, 2023; Linacre, 2024b; Myford & Wolfe, 2004a, 2004b). Specifically, inter and intra-judge consistency was evaluated using the following indices, which are discussed further in the results section:

- Judge severity measure
- Fair average of the most lenient judge (min) and most severe judge (max)
- Single-judge versus rest-of-judges (SJ/ROJ) point-measure correlation
- coefficients
- Observed percentage exact agreement and expected percentage
- agreement between judges
- Rasch-Kappa
- Infit Mean-square and Infit z-standardized (Infit Zstd)

3.3 External validity

External validation pertains to the extent to which cut scores are (i) applicable across various contexts and (ii) consistent when determined using an alternative standard setting method. It also encompasses the effects or implications these cut scores have on the overall learner population or specific learner subgroups. Table 7 outlines the components used to assess external validation.

Table 7: External evaluation elements

Evaluation element	**Description**
Comparisons to other standard setting methods	The agreement of cut scores across replications using other standard setting methods
Comparisons to other sources of information	The relationship between decisions made using the test to other relevant criteria (e.g., grades, performance on tests measuring similar constructs
Reasonableness of cut scores	The extent to which cut score recommendations are feasible or realistic (including pass/ fail rates and differential impact on relevant subgroups

Source: Cizek & Earnest (2016).

As the cut score workshops took place prior to the Avant STAMP for CEFR English proficiency test being made public, no external validation evidence was available at the time of this report. Thus, the elements outlined in the table above will be addressed in future validation studies.

4. Results

The results section presents the results from the analyses of the data gathered from the standard setting studies conducted for the Reading and Listening sections of the Avant STAMP for CEFR English proficiency test. This section highlights the outcomes of the pre-workshop familiarisation task and during-workshop tasks (Round 1 and Round 2), including the consistency of the judges' item ratings, the statistical validation of cut scores, and the consistency of judgments.

4.1 Familiarization pre-workshop task

To evaluate intra-judge consistency between the judges' ranking of CEFR descriptors and their actual levels in the assigned pre-work task 1 activity (refer to section 2.3.1 for details), a Misplacement Index (MPI) analysis was performed. The results of this activity are summarized in Table 8.

Table 8: Pre-workshop task MPI indices

	Reading	Listening
Minimum	0.96	0.81
Maximum	1.00	1.00
Mean	0.99	0.96

The MPI index confirmed that the panelists had a very good understanding of the CEFR levels and descriptors, as no panelist had an MPI value below the critical threshold of 0.70. All judges exhibited strong alignment, with MPI values ranging from 0.96 to 1.00 and from 0.81 to 1.00 in the Reading and Listening tasks, respectively. These results are highly desirable as they indicate the judges' preparedness to undertake the judgment tasks (i.e., Round 1 and Round 2) for the Reading and Listening sections.

4.2 Outlier investigation

In this study, judges were classified as outliers if their Round 2 fair average measure fell outside the 95% confidence interval (mean ± 1.96 × S.D.) of the overall judge mean. For instance, if the mean measure was 0.00 with a standard deviation (S.D.) of 1, any judge with a measure exceeding ±1.96 would be considered an outlier. In the Reading section, Judge 11 was identified as an outlier, while in the Listening section, Judges 4 and 14 were classified as outliers. Consequently, these judges were excluded from further analysis.

The following internal validation analyses is based on 13 judges for the Reading section and 12 judges for the Listening section.

4.3 Consistency within the method

In these workshops, the consistency within the method was evaluated using the guidelines from Cohen et al. (1999), who recommend that the SEj to RMSE (SEj/RMSE) ratio should not exceed 0.50. When this criterion is met, the likelihood of misclassification errors is significantly reduced (Cohen et al., 1999). The MFRM measures were derived from FACETS software (version 4.21.1, Linacre 2024a).

Table 9 summarizes the internal check for method consistency: the first row lists the SEj values for each round, the second row displays the RMSE of each item bank, and the third row presents the calculated SEj/RMSE ratios.

Table 9: Internal consistency check on Round 1 and Round 2 cut scores

Index	**Reading**		**Listening**	
	Round 1	**Round 2**	**Round 1**	**Round 2**
SEj	0.27	0.25	0.75	0.24
RMSE	0.64	0.64	0.64	0.64
SEj / RMSE	0.42	0.40	1.17	0.38

By the end of Round 2, the calculated ratio satisfied the internal check criterion (≤ 0.50) , adding internal validity to the cut scores.

4.4 Judge severity

The severity measures show how judges rated the Reading and Listening items. Table 10 provides a summary of the judges' severity measures during Round 1 and Round 2 judgments. The first column provides the measurement context (i.e., judge severity measures and precision of such measures) and the exact index reported, while columns two to five show the judges' values for each index.

Judges who assigned overall lower CEFR levels are associated with positive logit values (and thus more severe) whereas judges who assigned higher CEFR levels are associated with negative logit values (and thus more lenient). The fair average indicates the raw score that the Rasch model expected the judge to assign to the items they rated if severity or leniency were absent in their judgments. Therefore, it is possible to explore the degree to which each judge impacted the average measure of each item by examining (i) each judge's severity measure and (ii) the difference in the fair average between the most lenient and the most severe judge (see Appendix F for individual judge severity and precision measures). The *fair average* is a crucial concept in MFRM. Unlike raw scores or simple averages, which can be distorted by rater bias, fair averages adjust for the estimated severity or leniency of individual raters.

Table 10: Summary of judge severity and precision of measures

Index	**Reading**		**Listening**	
Inter-judge consistency	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Average measure (S.E.)	-1.86 (0.21)	0.34 (0.25)	2.16 (0.27)	0.83 (0.26)
Population S.D.	0.89	0.84	2.36	0.77
Measure min. (Model S.E.)	-3.22 (0.22)	-1.24 (0.25)	0.75 (0.26)	-0.61 (0.26)
Measure max. (Model S.E.)	-0.30 (0.21)	1.71 (0.25)	5.13 (0.29)	2.58 (0.27)
Fair average (min)	2.91	3.17	2.96	3.06
Fair average (max)	3.84	3.80	4.92	3.65

Upon reviewing the mean severity and precision of judge ratings for the Reading section in Table 10, it is evident that judges displayed a negative mean measure (-1.86) in Round 1. This indicates that judges tended to assign higher CEFR ratings to most Reading items during the initial evaluation. Following a discussion between rounds, some judges adjusted their ratings downward for certain items, resulting in a mean measure closer to neutral (0.34). The precision of ratings remained consistently high across both rounds, as demonstrated by the small standard errors (S.E. = 0.21 in Round 1; S.E. = 0.25 in Round 2).

A closer analysis of judge behavior revealed that the range between the most severe and the most lenient judge was similar across rounds as it was 2.92 logits [ranging from -3.22 (min) logits to -0.30 (max) logits] in Round 1 and 2.95 logits [ranging from -1.24 (min) logits to 1.71 logits (max)] in Round 2. However, when examining the difference in the fair average ranges across rounds, a reduction in variability was observed. In the Reading section, the range between *min* and *max*) dropped 0.93 raw score points in Round 1 to 0.63 in Round 2. These findings suggest that the discussion between rounds facilitated greater alignment among judges, thereby enhancing the internal validity of the Reading standard setting process.

The 0.63 raw score point difference reflects that the most lenient judge assigned ratings approximately half a CEFR level higher than the most severe judge. While such differences are small and quite common in human rating contexts, the application of the MFRM model effectively mitigated the influence of individual judge biases on the final CEFR item ratings as rater biasness is eliminated.

A similar trend was observed for the Listening section. Initially, the judges displayed a positive mean measure (2.16) in Round 1, indicating that they generally assigned lower CEFR ratings to the Listening items. After the subsequent discussion, some judges adjusted their ratings upward for certain items, resulting in a reduced mean measure (0.83). The precision of these ratings remained high across both rounds, as reflected by the small standard errors (S.E. = 0.27 in Round 1; S.E. = 0.26 in Round 2).

An analysis of judge behavior showed that the range between the most severe and the most lenient judge narrowed from 4.38 logits in Round 1 to 3.19 logits in Round 2. This reduction in variability translated into a decrease in the impact on raw scores, dropping from 1.96 raw score points in Round 1 to 0.59 in Round 2. This indicates that the discussions between rounds facilitated a greater alignment among the judges, enhancing the internal validity of the standard setting workshop for the listening items.

The remaining 0.59 raw score point difference suggests that the most lenient judge rated items approximately half a CEFR level higher than the most severe judge. While such differences are typical in human rating scenarios, the use of the MFRM model minimized the impact of individual judge biases on the final CEFR item ratings.

4.5 Inter-judge consistency

Inter-judge consistency, or interparticipant consistency, refers to the degree to which item ratings are consistent among judges (Cizek & Earnest, 2016; Hambleton & Pitoniak, 2006). Tables 11 and 12 presents four indices of inter-judge consistency derived from the MFRM analysis. The first column provides the measurement context, and the inter-judge consistency index reported, while columns two and three show the reported values for each index.

