TRINITY
COLLEGE LONDON

LINKING ISE DIGITAL TO THE CEFR: SETTING CUT SCORES AND PERFORMANCE STANDARDS

Paraskevi Kanistra

2025

Trinity Research Report 2024-01

About the Author

Paraskevi (Voula) Kanistra is Associate Director/Senior Researcher at Trinity College London and holds a PhD in Language Testing from the University of Bremen. She specialises in designing English language assessments that combine academic rigour with practical impact, ensuring that tests are fair, transparent, and aligned with the needs of learners and institutions worldwide.

Her expertise spans standard setting, CEFR alignment, validation, and measurement, and she has presented her work at major international conferences in Europe and Asia. She has published in leading language assessment journals.

She is the author of a forthcoming book on the Item Descriptor Matching method, to be published by Peter Lang. The book examines how this approach supports standard setting in both face-to-face and synchronous virtual environments, providing a systematic framework for linking candidate performance to CEFR levels and enhancing comparability across delivery modes.

About Trinity College London

Trinity College London is a leading international exam board and independent education charity that has been providing assessments around the world since 1877. We specialise in the assessment of communicative and performance skills covering music, drama, combined arts and English language. With over 850,000 candidates a year in more than 60 countries worldwide, Trinity's qualifications are specifically designed to help students progress. Our aim is to inspire teachers and candidates through the creation of assessments that are enjoyable to prepare for, rewarding to teach and that develop the skills needed in real life.

At the heart of Trinity's work is the belief that effective communicative and performance skills are life enhancing, know no boundaries and should be within reach of us all. We exist to promote and foster the best possible communicative and performance skills through assessment, content and training that is innovative, personal and authentic.

Trinity College London
Charity number England & Wales | 1014792
Charity number Scotland | SC049143
Patron | HRH The Duke of Kent KG
Chief Executive | Erez Tocker

Published by Trinity College London
First impression, 2025

Table of Contents

Tables

Figures

EXECUTIVE SUMMARY

This report documents the study conducted to link Trinity College London's Integrated Skills in English (ISE) Digital test to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR, Council of Europe, 2001). The study aimed to determine cut scores at CEFR levels A1 to C2 for each of the four skills (speaking, listening, reading, and writing), using the *Item Descriptor (ID) Matching* method (Ferrara, Perie, & Johnson, 2008; Ferrara & Lewis, 2012) and the *Principled Cut Score* approach (Kanistra, forthcoming). This methodology combines expert judgment with empirical data to provide a comprehensive alignment process. It particularly emphasises internal validity, focusing on consistency within the standard-setting process and the reliability of the classification decisions it produces.

Cut scores were successfully derived for each CEFR level and skill. Analyses were conducted to evaluate the internal consistency of the method, including agreement among expert judges and classification consistency of the resulting cut scores. The relative position of the cut scores was also examined in relation to the population mean and the CEFR level means of a retired diagnostic instrument that is widely accepted as a gold standard in CEFR linking methodology (DIALANG, https://wp.lancs.ac.uk/ltrg/projects/dialang-2-0/), offering further insight into their defensibility.

Findings across the four skills indicate that both linking methods (the *ID Matching* method and the *Principled Cut Score* approach) produced internally coherent and consistent results. Judgments showed high agreement, and the classification consistency estimates met or exceeded established benchmarks for high-stakes testing. The placement of cut scores reflected a logical progression across CEFR levels and aligned with the qualitative progression reflected in the CEFR scales of the CEFR Companion Volume (Council of Europe, 2020). Skill-specific analyses revealed some variation, as is typical in multi-skill assessments, but no evidence was found to suggest misalignment or method failure.

Overall, the evidence supports the validity and defensibility of the proposed CEFR cut scores for ISE Digital, as well as the interpretations based on them. The findings demonstrate that the resulting cut scores are consistent, interpretable, and appropriate for use in a high-stakes digital assessment context. This provides a robust foundation for further ISE Digital validity studies and its future operational use as a CEFR-aligned test.

1 Introduction

This report documents the CEFR alignment approach and standard-setting study to set cut scores for ISE Digital, Trinity's fully computer-delivered variant of the ISE qualification. The standard-setting study was conducted before the launch of the exam and comprised the following steps, as recommended in the manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (the Manual, Council of Europe, 2009):

- Familiarisation (of the panel members with the CEFR proficiency level descriptors and the CEFR categories)
- Specification (of the test tasks, items and content in relation to the CEFR)
- Standardisation, benchmarking and training in the method (training of panellists to gain a shared understanding of using the CEFR scales to relate tasks and performances to CEFR levels, including training of panellists on how to apply the method within the context of this study)
- Standard setting (the actual relation of tests or performances to CEFR levels)
- Validation (of the test, the panellist training, and the internal standard setting results)

Steps 1-4 were conducted during the virtual standard-setting workshop. Step 5 was conducted after the workshop, and the internal and external validity of the standard-setting procedure was evaluated.

The standard setting was conducted using data collected from the pilot phase. Two different methods were identified following a detailed analysis of the ISE Digital exam specifications, item creation processes, the available data, a systematic literature review of feasible standard-setting methods in the context of aligning exams to the CEFR, and the author's extensive research on standard setting:

- The *ID Matching* method (Ferrara, Perie, & Johnson, 2008; Ferrara & Lewis, 2012) operationalised as both an examinee-centred and a test-centred method (see Harsch & Kanistra, 2020). This was used for the speaking and writing modules.
- The *Principled Cut Score* approach (Kanistra, forthcoming, 2023) is situated within the Unified Alignment and Test Design (UATD) approach (Kanistra, forthcoming, 2023). This was used for the listening and reading modules.