Table 11: Inter-judge consistency indices for the Reading section

	Reading	
Inter-judge consistency	**Round 1**	**Round 2**
Overall SJ/ROJ	0.93	0.95
SJ/ROJ observed-(expected) minimum	0.90 (0.92)	0.91 (0.94)
SJ/ROJ observed-(expected) maximum	0.97 (0.92)	0.97 (0.94)
Overall Rasch-Kappa	-0.05	-0.05
Rasch-Kappa minimum	-0.12	-0.14
Rasch-Kappa maximum	0.07	0.14

Table 12: Inter-judge consistency indices for the Listening section

	Listening	
Inter-judge consistency	**Round 1**	**Round 2**
Overall SJ/ROJ	0.89	0.94
SJ/ROJ observed-(expected) minimum	0.78 (0.90)	0.91 (0.92)
SJ/ROJ observed-(expected) maximum	0.93 (0.90)	0.97 (0.93)
Overall Rasch-Kappa	-0.02	-0.06
Rasch-Kappa minimum	-0.11	-0.24
Rasch-Kappa maximum	0.03	0.11

Inter-judge consistency was evaluated using the single-judge versus rest-of-judges (SJ/ROJ) point-measure correlations (see Appendix E for individual judge measures). This metric, similar to the Pearson product-moment correlation, assesses how closely a single judge's ranking of item ratings aligns with the overall ranking of other judges for the same items (Myford & Wolfe, 2004a; Linacre, 2024b). A positive SJ/ROJ value indicates consistency among judges in their item rankings. For rating scales with multiple categories, SJ/ROJ correlations below 0.30 are considered low, while those above 0.70 are viewed as high. A near zero or negative SJ/ROJ correlation indicates substantial divergence between a judge's rankings and those of their peers. The FACETS software (Linacre, 2024a) enhances this analysis by providing expected SJ/ROJ correlation values based on the Rasch model, which serve as benchmarks. Observed SJ/ROJ values that closely align with these expected values confirm strong inter-judge consistency.

In this study, inter-judge consistency was high for both the Reading (see Table 11) and Listening (see Table 12) sections across the two rounds. The overall SJ/ROJ correlation values were 0.93 in Round 1 and 0.95 in Round 2 for the Reading section, and 0.89 in Round 1 and 0.94 in Round 2 for the Listening section. These findings indicate that judges consistently ranked items across both sections. Furthermore, an examination of individual SJ/ROJ values further supports this consistency. All judges demonstrated highly correlated rankings, with all SJ/ROJ values exceeding the critical threshold of 0.70. Even the judge with the lowest SJ/ROJ value recorded a strong correlation of 0.78 in Round 1 of the Listening section.

These results offer strong evidence of internal validity for the standard setting workshops held for the Reading and Listening sections. They indicate that judges reached a common understanding of the CEFR levels and descriptors, ensuring the reliability of their item ratings.

Inter-judge agreement was evaluated using the exact observed percentage (%) agreement and exact expected percentage (%) agreement at both individual and group levels, calculated by FACETS. These indices offer insights into the extent to which judges are aligned in their CEFR evaluations. At the individual level, the exact observed percentage agreement indicates the proportion of instances where a judge’s CEFR ratings match that of the other judges. Conversely, the exact expected percentage agreement represents the proportion of agreement that would be expected if judges’ ratings were perfectly aligned with the predictions of the Rasch model.

The observed percentage (%) agreement can be expected to slightly exceed the expected percentage (%) agreement for trained judges. This is because judges undergo rigorous training to develop a shared understanding of the evaluation criteria, such as the CEFR descriptors. When the observed and expected agreement percentages are closely aligned, it indicates that judges are functioning as independent experts in their evaluations, which is an ideal outcome for this study. Conversely, lower-than-expected observed agreement percentages may indicate insufficient training or inconsistencies in understanding the evaluation criteria.

In benchmarking and alignment exercises, adequately trained judges can be expected to produce observed agreement percentages that slightly exceed the expected ones (Kanistra, 2025, forthcoming; Kanistra & Kollias, 2024). However, within the RMT framework, excessively high observed agreement percentages – especially those surpassing 90% or significantly exceeding the expected agreement – can be problematic. Such high alignment may suggest that judges feel pressured to agree with one another or are restricted by overly rigid guidelines, which effectively diminish their autonomy and expertise (Linacre, 2024b). In these situations, judges risk operating as mechanical scorers instead of independent evaluators, thereby undermining the validity of the evaluation process. This balance between fostering agreement and independence emphasizes the significance of thoughtfully designed training that promotes both consistency and the maintenance of expert judgment.

In this study, the observed percentage agreement closely aligned with the expected values, with none of the observed agreements exceeding the critical threshold of 90%. These results indicate a desirable balance between inter-judge consistency and independent expert judgment. Tables 13 and 14 present the inter-judge agreement index derived from the MFRM analysis. The first column provides the measurement context, and the inter-judge agreement index reported, while columns two and three show the reported values for each index.

For the Reading section (see Table 13), the overall observed agreement in Round 1 was 46.7%, aligning closely with the expected agreement of 49.2%. In Round 2, the observed agreement increased to 57.5%, remaining well aligned with the expected 59.4%. An examination of individual observed and expected agreement values further underscores the consistency of the judges' CEFR item judgments. Even at their minimum, the range of agreement values across judge remained within acceptable limits and aligned with expected patterns.

Table 13: Inter-judge agreement indices for the Reading section

	Reading	
Inter-judge exact agreement	**Round 1**	**Round 2**
Overall exact observed % agreement (expected %)	46.7% (49.2%)	57.5% (59.4%)
exact observed % agreement (expected %) minimum	38.9% (45.3%)	47.1% (53.7%)
exact observed % agreement (expected %) maximum	55.0% (51.7%)	59.8% (56.5%)

The observed agreement values for Round 1 were slightly lower than the expected ones but remained closely aligned. Specifically, the minimum observed agreement was 38.9%, compared to an expected agreement of 45.3%, while the maximum observed agreement was 55.0%, slightly higher than the expected value of 51.7%. In Round 2, both the minimum and maximum values increased following the Round 1 discussion, reflecting improved alignment in item judgments. The minimum observed agreement rose to 47.1%, with an expected agreement of 53.7%, and the maximum observed agreement increased to 59.8%, slightly higher than the expected value of 56.5%.

Table 14: Inter-judge agreement indices for the Listening section

Inter-judge exact agreement	**Round 1**	**Round 2**
Overall exact observed % agreement (expected %)	48.7% (49.9%)	59.6% (61.8%)
exact observed % agreement (expected %) minimum	32.6% (30.6%)	54.0% (62.8%)
exact observed % agreement (expected %) maximum	54.3% (56.9%)	67.0% (62.8%)

Similarly, for the Listening section (see Table 14), the observed agreement in Round 1 was 48.7%, nearly matching the expected agreement of 49.9%. In Round 2, the observed agreement rose to 59.6%, remaining closely aligned with the Rasch model's expected agreement of 61.8%. When examining the minimum and maximum observed agreements, a similar trend was observed in the Listening section as that in the Reading section, namely that the agreement values among judges remained within acceptable limits and followed the expected patterns. In Round 1, the minimum observed agreement was 32.6%, slightly exceeding the expected agreement of 30.6%, while the maximum observed agreement was 54.3%, slightly lower than the expected value of 56.9%. In Round 2, the minimum observed agreement improved to 54.0%, closely aligning with the expected agreement of 62.8%, and the maximum observed agreement reached 67.0%, higher than its expected value of 62.8%.

These findings highlight the consistency and reliability of the judges' CEFR evaluations across rounds and components. The improvements observed after the Round 1 discussions demonstrate the effectiveness of the standard setting workshops in fostering alignment among judges while maintaining their independent expertise.

Such findings are highly desirable as they demonstrate that the panelists maintained their independence while applying CEFR descriptors, ensuring that their judgments were not overly influenced by group conformity. This alignment between observed and expected agreement provides strong evidence of internal consistency, further validating the internal consistency of the Reading and Listening standard setting workshops.

Inter-judge agreement was further evaluated using Rasch-Kappa, a variation of Cohen's kappa specifically adapted for Rasch measurement. A Rasch-Kappa value near zero indicates an appropriate level of agreement among judges, reflecting independent evaluations without excessive concordance. Positive Rasch-Kappa values suggest higher-than-expected agreement, whereas negative values signal divergence in CEFR judgments (Linacre, 2024b).

According to Taghvafard (cited in Linacre, 2024b), a Rasch-Kappa value within the range of -0.20 to +0.20 represents agreement consistent with the expectations of the Rasch model. Values between ±0.20 and ±0.40 suggest slightly higher or lower agreement than expected, while values of ±0.50 or greater indicate an unusually high level of agreement or disagreement. An excessive agreement may suggest that judges are operating as "rating machines," raising concerns about potential dependency or a lack of autonomy in their evaluations, which is problematic for the validity of standard setting exercises (Eckes, 2009).

While FACETS does not directly calculate Rasch-Kappa, it can be derived using the formula (Equation 5) provided by Linacre (2024b), providing additional insights into the dynamics of inter-judge agreement and independence.

$$Rasch - Kappa = \frac{(\text{Observed\%}-\text{Expected\%})}{(100-\text{Expected\%})} \tag{5}$$

The Rasch-Kappa values for most judges fell within the expected range of -0.20 to +0.20 (see Appendix G for individual Rasch-Kappa values), demonstrating a model-consistent level of agreement. In the Reading section, the values ranged from a minimum of -0.12 to a maximum of 0.07 in Round 1 and a minimum of -0.14 to a maximum of 0.14 in Round 2, with an average of -0.05 across both rounds. For the Listening section, the average Rasch-Kappa was -0.02 in Round 1 (with a minimum of -0.11 to a maximum of 0.03 in Round 1) and -0.06 in Round 2 (with a minimum of -0.24 to a maximum of 0.11). These results indicate that, apart from one judge who exhibited slightly less agreement than predicted by the Rasch model, all other judges demonstrated agreement levels consistent with model expectations.