1.1 ISE DIGITAL Overview

Trinity's Integrated Skills in English (ISE) exams provide an assessment of candidates' English language proficiency across four skills: speaking, listening, reading, and writing. ISE Digital is a multi-level, adaptive exam covering all six levels of the CEFR from A1 to C2. The exam has been designed to reflect the types of tasks and texts that students encounter within the educational domain or during their professional life. Preparing for ISE Digital helps develop authentic communicative abilities and transferable skills that are crucial for academic study and employment. These skills include synthesising information, participating in interactive discussions, and presenting on topics of personal interest.

ISE Digital is designed for young people and adults, typically those in school, college, or university, who are learning and using English in their academic studies. The typical ISE Digital candidate is aged between 12 and 19, but the exam is also suitable for working adults seeking a respected English language qualification.

Candidates taking the exam will first complete a short levelling test, which is used to select test content that is best suited to their language proficiency. Candidates will then complete a module for each skill. Although each module primarily focuses on one language skill, some tasks assess the skills together. This integrated approach reflects how language skills are used in real-life settings.

Table 1 provides an overview of the ISE Digital modules, tasks, and requirements. Detailed information about each module is available in the ISE Digital Exam information booklet.

Table 1.1:ISE Digital: modules, tasks and requirements

Module	Task	Task requirement
Speaking	**Responding to questions**	Describe objects, people or places and express opinions on a topic
	Delivering a prepared talk	Give a prepared talk on a topic of the candidate's choice and answer a follow-up question
	Interacting	Listen and respond to a scenario; respond to new information
	Summarising a talk or conversation	Listen to a conversation and give a summary with an opinion
Listening	**Listening to a description**	Listen to a description of people, places, objects or activities and answer questions
	Listening to a conversation	Listen to an informal conversation between two people and answer questions
	Listening to a discussion	Listen to a discussion between invited panellists and a host and answer questions
	Listening to a talk	Listen to a talk followed by a retelling of the talk by a second speaker and answer questions
Reading	**Reading a visual text**	Read a short text with visuals (eg a poster/leaflet) and answer questions
	Reading a single text	Read a single text on a topic and answer questions
	Reading a paired text	Read two texts on the same theme and answer questions
Writing	**Written online communication**	Write a short contribution to an opinion-based discussion, give suggestions or feedback, or respond to a group chat
	Writing from sources	Read two or three source texts and write an essay/ report in response to a prompt, synthesising relevant information from the source texts and adding own ideas and stance on the topic

The reading and listening tasks comprise reading or listening input materials accompanied by multiple-choice questions. For each reading and listening multiple-choice question, only one option is correct. A computer marks the candidates' answers. The candidate's speaking and writing performances are evaluated by professional language assessors who use rating scales specifically developed for the exam. The rating scales are available in the ISE Digital Information Booklet (pp. 32-46).

The adaptive nature of the exam ensures that, depending on their ability, candidates will be directed to an A1-A2, B1-B2 or C1-C2 route. Candidates will see tasks that are suitable for their level of English language proficiency. They may see some task types and not others. This ensures that the candidate receives a challenge that is appropriate for their level.

The ISE Digital results report provides candidates with a score for each language skill (speaking, listening, reading, and writing) on a scale of zero to 150, along with the corresponding CEFR level. The results report also includes a diagnostic profile of the candidate's performance in each skill, showing the areas where they performed well and the areas where they might wish to practise and develop further.

All tasks were developed by drawing extensively from the relevant literature underpinning language assessment and second language acquisition. The theoretical foundations for the exam have been published as framework documents (Trinity College London, 2025a, 2025b, 2025c, 2025d, 2025e). The whole test development and design process was also situated within the *Principled Assessment Approaches Design and Implementation* (PADDI) process (Ferrara, Lai, & Nichols, 2016; Lewis & Cook, 2020), referencing the CEFR and mapping specific skills and subskills to the CEFR to the extent possible. Furthermore, the CEFR-informed construct and tasks are consequently reflected in the assessment criteria used to evaluate candidates' written and spoken performances.

To ensure a systematic alignment of live items with the CEFR, all examination content is developed in accordance with step 4 of the UATD approach (Kanistra, forthcoming), whereby each task targets specific CEFR levels. To generate CEFR-aligned tasks and items, all listening and reading texts are written within a specific range of readability indices, which have been pegged to the targeted CEFR level. The topic and domain coverage of the reading and listening texts also align with the targeted CEFR levels. The development of speaking and writing prompts follows a similar process, and item writers and reviewers ensure that topics, domains, content, and other readability indices – particularly in the presence of longer text input – align with the targeted CEFR level specifications. Item writers use EDIA Papyrus (https://www.edia.nl/papyrus) to verify the alignment of longer texts with the CEFR. All item writers undergo a rigorous training session, including familiarisation with the test construct and the relevant CEFR scales and descriptors. All items are piloted before being included in a live administration, and the listening and reading item banks are calibrated through Rasch Measurement Theory (RMT).

1.2 The Common European Framework of Reference for Languages

The CEFR is the outcome of projects funded by the Council of Europe (https://www.coe.int/en/web/language-policy/cefr). It was first published in 2001 (Council of Europe, 2001) and updated as a Companion Volume in 2020 (Council of Europe, 2020). It is intended to guide the preparation of language syllabuses, curriculum guidelines, teaching and learning materials, and assessment. As such, it provides a common basis for comparing language courses, syllabuses, and qualifications, offering a transparent tool for discussion and reflection. The CEFR is language independent. Though primarily used in Europe, it has also been applied in other continents.

The CEFR describes language proficiency at six levels, ranging from A1 (Breakthrough) to C2 (Mastery). The primary focus is on communicative language competences, activities, and strategies in the comprehension and production of language, in interaction using language, and in the mediation of information and opinions using language. There is a global scale that offers a snapshot of language proficiency at each CEFR level, along with 53 illustrative scales. Each of these scales provides a detailed description of how specific language competences, activities, and strategies progress in relation to increases in language proficiency. Figure 1, reproduced from the CEFR Companion Volume (Council of Europe, 2020, p. 36), presents a view of how each CEFR level is nested within higher levels, illustrating the incremental and cumulative gains that learners make as they progress on their language learning journey.

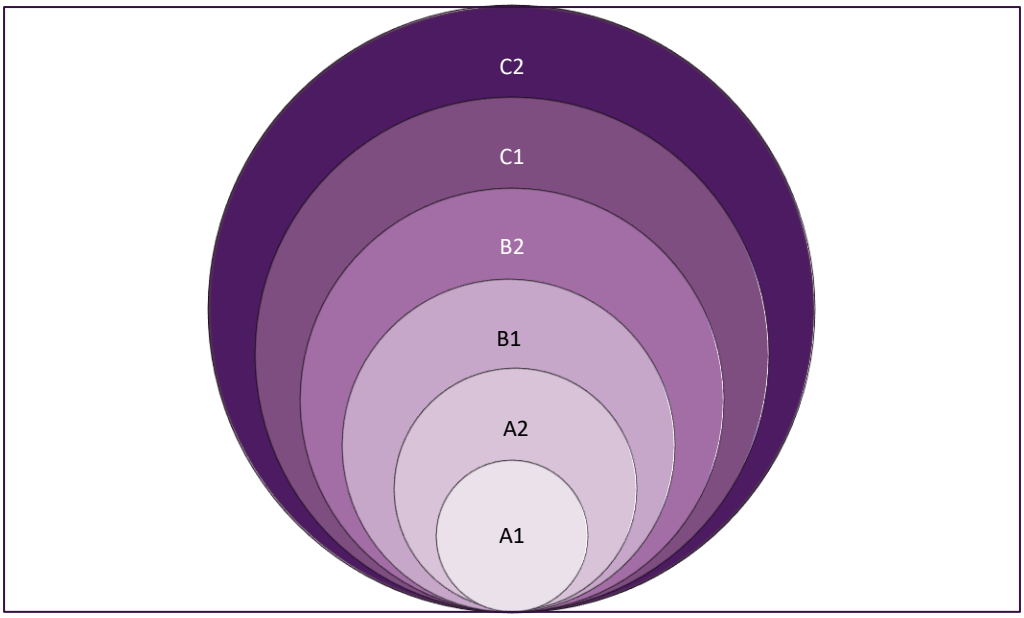


Figure 1.1: CEFR Common Reference Levels

The Council of Europe has also supported the creation of various ancillary materials, including the European Language Portfolio, a resource platform for plurilingual and intercultural education, tools for applying the CEFR in the classroom, and several manuals and guides to assist test designers in aligning their exams with the CEFR. In 2003, the Council of Europe published a pilot version of the Manual for Relating Language Examinations to the CEFR (Manual, Council of Europe, 2003) along with a Reference Supplement. The final edition of the Manual was published in 2009 (Council of Europe, 2009). In 2022, a handbook for aligning language education with the CEFR was published (EALTA, UKALTA, the British Council, and ALTE, 2022). This handbook has influenced Trinity's approach to relating ISE Digital to the CEFR.

1.3 Structure of this Report

In addition to the introduction, this report comprises eight sections. The next section offers an overview of standard setting procedures, especially in relation to the CEFR, and details the specific approach taken for ISE Digital. Section 3 describes the post-hoc validation methods used in this study. Section 4 covers the CEFR familiarisation methodology and outcomes. This is followed by sections describing the process of deriving and validating cut scores for each language skill. The skills are reported in the order in which they appear on the exam. The final section presents a reflection of the findings.

2 Standard Setting Methodology

In language testing and assessment, the CEFR (Council of Europe, 2001, 2020) has significantly impacted how language test results are reported in Europe and beyond. Most exams define and align their achievement levels with the six proficiency levels of the CEFR. Several alignment and standard-setting procedures are explained in the Manual (Council of Europe, 2009). The Manual also outlines the steps required to classify exam results into achievement levels, which are defined by CEFR proficiency levels and their corresponding descriptors. The Manual recommends five steps for any alignment process (Figure 2.1), with each step being evaluated after completion.