These Rasch-Kappa values align with the exact observed percentage agreement analysis results, confirming that judges achieved an appropriate level of inter-judge agreement while maintaining their independence as evaluators. This balance underscores the credibility of the CEFR item judgments and supports the validity of the derived cut scores, strengthening the overall reliability of the standard setting process.

4.6 Intra-judge consistency

Intraparticipant consistency, also referred to as intra-judge consistency, measures how well judge ratings (i) align with the empirical difficulties of the items and (ii) vary across different rounds (Cizek & Earnest, 2016; Hambleton & Pitoniak, 2006). The alignment between empirical difficulties and judge estimates reinforces the appropriateness of the selected standard setting method and the validity of the established cut scores (Kaftandjieva, 2010).

This section evaluates intra-judge consistency within the RMT and CTT frameworks as applied in this study. The analysis focuses on how consistently judges assign CEFR ratings to items. Intra-judge consistency in the RMT framework was assessed using two key indices: the Infit Mean-square (Infit Mnsq) and the Infit z-standardized (Infit Zstd), both at the individual and group levels. In the CTT framework, intra-judge consistency was evaluated through Spearman (ρo), and MPI indices. MPI indices are also reported at the group and individual judge levels.

The results of the intra-judge consistency analysis are shown in Tables 15 and 16. As in other tables in this report, the first column describes the measurement context and the specific index being evaluated. The remaining columns present the values obtained from the internal consistency analyses for the Reading and Listening sections.

Table 15: Intra-judge consistency indices for the Reading section

	Reading	
Intra-judge consistency	**Round 1**	**Round 2**
Mean Infit Mnsq; S.D. (Zstd)(Group)	0.98; 0.19 (-0.01)	0.96; 0.24 (-0.40)
Minimum Infit Mnsq (Zstd)	0.51 (-3.80)	0.45 (-4.20)
Maximum Infit Mnsq (Zstd)	1.21 (1.20)	1.31 (1.70)
Overall MPI	0.89	0.91
MPI minimum	0.87	0.88
MPI maximum	0.92	0.92
Spearman (ρo) mean	0.96	
Spearman minimum	0.91	
Spearman maximum	0.99	

Table 16: Intra-judge consistency indices for the Listening section

	Listening	
Intra-judge consistency	**Round 1**	**Round 2**
Mean Infit Mnsq; S.D. (Zstd)(Group)	0.98; 0.19 (-0.01)	0.96; 0.24 (-0.40)
Minimum Infit Mnsq (Zstd)	0.51 (-3.80)	0.45 (-4.20)
Maximum Infit Mnsq (Zstd)	1.21 (1.20)	1.31 (1.70)
Overall MPI	0.89	0.91
MPI minimum	0.87	0.88
MPI maximum	0.92	0.92
Spearman (ρo) mean	0.96	
Spearman minimum	0.91	
Spearman maximum	0.99	

The Infit indices measure the consistency of judges' ratings relative to the expectations of the Multifaceted Rasch Measurement model. Ideally, Infit values are equal to 1, representing perfect alignment between observed judgments and model predictions. However, they can range from 0 to infinity.

- Infit values close to 1 indicate that observed judgments align well with model predictions.
- Values below 1 suggest overfit, meaning the judgments are more consistent than expected, potentially reflecting limited variability in the ratings.
- Values above 1 indicate misfit, signifying greater variability than predicted by the model. Misfit values are particularly concerning as they reflect deviations that are difficult to explain and may undermine the reliability of the ratings (Myford & Wolfe, 2004a).

Wright and Linacre (1994) recommend that acceptable Infit Mnsq values should fall between 0.40 and 1.20 in contexts where rater agreement is critical. Infit values outside the suggested range are statistically significant if they are associated with Infit Zstd values larger than ±2. Linacre (2024b) further emphasizes that lower Infit Mnsq values demonstrate strong intra-judge consistency, indicating that a judge's ratings for one item can reliably predict their ratings for other items of similar ability.

In this study, as shown in Tables 15 and 16, the mean Infit Mnsq values for the group of judges were very close to the ideal value of 1.00, ranging from 0.90 to 0.98 across rounds and sections (See Appendix F for individual judge fit statistics). **These results indicate that the judges demonstrated appropriate intra-judge consistency during the Reading and Listening standard setting workshops, thus adding internal validity to cut scores.**

The Infit Mnsq values for individual judges were also within acceptable limits for trained raters. Even the maximum values that exceeded the threshold of 1.20 were associated with Zstd values lower than ±2, suggesting that these deviations were not substantial enough, and as such, they did not affect the overall reliability of the CEFR item judgments. **These findings align with earlier evidence of internal consistency, reinforcing the credibility of the judges' evaluations.**

The MPI index further corroborated the findings from the MFRM analysis, as no panelist was associated with an MPI value below the critical threshold of 0.70. On the contrary, all judges demonstrated strong alignment, with MPI values ranging from 0.86 to 0.93 across rounds and sections (see Appendix H for individual judge MPI indices).

These results provide additional evidence of intra-judge validity, clearly demonstrating that the judges' ratings were consistent with the empirical item difficulties. This alignment reinforces the reliability and accuracy of the standard setting process and underscores the judges' ability to evaluate items in accordance with the CEFR descriptors.

Intra-judge consistency was further assessed by analyzing changes in item ratings between rounds of the Reading and Listening sections. In this context, high correlation indices were anticipated, as standard setting methods that organize items in an Ordered Item Booklet (OIB) and incorporate empirical item difficulty data into the information provided to judges typically result in a low proportion of rating changes between rounds (Smith, Davis-Becker, & O'Leary, 2014).

The Spearman correlation was used to measure the extent to which judges adjusted their ratings across rounds. According to Hambleton, Pitoniak, and Copella (2012), if judges make no changes to their ratings between rounds, it may indicate that they are not fully considering the feedback or discussion provided during the process. However, the ID Matching method employed in this study does not prompt judges to make substantial changes across rounds, as the focus is on refining judgments around the threshold regions rather than overhauling them.

The minor adjustments observed in this study, reflected by correlation indices ranging from 0.81 to 1.00 (rounded), suggest that judges carefully considered the feedback and discussions between rounds (see Appendix I for individual judge correlations). **These findings provide further evidence of intra-judge consistency and confirm that judges engaged meaningfully with the feedback to refine their item ratings while maintaining alignment with the CEFR descriptors.**

In summary, the judges exhibited high internal consistency throughout the workshops. This consistency supports the reliability of the item CEFR ratings derived from the standard setting process and confirms that they are both qualitatively and quantitatively robust representations of the targeted CEFR levels.

Table 17 displays the summary of person and item measures for both the Reading and Listening sections. The measures were retrieved from a Rasch analyses using jMetrik software (Meyer, 2018).

Table 17: Psychometric characteristics of the item banks

	Reading		**Listening**	
	Person	**Items**	**Person**	**Items**
No. of items	75		81	
No. of test takers	1245		1093	
S.D.	2.10	2.12	2.23	2.46
RMSE	0.64	0.19	0.64	0.25
Strata	4.52	11.18	3.35	9.95
Reliability	0.91	0.99	0.92	0.91

The Reading and Listening item banks demonstrated Rasch person reliability indices of 0.91 and 0.92, with corresponding person strata values of 4.52 and 4.80, respectively. These values indicate that the items were statistically capable of distinguishing approximately five distinct levels of test-taker ability. Person strata, rather than person separation, were used to represent the number of distinct test-taker levels, particularly when extremely low and high scores accurately reflect ability levels.

The item reliability for both the Reading and Listening sections were 0.99 and 0.91 respectively, exceeding the minimum threshold of 0.90. This suggests that the sample sizes of each item bank were sufficient to support their construct validity. **As a result, the psychometric properties of both item banks confirm that valid and reliable cut scores could be established.**

4.7 Decision consistency and accuracy

Decision consistency (DC) refers to the likelihood that learners would receive the same classification if they were assessed on two separate occasions (Kaftandjieva, 2010) and should be investigated and reported (AERA/APA/NCME, standard 2.16, 2014). While calculating these coefficients typically requires learners to retake the same test, such an approach is often impractical. To overcome this limitation, various methods have been developed to estimate decision consistency and decision accuracy based on a single test administration (Hanson & Brennan, 1990; Livingston & Lewis, 1995; Subkoviak, 1988). These methods calculate the probability that a learner would be classified consistently in a hypothetical second test administration (Cizek & Bunch, 2007).