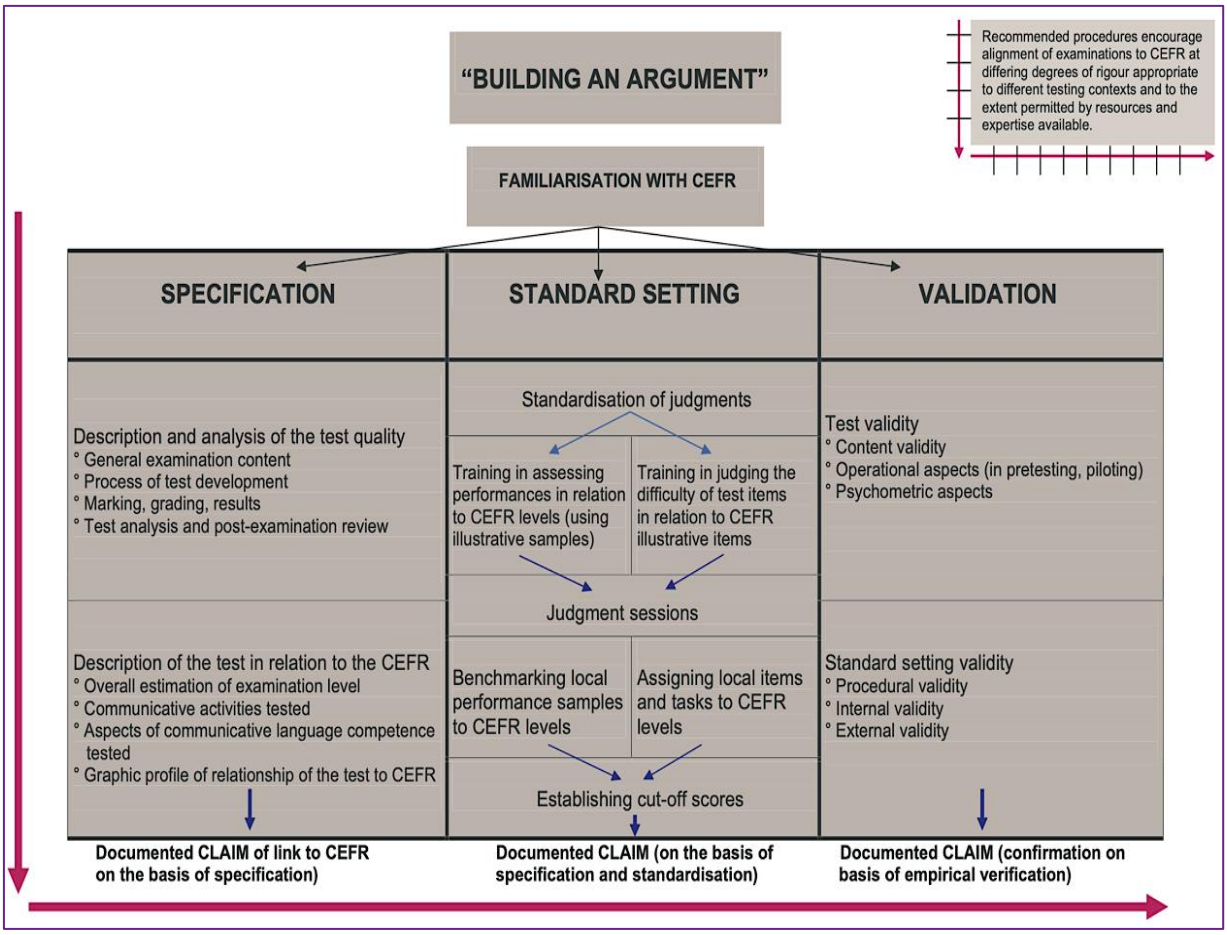


Figure 2.1: Visual representation of the procedures for relating examinations to the CEFR (Council of Europe, 2009, p. 15)

In summary, the five steps described in the Manual are:

- Familiarisation (of the panel members with the CEFR proficiency level descriptors and the CEFR categories)
- Specification (of the test tasks, items and content in relation to the CEFR)
- Standardisation, benchmarking and training in the method (training of panellists to gain a shared understanding of using the CEFR scales to relate tasks and performances to CEFR levels, including training of panellists on how to apply the method within the context of this study)
- Standard setting (the actual relation of tests or performances to CEFR levels)
- Validation (of the test, the panellist training, and the internal standard setting results)

The underlying premise of the Manual (2009) is that the linking process is conducted on a valid, reliable, and stable examination. Kanistra (forthcoming) demonstrated how the Item Descriptor (ID) Matching method enhances the scope and depth of the *Standard setting* stage in CEFR alignment studies and proposed an expansion to the alignment process proposed by the Manual (2009) and O'Sullivan (2013). A key advantage of the ID Matching method is that

it encourages panellists to consider the exam's underlying construct while linking items to CEFR levels, effectively conducting a "bottom-up content analysis" of an existing exam/test. By allowing panellists to access an examination's construct, the CEFR alignment model is expanded in several ways: (1) it incorporates the exam's construct into the *familiarisation* stage, (2) it aligns the activities in the *familiarisation stage* with the cognitive processes involved in the *ID Matching method*, (3) it assesses the success of the *familiarisation stage* through the panellist training during the *training-in-the-method* stage, and (4) it adds an extra evaluation step to the *specification stage* by assessing the consistency of the CEFR item mapping between the *specification* and *standard setting* stages (see Figure 2.2[1]). Ensuring consistency of item judgements gathered during these stages adds external validity to the *Specification* phase. These adjustments strengthen the alignment process by introducing additional checks and addressing discrepancies before final cut scores are established.

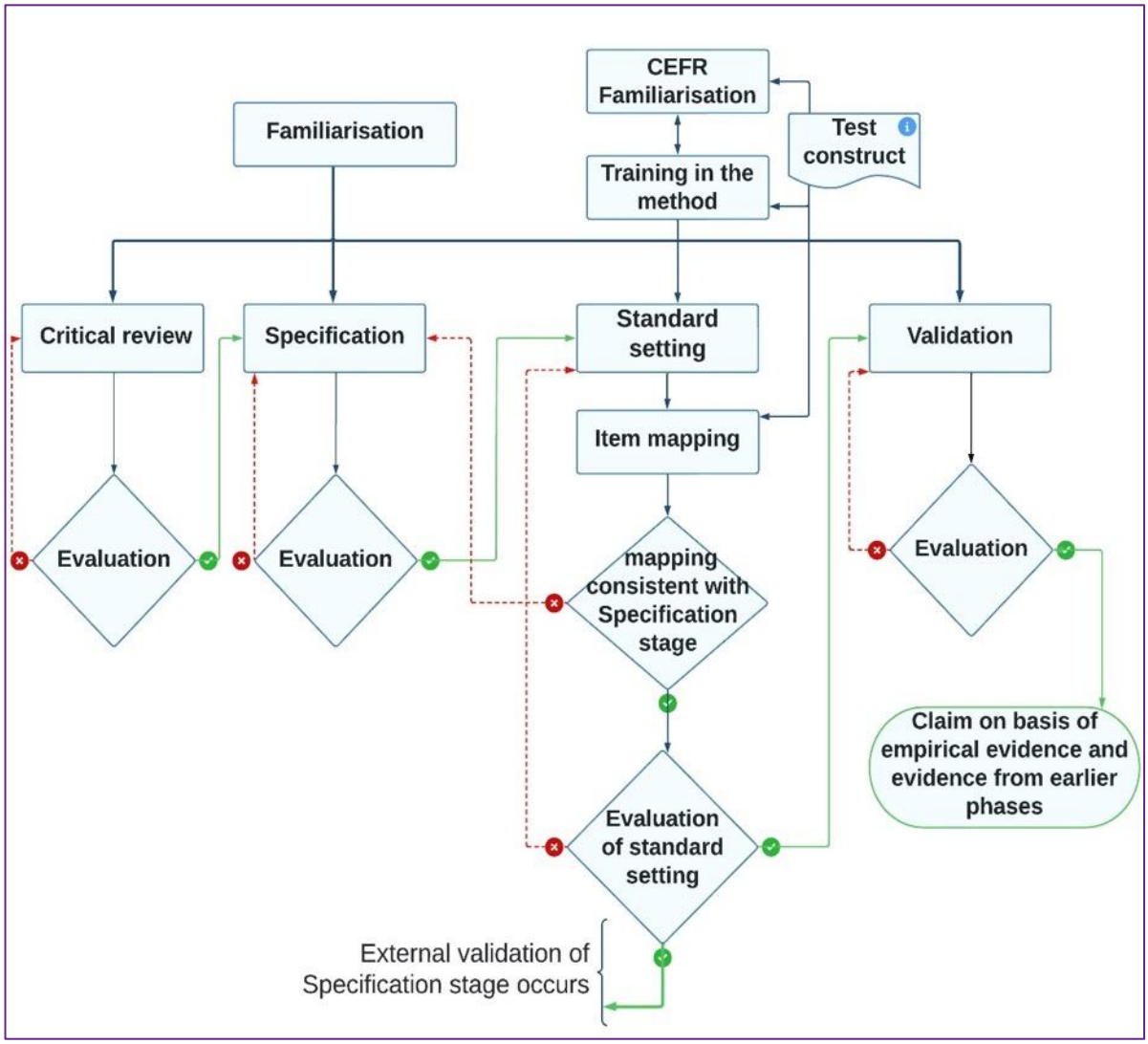


Figure 2.2: Expanded CEFR alignment model (Kanistra, forthcoming; adapted from O'Sullivan, 2013

The Manual (Council of Europe, 2009) cautions that, to ensure appropriate standards, the standard-setting process must be followed from the outset of an alignment study, requiring high-quality data and careful decision-making. Additionally, it is customary in standard setting

[1] a successful evaluation outcome (✓) signals the beginning of the next stage of the alignment process. A negative evaluation outcome (✗) implies that either the current or a previous stage needs to be repeated or revisited

to refer to content standards (which define the subject matter for exams) and to performance standards (which are specific to an examination). These performance standards are typically referred to as Performance Level Descriptors (PLDs) or Achievement Level Descriptors (ALDs) and serve as a reference framework for exam descriptions, expressing the minimum performance levels expected. In this sense, PLDs and ALDs are synonymous with cut scores.

Unlike in other contexts, where PLDs must be developed specifically for an examination, the CEFR provides content standards and qualitative PLDs. Therefore, it is paramount that the CEFR be referenced throughout the linking process (see Figure 2.3).