This study evaluated decision consistency (DC) and accuracy (DA) using two IRT-based methods: Lee (2010) via the IRT-CLASS software (v2.0, Lee & Kolen, 2008) and Rudner (2001) via *cacIRT* software (version 1.4, Lathrop, 2015)

Tables 18 (Reading section) and 19 (Listening section) present the decision consistency and accuracy indices for each CEFR cut score. The tables display the values obtained using the Lee method, with values from the Rudner method provided in parentheses where applicable

Table 18: Decision consistency and accuracy: Reading section

Cut scores	**A2**	**B1**	**B2**	**C1**
Classification consistency (phi)	0.99 (0.99)	0.96 (0.96)	0.93 (0.93)	0.95 (0.94)
probability of misclassification	0.01 (0.01)	0.04 (0.04)	0.07 (0.07)	0.05 (0.06)
chance probability	0.89	0.64	0.50	0.66
kappa	0.88	0.89	0.86	0.85
classification accuracy	0.99 (0.99)	0.97 (0.97)	0.95 (0.95)	0.96 (0.96)
false negative error rate	0.01	0.02	0.03	0.01
false positive error rate	0.00	0.01	0.02	0.03

All classification consistency indices for the Reading cut scores ranged from 0.93 to 0.99, while classification accuracy measures ranged from 0.95 to 0.99. Both sets of indices far exceeded the recommended minimum criterion of 0.85 for certification examinations at each CEFR level (Subkoviak, 1988). Additionally, the kappa agreement values were either higher than or almost equal to the chance probability, implying that the classification of these cut scores were consistent, and as such, offered further evidence of decision consistency.

According to Subkoviak (1988), chance probability increases when cut scores are placed near the lower or upper ends of the test-taker ability range because test takers at these extremes would likely perform similarly even on tests that are not perfectly parallel. The kappa index measures the extent to which classification consistency can be attributed to the test instrument, correcting for chance agreement (Huynh, 1976, 1990; Subkoviak, 1980, 1988). Kappa values range from 0 to 1, with higher values indicating stronger classification reliability. In this study, all kappa values exceeded 0.85, indicating that test-taker classifications primarily resulted from the test instrument rather than chance.

These results highlight the robustness of the Reading cut scores and confirm their strong alignment with CEFR standards.

Table 19: Decision consistency and accuracy: Listening section

CEFR level	A2	B1	B2	C1
Classification consistency (phi)	0.98 (0.99)	0.96 (0.96)	0.93 (0.93)	0.94 (0.94)
probability of misclassification	0.02 (0.01)	0.04 (0.04)	0.07 (0.07)	0.06 (0.06)
chance probability	0.92	0.67	0.50	0.57
kappa	0.73	0.89	0.87	0.85
classification accuracy	0.98 (0.99)	0.97 (0.97)	0.95 (0.95)	0.96 (0.96)
false negative error rate	0.01	0.01	0.03	0.02
false positive error rate	0.01	0.02	0.02	0.03

Similarly, the Listening cut scores (see Table 19) showed classification consistency indices ranging from 0.93 to 0.99, while classification accuracy measures ranged from 0.95 to 0.99. Both sets of indices exceeded Subkoviak's recommended threshold of 0.85 (Subkoviak, 1988) for certification examinations across all CEFR levels. Moreover, the kappa agreement index values—except for A2—were above chance probability, indicating a high level of decision consistency.

According to Subkoviak (1988), chance probability tends to increase when cut scores are set near the lower or upper extremes of test-taker ability, as both the least and most capable participants are likely to perform similarly even on non-parallel tests. The kappa index measures the extent of classification consistency that can be attributed to the test itself, while accounting for chance agreement (Huynh, 1976, 1990; Subkoviak, 1980, 1988). Ranging from 0 to 1, higher kappa values indicate stronger reliability in classification decisions. In this instance, all kappa values, apart from A2, exceeded 0.85, suggesting that test-taker classifications were primarily driven by the test instrument rather than random factors. The A2 cut score had a high chance probability as it was at the lower end of the test-taker ability scale, thus, yielding a low kappa as expected. Nonetheless, the very high classification accuracy of and consistency of the A2 cut score supported its robustness.

Overall, these findings underscore the robustness of the Listening cut scores and affirm their alignment with CEFR standards.

4.8 Cut score accuracy and reliability

The Reading and Listening cut scores were also evaluated for their precision, accuracy, and reliability through the conditional SEM (cSEM), which shows the SEM at the specific cut score point (Sireci et al., 2008), and their conditional reliability (cReliability) (Nicewander, 2018, 2019). Tables 20 and 21 present the accuracy and reliability measures of the Reading and Listening section cut scores.

Table 20: Reading section cut scores

	Cut score measure	cSEM	cReliability
A2	-1.94	0.32	0.91
B1	0.23	0.23	0.95
B2	1.68	0.30	0.92
C1	3.31	0.50	0.80

For the Reading section (see Table 20), the conditional standard error of measurement (cSEM) for all four cut scores ranged from 0.23 to 0.50, which is notably lower than the RMSE value of 0.64 for the Reading section. Moreover, the high internal test reliability of 0.91 guarantees that any potential errors in the cut scores have only a minor impact on test-taker classifications. Additionally, the conditional reliability (cReliablity) values for all four cut scores met or exceeded the recommended minimum criterion of 0.80 for conditional reliability in language proficiency examinations (Nicewander, 2018, 2019).

Table 21: Listening section cut scores

	Cut score measure	cSEM	cReliability
A2	-2.18	0.35	0.89
B1	-0.36	0.23	0.95
B2	1.60	0.29	0.92
C1	3.18	0.51	0.80

Similarly, for the Listening section (see Table 21), the cSEM for all four cut scores ranged from 0.23 to 0.51, staying well below the RMSE of 0.64. In addition, the high internal test reliability (0.92) minimizes the impact of cut score errors on test-taker classifications. Moreover, the cReliability for each of the four cut scores met or exceeded the 0.80 threshold recommended for language proficiency examinations (Nicewander, 2018, 2019).

In summary, the evidence presented – from the classification consistency and accuracy indices to the conditional error and reliability measures – strongly supports the robustness and appropriateness of the Reading and Listening cut scores. The high internal test reliability and appropriate conditional reliability values underscore that the selected cut scores are reliable and valid, minimizing misclassification and supporting fair decision-making for both the Reading and Listening sections. Together, these findings confirm the alignment of the Reading and Listening cut scores with recommended proficiency standards, reinforcing the fairness and validity of the assessment outcomes.

4.9 Conclusion

The outcomes confirm that the established cut scores for the Reading and Listening sections are both valid and reliable, effectively aligning with CEFR standards. The rigorous methodology, supported by the expertise of the judges and robust statistical validation, ensures that the Avant STAMP for CEFR English proficiency test serves as a dependable tool for assessing English language proficiency. These results underscore the test's value for guiding educational and professional decisions on a global scale.

References

American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement (NCME). (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59-88.

British Council, UKALTA, EALTA, ALTE. (2022). *Aligning language education with the CEFR: A Handbook.* London: European Association for Language Testing and Assessment (EALTA), the UK Association for Language Testing and Assessment (UKALTA), the British Council, and the Association for Language Testers in Europe (ALTE).

Cizek, G. J. (2012). The forms and functions of evaluations in the Standard Setting Process. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (pp. 165-178). New York: Routledge.

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

Cizek, G. J., & Earnest, D. S. (2016). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 212-237). New York: Routledge.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalised examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education, 12*(4), 343-366.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual.* Strasbourg: Language Policy Division.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume.* Strasburg: Council of Europe Publishing.

DIALANG, Council of Europe CD. (2005). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Reading and Listening Items and Tasks: Pilot Samples.

Eckes, T. (2009). Many-Facet Rasch Measurement . In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment, (Section H).* Strasbourg: Council of Europe/Language Policy Division.

Eckes, T. (2015). *Introduction to Many-facet Rasch measurement: Analysing and evaluating rater-mediated assessments* (2 Revised and updated edition ed.). Frankfurt: Peter Lang.

Engelhard, G. (2009). Evaluating the judgements of standard-setting panellists using Rasch measurement theory. In J. Everett V. Smith, & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 312-346). Maple Grove: JAM Press.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* New York: Routledge.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments.* New York: Routledge.

Ferrara, S., & Lewis, D. M. (2012). The Item-Descriptor (ID) Matching method. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (pp. 255-282). New York: Routledge.

Hambleton, R. K., & Eignor, D. R. (1979). *A practitioner's guide to criterion-referenced test development, validation, and test score usage (Report No. 70).* Washington, DC: Prepared for the National Institute of Education and Department of Health, Education, and Welfare. Retrieved from http://files.eric.ed.gov/fulltext/ED249269.pdf

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport: Praeger Publishers.

Hambleton, R. K., Pitoniak, M. J., & Coppella, J. M. (2012). Essential steps in setting performance standards. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 47-76). New York: Routledge.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*(4), 345-359. doi:10.1111/j.1745-3984.1990.tb00753.x

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly, 12*(4), 333-362. doi:10.1080/15434303.2015.1092545

Harsch, C., & Kanistra, P. V. (2020). Using an innovative standard-setting approach to align integrated and independent writing tasks to the CEFR. *Language Assessment Quarterly*, 262-281.

Huynh, H. (1976). On the reliability of the decisions in domain-referenced testing. *Journal of Educational Measurement, 13*(4), 253-264.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. (A. E. Association, Ed.) *Journal of Educational Statistics, 15*(4), 353-368.

Kaftandjieva, F. (2004). Standard Setting. In S. Takala (Ed.), *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (Section B).* Strasbourg: Council of Europe/Language Policy Division.

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL.* Arnhem: Cito. Retrieved from http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf

Kanistra, P. (2025, forthcoming). *Evaluating the Item Descriptor (ID) Matching Method in a Face-to-Face and a Synchronous Virtual Environment.* Berlin: Peter Lang.