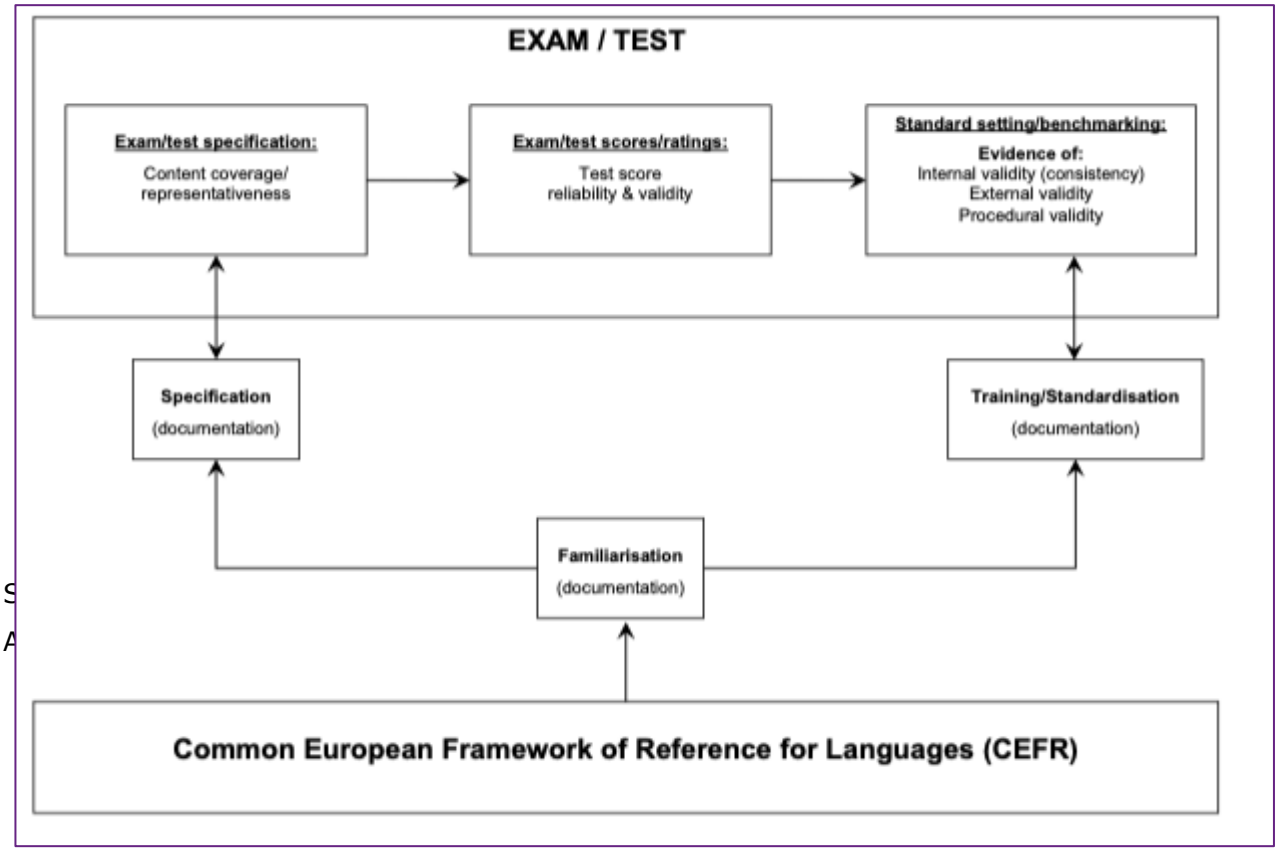


Figure 2.3: Validity evidence of linkage of examination/test results to the CEFR

2.1 Approaches to Test Development

Systematised test development approaches, such as Evidence-Centred Design (ECD) (Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006) and newer approaches, such as the Principled Approaches to Assessment Design, Development, and Implementation (PADDI, Ferrara, Lai, Reilly, & Nichols, 2017), offer a ripe environment for qualifications to be developed in alignment with the CEFR from the outset.

2.1.1 Evidence Centred Design (ECD)

ECD is a comprehensive framework that provides a structured approach to developing assessments grounded in explicit theories and models of learning and cognition. It formalises the assessment argument, presenting claims about test takers' knowledge and abilities based on evidence generated during the assessment process. The ECD organises assessment design into five interconnected layers: domain analysis, domain modelling, conceptual assessment framework, assessment implementation, and assessment delivery. This iterative process enables continuous refinement throughout the design, development, and implementation stages. Briefly, the five stages in the ECD are as follows:

1. **Domain Analysis:** Test designers gather information about the target domain, including learning models, performance theories, terminology, and relevant tools or technologies. This foundational layer informs subsequent decisions.
2. **Domain modelling:** Information from domain analysis is structured into a design document, which specifies key elements such as knowledge, skills, and abilities (KSAs), content, and performance requirements for assessment development
3. **Conceptual assessment framework:** Designers develop three interrelated models:
    - **Student model:** Defines the attributes and abilities the assessment aims to infer.
    - **Task model:** Outlines tasks and content designed to elicit evidence about the KSAs.
    - **Evidence model:** Specifies how student responses (work products) will be evaluated and scored, including rubrics, evidence rules, and statistical models.
4. **Assessment implementation:** Tools and specifications from the conceptual framework are used to create tasks, rubrics, and scoring systems, ensuring alignment with the intended inferences.
5. **Assessment administration:** The final stage involves administering the assessment, analysing results, and reporting outcomes using a four-process model for practical application.

2.1.2 Principled Approaches to Assessment Design, Development, and Implementation (PADDI)

The PADDI approach (Ferrara, Lai, Reilly, & Nichols, 2017) emphasises the integration of evidence to construct validity arguments. The process is structured into three key steps:

1. **Defining assessment targets and uses involves identifying the intended score interpretations and uses,** and setting clear assessment targets. These foundational steps guide the entire design process.
2. **Developing a test blueprint** involves selecting or developing models of cognition, learning, or performance, and aligning them with appropriate measurement models. This ensures that the design remains focused on accurately and validly capturing the desired constructs.
3. **Manipulating assessment items and tasks** involves generating and refining test items and tasks in accordance with the blueprint. Field testing, scaling, and psychometric analyses are conducted to ensure the reliability and validity of the items.

Recent advances in standard setting theory are more closely aligned with the CEFR linking process detailed in the Manual, and they apply standard setting methodology at the test design and development stages. This ensures that the link to the chosen framework (in this case, the CEFR) is threaded through the core of the exam. To this end, Lewis and Cook (2020) have proposed the *Embedded Standard Setting* (ESS) method, which integrates standard-setting practices directly within the assessment development process, aligning it with the ECD and PADDI approaches (Ferrara, Lai, & Nichols, 2016). Lewis and Cook (2022) situated the ESS method within the PADDI process.

- **Step 1:** The intended uses of an examination are established
- **Step 2:** The interpretative standards are identified through measurement targets, academic content standards, performance standards or achievement standards.

These two steps guide the entire test development process, ensuring that items are aligned with performance standards from inception.

- **Step 3:** In this final step, the qualitative performance standards set in Step 2 are translated to quantitative standards by identifying the critical numeric scores that imply a different interpretation. These critical scores, in essence, serve as the cut scores. Any items misaligned (ie associated with empirical difficulties inconsistent with the targeted item domains —PLD/ALD alignment) are reviewed by subject matter experts to identify and resolve the source of the misalignment.

This approach eliminates the need for traditional, subjective standard setting workshops since the PLDs and ALDs are produced before the standard setting workshop occurs.

2.1.3 The Unified Alignment and Test Design process (UATD)

Kanistra (2023, forthcoming) demonstrated that it is possible to derive cut scores with items and tasks aligned with the CEFR, adding further evidence in favour of the ESS approach. She adapted and expanded on Lewis and Cook's (2022) PADDI approach, developing the *Unified Alignment and Test Design* process (UATD, Figure 2.4) to address the fact that conceptual proficiency frameworks such as the CEFR provide qualitative PLDs that are not specific to any examination and do not quantify the number of KSAs a candidate needs to show.