Kanistra, P., & Kollias, C. (2024). *Aligning the ICLE 500 Written Scripts to the CEFR: The Technical Report*. Retrieved from https://dataverse.uclouvain.be/file.xhtml?fileId=25720&version=1.2

Kollias, C. (2023). *Virtual standard setting: Setting cut scores.* Frankfurt, Peter Lang.

Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R package ccIRT. *Practical assessment research & evaluation, 20*(18), 1-5. doi:10.7275/43vm-p442

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using Item Response Theory. *Journal of Educational Measurement, 47*(1).

Lee, W. C., & Kolen, M. J. (2008). IRT-CLASS: IRT classification consistency and accuracy v2.0. University of Iowa.

Linacre, J. M. (1989, 1994). *Many-facet Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (2024a). A user's guide to FACETS Rasch-model computer programs (Program manual 4.21.1). Retrieved from http://www.winsteps.com/manuals.htm

Linacre, J. M. (2024b). Facets (Many-Facet Rasch Measurement) computer program (Version 4.21.1) [Computer software]. Retrieved from www.winsteps.com

Linacre, J. M. (2024c). Winsteps® (Version 5.8.3) [Computer Software]. Beaverton. Retrieved from www.winsteps.com

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement, 32*(2), 179-197. doi:10.1111/j.1745-3984.1995.tb00462.x

Maurer, T. J., & Alexander, R. A. (1992). Methods of improving employment test critical scores derived by judging test content: A review and critique. *Personnel Psychology, 45*(4), 727 - 762.

Meyer, P. J. (2018). jMetrik Computer programme (Version 4.1.1) [Computer software]. Retrieved from https://itemanalysis.com/jmetrik-download/

Myford, C. M., & Wolfe, E. W. (2004a). Detecting and measuring rater effects using Many-Facet Rasch measurement: Part 1. In E. V. Smith, & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460-517). Maple Grove: JAM Press.

Myford, C. M., & Wolfe, E. W. (2004b). Detecting and measuring rater effects using Many-Facet Rasch measurement. Part 2. In E. V. Smith, & R. M. Smith (Eds.), *Introduction to Rasch measurement.* Maple Grove: JAM Press.

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods, 23*(2), 351-362.

Nicewander, W. A. (2019). Conditional precision of measurement for test scores: Are conditional standard errors sufficient? *Educational and Psychological Measurement, 79*(1), 5-18. doi:10.1177/00131644187538373

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Mahwah: Lawrence Erlbaum Associates.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research and Evaluation, 7*(14). doi:10.7275/an9m-2035

Sireci, S. G., Peter Baldwin, A. M., Zenisky, A. L., Kaira, L., Lam, W., Shea, C. L., . Hambleton, R. K. (2008). *Massachusetts Adult Proficiency Tests Technical Manual.* University of Massachusetts, Centre for Educational Assessment, Amherst.

Smith, R. W., Davis-Becker, S. L., & O'Leary, L. S. (2014). Combining the best of two standard setting methods: The Ordered Item Booklet Angoff. *Journal of Applied Testing Technology, 15*(1), 18-26.

Subkoviak, M. J. (1980). Decision-consistency approaches. In *Criterion-referenced measurement: The state of the art.* (pp. 129-185). Baltimore, London: The John Hopkins University Press.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement in Education*, 47-55.

Wolfe, E. W. (2004). Equating and item banking with the Rasch model. In J. Everett V. Smith, & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 366-390). Maple Grove: JAM Press.

Wolfe, E. W., & Everett V. Smith, J. (2007a). Instrument development tools and activities for measure validation using Rasch Models: Part I - Instrument development tools. In J. Everett V. Smith, & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialised applications* (pp. 202-242). Maple Grove: JAM Press.

Wolfe, E. W., & Everett V. Smith, J. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II - Validation activities. In J. Everett V. Smith, & R. M. Smith (Eds.), *Rasch Measurement: Advanced and specialised applications* (pp. 243-290). Maple Grove: JAM Press.

Wright, B. D., & Linacre, M. J. (1994). *Reasonable mean-square fit values.* Retrieved from Rasch Measurement Transactions: https://www.rasch.org/rmt/rmt83b.htm

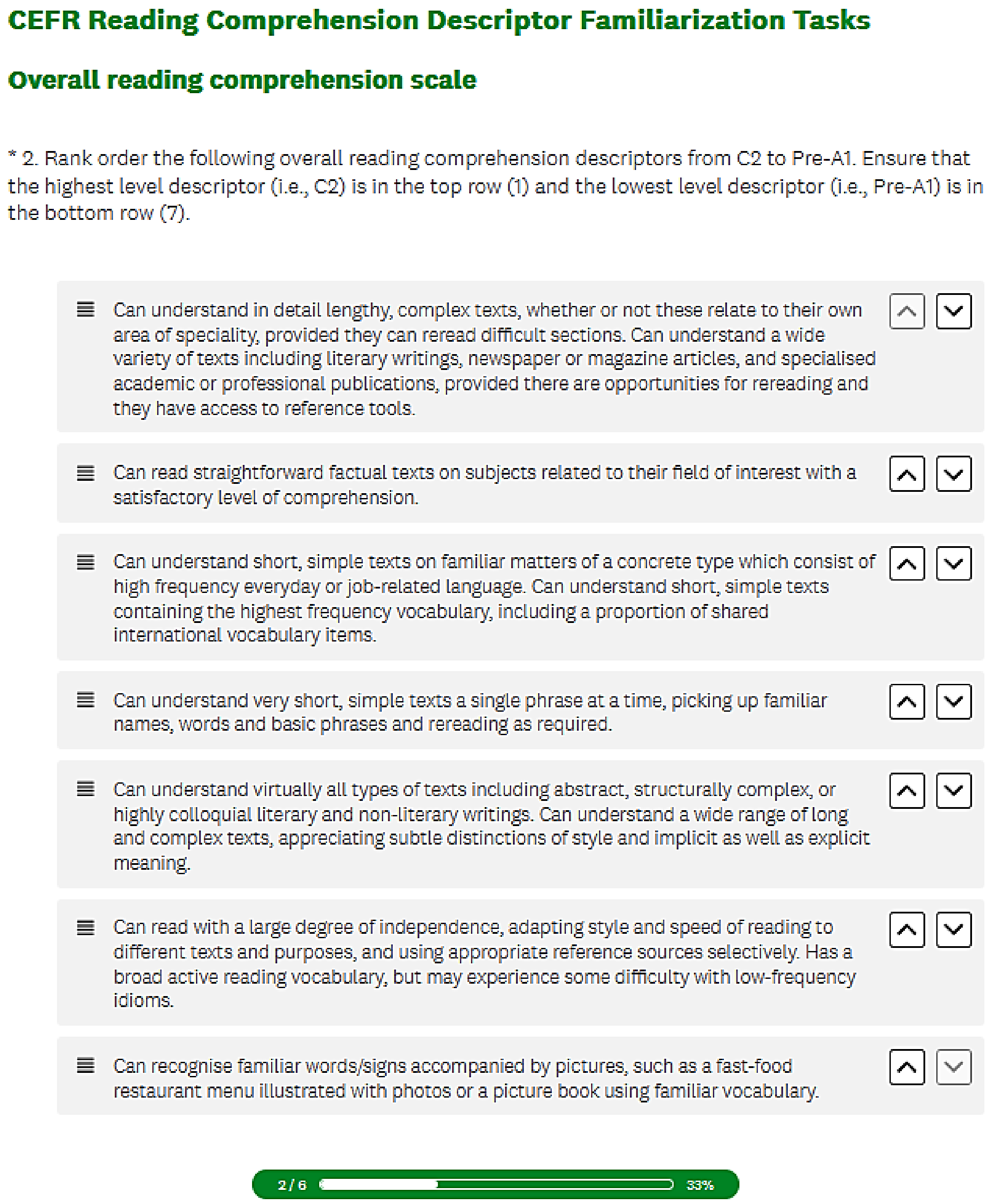
Appendices

Appendix A: Agenda

Table A1: Workshop agenda

Session	Workshop date	Description	Length
Session 1 (5.5 hours)	Reading workshop April 27th , 2024 Listening workshop May 18th , 2024	Welcoming, Introductions and overview of workshop	1 hour
		Feedback and discussion on pre-workshop descriptor matching task	1 hour
		Break	30 minutes
		Familiarisation task	2 hours
		Break	15 minutes
		Evaluation 1	15 minutes
		Training in the method	40 minutes
Session 2 (6 hours)	Reading workshop April 28th , 2024 Listening workshop May 19th , 2024	Discussion on previous day	30 minutes
		Training practice task & discussion	2 hours
		Evaluation 2	15 minutes
		Break	1 hour
		Round 1 SET A & SET B	2 hours 15 minutes
Session 3 (4 hours)	Reading workshop May 4th , 2024 Listening workshop May 25th , 2024	Round 1 SET A & SET B cont.	1.5 hours
		Evaluation 3	15 minutes
		Round 1 SET A & B discussion	1.5 hours
		Break	15 minutes
		Round 1 SET A & B discussion cont.	30 minutes
Session 4 (4 hours)	Reading workshop May 11th , 2024 Listening workshop May 26th , 2024	Round 1 SET A & B discussion cont.	1hour
		Break	15 minutes
		Round 1 SET A & B discussion cont.	1 hour 15 minutes
		Round 2	60 minutes
		Evaluations 4 & 5	20 minutes
		Closure	10 minutes

Appendix B: CEFR Pre-Workshop CEFR Tasks



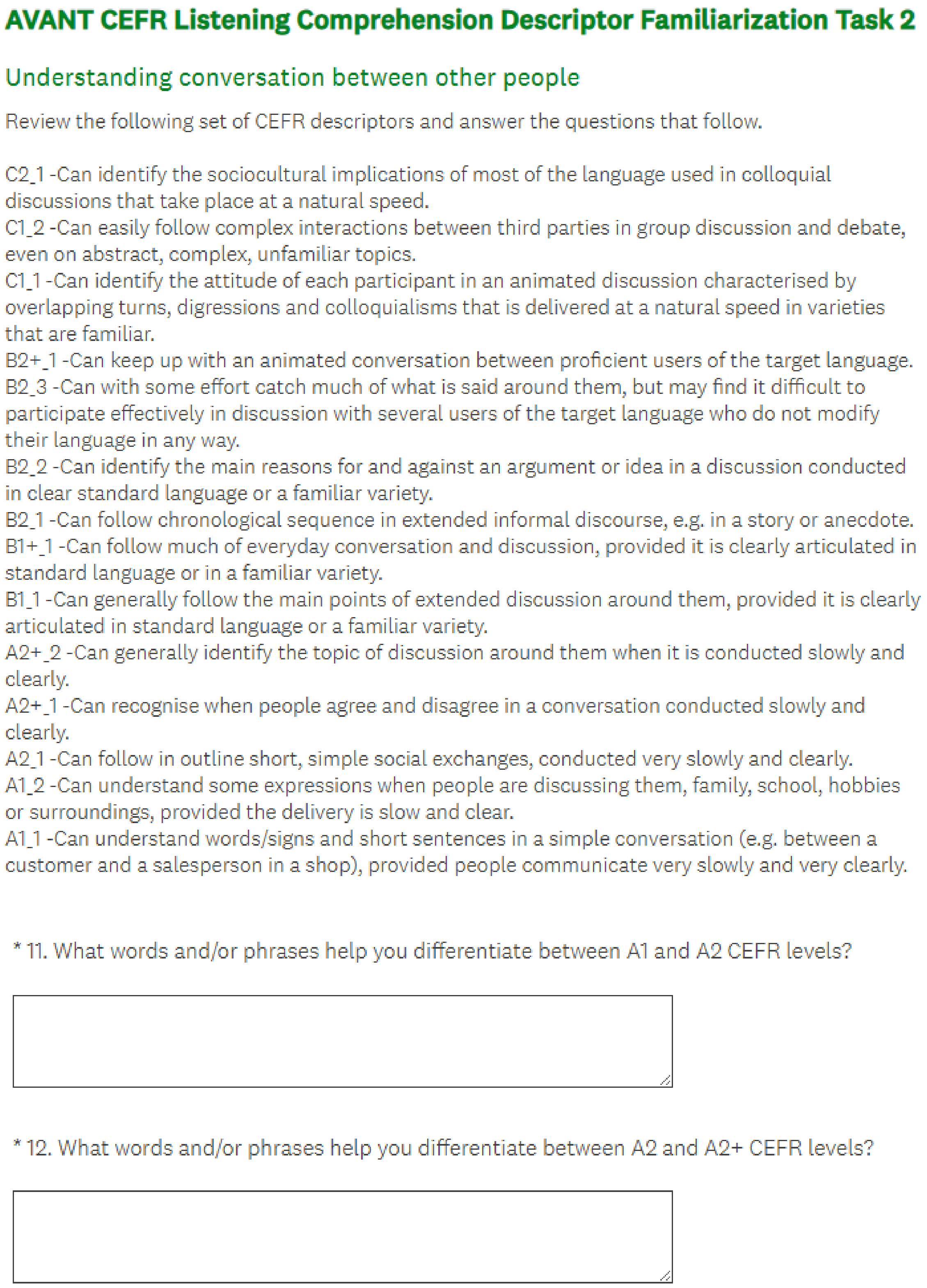
CEFR Reading Comprehension Descriptor Familiarization Tasks

Overall reading comprehension scale

* 2. Rank order the following overall reading comprehension descriptors from C2 to Pre-A1. Ensure that the highest level descriptor (i.e., C2) is in the top row (1) and the lowest level descriptor (i.e., Pre-A1) is in the bottom row (7).

- Can understand in detail lengthy, complex texts, whether or not these relate to their own area of speciality, provided they can reread difficult sections. Can understand a wide variety of texts including literary writings, newspaper or magazine articles, and specialised academic or professional publications, provided there are opportunities for rereading and they have access to reference tools.
- Can read straightforward factual texts on subjects related to their field of interest with a satisfactory level of comprehension.
- Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language. Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.
- Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.
- Can understand virtually all types of texts including abstract, structurally complex, or highly colloquial literary and non-literary writings. Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.
- 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 idioms.
- Can recognise familiar words/signs accompanied by pictures, such as a fast-food restaurant menu illustrated with photos or a picture book using familiar vocabulary.

Figure B1: Example of pre-workshop task 1 CEFR descriptor matching activity



AVANT CEFR Listening Comprehension Descriptor Familiarization Task 2

Understanding conversation between other people

Review the following set of CEFR descriptors and answer the questions that follow.

C2_1 -Can identify the sociocultural implications of most of the language used in colloquial discussions that take place at a natural speed.
C1_2 -Can easily follow complex interactions between third parties in group discussion and debate, even on abstract, complex, unfamiliar topics.
C1_1 -Can identify the attitude of each participant in an animated discussion characterised by overlapping turns, digressions and colloquialisms that is delivered at a natural speed in varieties that are familiar.
B2+_1 -Can keep up with an animated conversation between proficient users of the target language.
B2_3 -Can with some effort catch much of what is said around them, but may find it difficult to participate effectively in discussion with several users of the target language who do not modify their language in any way.
B2_2 -Can identify the main reasons for and against an argument or idea in a discussion conducted in clear standard language or a familiar variety.
B2_1 -Can follow chronological sequence in extended informal discourse, e.g. in a story or anecdote.
B1+_1 -Can follow much of everyday conversation and discussion, provided it is clearly articulated in standard language or in a familiar variety.
B1_1 -Can generally follow the main points of extended discussion around them, provided it is clearly articulated in standard language or a familiar variety.
A2+_2 -Can generally identify the topic of discussion around them when it is conducted slowly and clearly.
A2+_1 -Can recognise when people agree and disagree in a conversation conducted slowly and clearly.
A2_1 -Can follow in outline short, simple social exchanges, conducted very slowly and clearly.
A1_2 -Can understand some expressions when people are discussing them, family, school, hobbies or surroundings, provided the delivery is slow and clear.
A1_1 -Can understand words/signs and short sentences in a simple conversation (e.g. between a customer and a salesperson in a shop), provided people communicate very slowly and very clearly.

* 11. What words and/or phrases help you differentiate between A1 and A2 CEFR levels?

* 12. What words and/or phrases help you differentiate between A2 and A2+ CEFR levels?

Figure B2: Example of pre-workshop task 2 CEFR descriptor matching activity

Appendix C: Coded CEFR scale

Table C1: Example of a coded scale

OVERALL READING COMPREHENSION		
C2	**ORC_C2_2**	Can understand virtually all types of texts including abstract, structurally complex, or highly colloquial literary and non-literary writings.
C2	**ORC_C2_1**	Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.
C1	**ORC_C1_2**	Can understand in detail lengthy, complex texts, whether or not these relate to their own area of speciality, provided they can reread difficult sections.
C1	**ORC_C1_1**	Can understand a wide variety of texts including literary writings, newspaper or magazine articles, and specialised academic or professional publications, provided there are opportunities for rereading and they have access to reference tools.
B2	**ORC_B2_1**	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 idioms.
B1	**ORC_B1_1**	Can read straightforward factual texts on subjects related to their field of interest with a satisfactory level of comprehension.
A2+	**ORC_A2+_1**	Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language.
A2	**ORC_A2_1**	Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.
A1	**ORC_A1_1**	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.
Pre-A1	**ORC_Pre-A1_1**	Can recognise familiar words/signs accompanied by pictures, such as a fast-food restaurant menu illustrated with photos or a picture book using familiar vocabulary.

Appendix D: Survey evaluations

Table D1: End of orientation session (Evaluation 1) survey

Statements	Section	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
The orientation session provided a clear overview of the purpose of the standard setting for the AVANT multistage assessment.	Reading	0	0	0	6	8	14
	Listening	0	0	0	4	10	14
The orientation session answered questions I had about standard setting for the AVANT multistage assessment.	Reading	0	0	0	8	6	14
	Listening	0	0	0	6	8	14
I have a good understanding of my role in this standard setting activity.	Reading	0	0	1	6	7	14
	Listening	0	0	1	4	9	14
I have a good understanding of the CEFR Reading/Listening scales.	Reading	0	0	0	5	9	14
	Listening	0	0	0	8	6	14
I have a good understanding of the CEFR Reading/ Listening descriptors.	Reading	0	0	0	6	8	14
	Listening	0	0	1	9	4	14
Reviewing the AVANT multistage assessment content before the first online session helped me understand the standard setting task.	Reading	0	0	1	4	9	14
	Listening	0	0	0	5	9	14
Experiencing the AVANT multistage assessment online helped me understand the difficulty, content, and other aspects of the multistage assessment.	Reading	0	0	0	6	8	14
	Listening	0	0	0	6	8	14
The timing of the orientation session was appropriate.	Reading	0	1	1	6	6	14
	Listening	0	0	2	4	8	14
The pace of the orientation session was appropriate.	Reading	0	1	2	4	7	14
	Listening	0	1	1	5	7	14

Table D2: End of training session (Evaluation 2) survey

Statements	Section	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
I have a good understanding of the CEFR Reading/ Listening Scales	Reading	0	0	0	9	5	14
	Listening	0	0	0	3	11	14
I have a good understanding of the CEFR Reading/ Listening descriptors	Reading	0	0	0	8	6	14
	Listening	0	0	0	4	10	14
The training in the standard setting method was clear.	Reading	0	0	0	7	7	14
	Listening	0	0	0	4	10	14
The practice using the standard setting method helped me understand how to apply the method.	Reading	0	0	0	6	8	14
	Listening	0	0	0	7	7	14
I am comfortable with my ability to apply the standard setting method.	Reading	0	0	0	7	7	14
	Listening	0	0	0	6	8	14
I understand the feedback that will be provided to me during the standard setting process.	Reading	0	0	0	4	10	14
	Listening	0	0	2	3	9	14
The timing of the method training was appropriate.	Reading	0	0	1	6	7	14
	Listening	0	0	0	8	6	14
The pace of the training session was appropriate.	Reading	0	0	2	4	8	14
	Listening	0	0	0	7	7	14

Table D3: End of Round 1 (Evaluation 3) survey

Statements	Section	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
I understood how to complete my Round 1 ratings.	Reading	0	0	0	4	10	14
	Listening	0	0	0	2	12	14
I am confident in my Round 1 ratings.	Reading	0	0	1	12	1	14
	Listening	0	0	2	8	4	14
I had the opportunity to ask questions while working on my Round 1 ratings.	Reading	0	0	0	4	10	14
	Listening	1	0	1	4	8	14
The facilitator helped to answer questions and to ensure everyone's input was respected and valued.	Reading	0	0	0	3	11	14
	Listening	0	0	1	2	11	14
The technologies were helpful and functioned well.	Reading	0	1	3	6	4	14
	Listening	0	0	3	5	6	14
The timing of Round 1 was appropriate.	Reading	0	0	1	5	8	14
	Listening	0	0	0	9	5	14
The pace of Round 1 was appropriate.	Reading	0	0	0	6	8	14
	Listening	0	0	1	8	5	14

Table D4: End of Round 2 (evaluation 4) survey

Statements	Section	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
The instructions on how to use the item performance information (i.e., the numbers between brackets) were clear.	Reading	0	0	0	6	8	14
	Listening	0	0	0	3	11	14
The normative information (i.e., my ratings sent to me and the ratings of other judges) provided before the beginning of Round 2 was helpful.	Reading	0	0	0	5	9	14
	Listening	0	0	1	2	11	14
The instructions on how to use the normative information were clear.	Reading	0	0	0	5	9	14
	Listening	0	0	1	3	10	14
The discussion of Round 1 ratings and instructions helped me understand what I needed to do to complete Round 2.	Reading	0	0	0	4	10	14
	Listening	0	0	0	2	12	14
I understood how to complete my Round 2 ratings.	Reading	0	0	0	3	11	14
	Listening	0	0	0	1	13	14
I am confident in my Round 2 Ratings.	Reading	0	0	0	8	6	14
	Listening	0	0	0	5	9	14
I had the opportunity to ask questions while working on my Round 2 ratings.	Reading	0	0	0	2	12	14
	Listening	0	0	0	2	12	14
The technologies were helpful and functioned well.	Reading	0	0	4	6	4	14
	Listening	0	0	3	6	5	14
The timing of Round 2 was appropriate.	Reading	0	0	2	5	7	14
	Listening	0	0	0	7	7	14
The pace of Round 2 was appropriate.	Reading	0	0	1	7	6	14
	Listening	0	0	1	6	7	14

Table D5: Final (Evaluation 5) survey

Statements	Section	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Overall, the training in the standard setting purpose and method was clear.	Reading	0	0	0	9	5	14
	Listening	0	0	0	3	11	14
Overall, I am confident that I was able to apply the standard setting method appropriately.	Reading	0	0	0	7	7	14
	Listening	0	0	0	4	10	14
Overall, the standard setting procedures allowed me to use my experience and expertise to align items to the CEFR scales.	Reading	0	0	0	3	11	14
	Listening	0	0	0	2	12	14
Overall, the facilitators helped to ensure that everyone was able to contribute to the group discussions and that no one unfairly dominated the discussions.	Reading	0	0	0	4	10	14
	Listening	0	0	0	3	11	14
Overall, I was able to understand and use the scales scores provided (i.e., the scaled scores provided in square brackets).	Reading	0	0	1	4	9	14
	Listening	0	0	0	5	9	14
The technologies were helpful and functioned well.	Reading	0	1	3	6	4	14
	Listening	0	0	1	6	7	14

Appendix E: Pre-workshop Task 1 MPI indices

Table E1: Pre-workshop task MPI indices

	Reading	**Listening**
J01	1.00	1.00
J02	1.00	1.00
J03	1.00	0.97
J04	0.99	0.96
J05	0.99	0.99
J06	1.00	0.89
J07	0.99	0.98
J08	0.99	0.99
J09	0.98	-
J10	1.00	1.00
J11	1.00	-
J12	1.00	1.00
J13	1.00	-
J14	0.96	0.95
J15	-	0.97
J16	-	0.96
J17	-	0.81
Average	0.99	0.96

Appendix F: Individual judge severity and precision of measures

Table F1: Reading Round 1

	Observed Average (Fair Average)	Measure (S.E.)	Infit (Zstd)	Outfit (Zstd)	Correlation Ptmea (PtExp)	Obs% (Exp%)
J01	3.72 (3.76)	-2.95 (0.21)	1.01 (0.0)	1.04 (0.2)	0.91 (0.93)	46.9 (47.6)
J02	3.67 (3.71)	-2.77 (0.21)	0.97 (-0.1)	0.98 (0.0)	0.94 (0.92)	45.7 (48.7)
J03	3.01 (2.91)	-0.30 (0.21)	0.83 (-1.0)	0.82 (-1.1)	0.95 (0.92)	43.6 (44.7)
J04	3.82 (3.84)	-3.22 (0.22)	1.10 (0.6)	1.17 (1.0)	0.9 (0.92)	38.9 (45.3)
J05	3.44 (3.36)	-1.67 (0.21)	1.20 (1.2)	1.21 (1.2)	0.91 (0.92)	48.3 (51.6)
J06	3.44 (3.44)	-1.91 (0.21)	1.19 (1.2)	1.14 (0.8)	0.9 (0.92)	49.2 (51.7)
J07	3.37 (3.36)	-1.65 (0.21)	0.51 (-3.8)	0.49 (-3.9)	0.95 (0.92)	55.0 (51.7)
J08	3.16 (3.09)	-0.87 (0.21)	0.97 (-0.1)	1.02 (0.1)	0.94 (0.92)	46.9 (48.9)
J09	3.69 (3.71)	-2.75 (0.22)	1.21 (1.2)	1.19 (1.1)	0.94 (0.93)	43.9 (49.1)
J10	3.51 (3.53)	-2.17 (0.21)	0.89 (-0.6)	0.89 (-0.6)	0.97 (0.92)	49.4 (51.3)
J11*	-	-	-	-	-	-
J12	3.06 (2.96)	-0.47 (0.21)	0.82 (-1.1)	0.87 (-0.7)	0.92 (0.92)	44.9 (46.2)
J13	3.31 (3.28)	-1.43 (0.21)	1.06 (0.4)	1.07 (0.4)	0.91 (0.92)	46.0 (51.3)
J14	3.55 (3.47)	-1.98 (0.22)	1.04 (0.2)	1.05 (0.3)	0.94 (0.92)	47.1 (51.0)
Mean	3.44 (3.42)	-1.86 (0.21)	0.98 (-0.1)	0.99 (-0.1)	0.93	
***S.D.* (popul.)**	0.25 (0.29)	0.89 (0.00)	0.19 (1.3)	0.19 (1.3)	0.02	

* J11 dropped from the analysis

Table F2: Reading Round 2

	Observed Average (Fair Average)	Measure (S.E)	Infit (Zstd)	Outfit (Zstd)	Correlation Ptmea (PtExp)	Obs% (Exp%)
J01	3.60 (3.80)	-0.89 (0.25)	0.82 (-1.1)	0.80 (-1.1)	0.94 (0.94)	58.1 (56.5)
J02	3.52 (3.63)	-0.12 (0.25)	0.92 (-0.4)	0.87 (-0.6)	0.95 (0.94)	58.2 (60.8)
J03	3.16 (3.17)	1.71 (0.25)	0.55 (-3.3)	0.47 (-3.5)	0.97 (0.94)	59.8 (56.5)
J04	3.68 (3.86)	-1.24 (0.25)	1.25 (1.5)	1.28 (1.4)	0.91 (0.94)	47.1 (53.7)
J05	3.35 (3.47)	0.49 (0.25)	1.03 (0.2)	1.15 (0.8)	0.94 (0.94)	57.9 (61.3)
J06	3.49 (3.58)	0.10 (0.25)	1.07 (0.4)	1.07 (0.4)	0.93 (0.94)	58.5 (61.2)
J07	3.36 (3.44)	0.61 (0.25)	0.45 (-4.2)	0.39 (-4.2)	0.97 (0.94)	66.8 (61.5)
J08	3.35 (3.34)	0.97 (0.25)	0.98 (0.0)	1.04 (0.2)	0.96 (0.94)	57.7 (59.8)
J09	3.52 (3.58)	0.08 (0.26)	1.18 (1.0)	1.13 (0.6)	0.93 (0.94)	55.8 (61.2)
J10	3.58 (3.71)	-0.47 (0.25)	1.00 (0.0)	0.98 (0.0)	0.97 (0.94)	56.5 (59.6)
J11*	-	-	-	-	-	-
J12	3.19 (3.21)	1.51 (0.24)	0.92 (-0.4)	0.86 (-0.7)	0.94 (0.94)	56.8 (57.9)
J13	3.29 (3.34)	0.97 (0.25)	0.97 (-0.1)	1.03 (0.2)	0.94 (0.94)	58.0 (60.6)
J14	3.39 (3.42)	0.67 (0.25)	1.31 (1.7)	1.35 (1.7)	0.95 (0.94)	55.5 (61.6)
Mean	3.42 (3.51)	0.34 (0.25)	0.96 (-0.4)	0.96 (-0.4)	0.95	
***S.D.* (popul.)**	0.15 (0.21)	0.84 (0.00)	0.24 (1.7)	0.27 (1.7)	0.02	

* J11 dropped from the analysis

Table F3: Listening Round 1

	Observed Average (Fair Average)	Measure (S.E)	Infit (Zstd)	Outfit (Zstd)	Correlation Ptmea (PtExp)	Obs% (Exp%)
J01	3.53 (3.81)	1.95 (0.26)	0.87 (-0.7)	0.8 (0.0)	0.91 (0.88)	54.3 (55.9)
J02	3.43 (3.66)	1.28 (0.26)	0.94 (-0.3)	0.76 (0.0)	0.89 (0.89)	55.2 (55.6)
J03	3.56 (3.98)	2.89 (0.27)	1.06 (0.4)	1.15 (0.4)	0.93 (0.86)	46.8 (52.2)
J04*	-	-	-	-	-	-
J05	4.34 (4.92)	7.75 (0.31)	1.07 (0.4)	0.66 (0.2)	0.78 (0.8)	32.6 (30.6)
J06	2.85 (2.96)	-2.31 (0.26)	0.98 (0.0)	0.80 (0.0)	0.94 (0.93)	34.1 (33.5)
J07	3.63 (3.86)	2.21 (0.26)	1.24 (1.4)	1.02 (0.4)	0.89 (0.88)	55.3 (56.9)
J08	3.33 (3.53)	0.77 (0.26)	0.75 (-1.5)	0.59 (-0.4)	0.92 (0.91)	55.2 (54.6)
J10	3.35 (3.71)	1.49 (0.26)	0.91 (-0.5)	0.76 (-0.1)	0.91 (0.9)	55 .0(53.5)
J12	3.29 (3.52)	0.75 (0.26)	0.83 (-0.9)	0.64 (-0.4)	0.93 (0.9)	51.2 (53.2)
J14*	-	-	-	-	-	-
J15	4.09 (4.44)	5.13 (0.29)	0.92 (-0.3)	0.68 (-0.1)	0.85 (0.87)	51.2 (53.2)
J16	3.73 (3.97)	2.85 (0.28)	1.18 (1.0)	1.04 (0.5)	0.86 (0.88)	42.0 (43.5)
J17	3.44 (3.64)	1.20 (0.29)	1.06 (0.3)	0.96 (0.2)	0.91 (0.91)	51.9 (55.7)
Mean	3.55 (3.83)	2.16 (0.27)	0.98 (-0.1)	0.82 (0.1)	0.89	
***S.D.* (popul.)**	0.37 (0.47)	2.36 (0.02)	0.14 (0.8)	0.17 (0.3)	0.04	

* J04 & J14 dropped from the analysis

Table F4: Listening Round 2

	Observed Average (Fair Average)	**Measure (S.E)**	**Infit (Zstd)**	**Outfit (Zstd)**	**Correlation Ptmea (PtExp)**	**Obs% (Exp%)**
J01	3.52 (3.19)	0.61 (0.25)	1.03 (0.2)	1.00 (0.0)	0.93 (0.93)	60.1 (63.0)
J02	3.49 (3.17)	0.43 (0.25)	0.70 (-2.1)	0.68 (-1.4)	0.94 (0.93)	65.1 (63.1)
J03	3.89 (3.65)	2.58 (0.27)	0.91 (-0.4)	0.78 (-0.8)	0.96 (0.94)	56.5 (56.6)
J04*	-	-	-	-	-	-
J05	3.66 (3.33)	1.30 (0.26)	0.54 (-3.5)	0.45 (-2.8)	0.97 (0.93)	67.0 (62.8)
J06	3.73 (3.21)	0.69 (0.28)	1.35 (1.8)	1.31 (1.2)	0.94 (0.93)	55.6 (63.4)
J07	3.57 (3.17)	0.47 (0.25)	0.81 (-1.2)	0.72 (-1.3)	0.91 (0.92)	62.8 (62.0)
J08	3.48 (3.13)	0.17 (0.25)	0.77 (-1.6)	0.69 (-1.4)	0.92 (0.93)	62.3 (61.7)
J10	3.58 (3.28)	1.07 (0.26)	1.31 (1.8)	1.26 (1.0)	0.90 (0.93)	55.3 (63.4)
J12	3.29 (3.06)	-0.61 (0.26)	0.89 (-0.6)	0.8 (-0.8)	0.94 (0.93)	56.4 (57.9)
J14*	-	-	-	-	-	-
J15	3.70 (3.39)	1.57 (0.26)	0.72 (-1.8)	0.59 (-1.9)	0.95 (0.93)	63.4 (62.0)
J16	3.67 (3.31)	1.23 (0.26)	1.25 (1.5)	1.20 (0.8)	0.93 (0.93)	55.7 (63.0)
J17	3.52 (3.17)	0.46 (0.26)	1.30 (1.7)	1.27 (1.1)	0.96 (0.93)	54.0 (62.8)
Mean	3.59 (3.25)	0.83 (0.26)	0.97 (-0.4)	0.9 (-0.5)	0.94	
***S.D.* (popul.)**	0.15 (0.15)	0.77 (0.01)	0.27 (1.8)	0.28 (1.3)	0.02	

* J04 & J14 dropped from the analysis

Appendix G: Rasch-Kappa indices

Table G1: Rasch-Kappa indices

	Reading		**Listening**	
	Round 1	**Round 2**	**Round 1**	**Round 2**
J01	-0.01	0.04	-0.04	-0.08
J02	-0.06	-0.07	-0.01	0.05
J03	-0.02	0.08	-0.11	0.00
J04	-0.12	-0.14	dropped	Dropped
J05	-0.07	-0.09	0.03	0.11
J06	-0.05	-0.07	0.01	-0.21
J07	0.07	0.14	-0.04	0.02
J08	-0.04	-0.05	0.01	0.02
J09	-0.10	-0.14	-	
J10	-0.04	-0.08	0.03	-0.22
J11	dropped	dropped	-	-
J12	-0.02	-0.03	-0.04	-0.04
J13	-0.11	-0.07	-	-
J14	-0.08	-0.08	dropped	dropped
J15	-	-	-0.03	0.04
J16	-	-	-0.09	-0.20
J17			-0.10	-0.24
Mean	-0.05	-0.05	-0.02	-0.06

Appendix H: Round 1 and Round 2 MPI indices

Table H1: MPI indices

	Reading		**Listening**	
	Round 1	**Round 2**	**Round 1**	**Round 2**
J01	0.88	0.92	0.93	0.93
J02	0.90	0.92	0.89	0.90
J03	0.90	0.92	0.90	0.91
J04	0.88	0.90	dropped	dropped
J05	0.90	0.91	0.88	0.91
J06	0.89	0.88	0.90	0.90
J07	0.92	0.92	0.89	0.91
J08	0.89	0.90	0.92	0.92
J09	0.89	0.90	-	-
J10	0.89	0.90	0.89	0.90
J11	dropped	dropped	-	-
J12	0.92	0.92	0.91	0.91
J13	0.87	0.89	-	-
J14	0.89	0.90	dropped	Dropped
J15	-	-	0.94	0.89
J16	-	-	0.86	0.87
J17	-	-	0.91	0.91
Mean	0.89	0.91	0.90	0.91

Appendix I: Spearman correlations

Table I1: R1-R2 Spearman Correlations

	Reading	Listening
J01	0.91	0.99
J02	0.96	0.99
J03	0.98	0.99
J04	0.91	dropped
J05	0.93	0.81
J06	0.98	0.89
J07	0.98	0.94
J08	0.98	0.98
J09	0.93	-
J10	0.99	0.98
J11	dropped	-
J12	0.96	1.00
J13	0.95	-
J14	0.98	dropped
J15	-	0.92
J16	-	0.97
J17	-	0.99
Mean	0.96	0.95