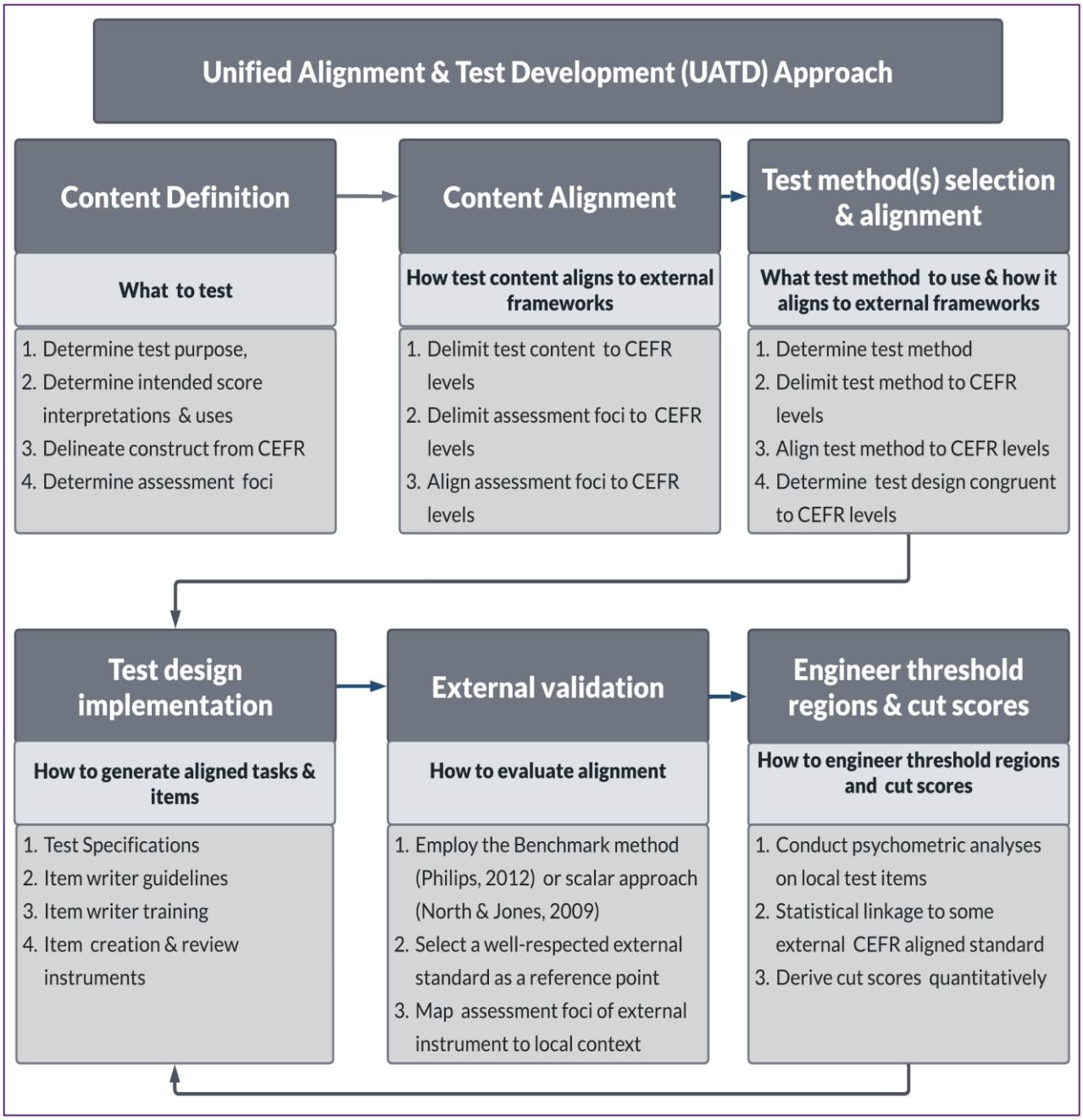


Figure 2.4: Structure of the Unified Alignment and Test Design (UATD) approach (Kanistra, 2023, forthcoming

The UATD approach demonstrates how the CEFR (or any other external framework) can be integrated throughout the assessment cycle, aligning with the underlying principles outlined in the Manual (2009) and depicted in Figure 2.4. In summary, the UATD approach consists of the following six steps:

1. **Content definition:** Definition of what an examination measures, ranging from its construct(s), purpose, and intended score interpretation and use in terms of CEFR levels.

2. **Content alignment:** Alignment and consistency of test content and assessment foci with the targeted CEFR level(s) demands.
3. **Test method(s) selection & alignment:** Selection of assessment methods aligned with the cognitive demands of the CEFR level(s), ensuring congruence with the content and objectives of the CEFR framework.
4. **Test design implementation:** Creation of test specifications, item writer guidelines, training, and item review instruments to ensure content is aligned with CEFR.
5. **External validation:** CEFR alignment validation using methodologies such as the Benchmark method (Philips, 2012), coupled with external instruments mapped to the local context (North and Jones, 2009). This step may include evaluating items from an external expert panel.
6. **Calculate threshold regions & cut scores:** Psychometric analyses and statistical linking are used to determine cut scores, ensuring they align quantitatively with an external CEFR-aligned standard.

The UATD approach expands on the ESS approach by incorporating external validation evidence into the alignment process, drawing on aspects of Philips' (2012) Benchmark method and North and Jones' (2009) data-based scalar approaches to setting CEFR-aligned cut scores. Philips (2012) proposed using benchmarks to statistically link the National Assessment of Educational Progress (NAEP) scale across individual States in the United States of America or other international scales such as the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) or Progress in International Reading Literacy Study (PIRLS) to support comparable score interpretations. Furthermore, North and Jones (2009) argue that using CEFR illustrative examples as anchors enhances alignment by providing benchmarks for validating, calibrating, and standardising proficiency levels across tests and languages. These examples ensure consistency in interpreting proficiency levels and maintain standards over time through IRT scaling. This strengthens the validity, reliability, and comparability of language proficiency assessments, ensuring they accurately reflect CEFR-defined communicative competencies.

Indeed, CEFR alignment studies can benefit greatly from Philips' (2012) Benchmark standard-setting approaches and North and Jones' (2009) data-based scalar approaches to setting cut-off CEFR points, as the addition of an external CEFR-aligned and validated test instrument can act as a reference and calibration point through the common item linking technique. Including common items allows test developers to explore how their items compare in terms of difficulty to those from external test instruments already aligned to the CEFR. It can also facilitate the direct comparison of cut scores set across various examinations aiming at the same CEFR level(s), thus indirectly allowing stakeholders to evaluate the interpretations made from these cut scores. More importantly, cut scores can be derived quantitatively in a principled manner when items and tasks are designed within the UATD approach. Including items from an external test instrument ensures that cut scores established in this principled approach are simultaneously externally validated and calibrated against this external criterion.

2.2 ISE Digital Development Process

To ensure rigorous alignment with the CEFR, Trinity developed ISE Digital using the ECD (Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006) and PADDI approaches (Ferrara, Lai, Reilly, & Nichols, 2017) described in Section 2.1. Figure 2.5 illustrates how the test development process was situated within an ECD/PADDI framework, ensuring that its intended score interpretation and uses were aligned a priori with the six levels of the CEFR (A1-C2). Test development began with a domain analysis of the target language use domains (education, workplace, and migration), reviewing the current literature on assessment theories and learning, and identifying the CEFR-aligned knowledge, skills, and abilities (KSAs) to be assessed. These were then mapped to real-world language use scenarios. During domain modelling, these KSAs were translated into a design document which mapped the different assessment targets to specific tasks and activities. Three interrelated models were established through the conceptual assessment framework: the student model, which defines the proficiencies to be inferred; the task model, which designs the tasks that elicit evidence of those proficiencies; and the evidence model, which establishes scoring rubrics, weighting rules,

and statistical models for evaluating candidate responses and interpreting scoring information. Test and form specifications were created during the implementation stage, and tasks and items were authored and field-tested. The data from the field tests were psychometrically analysed to define the adaptive algorithm and refine the assessment.

During the implementation stage, Trinity commissioned an interim critical review of the alignment between the CEFR and a subset of the initial tasks developed (Griffiths, 2023). The outcomes of this review informed revisions to the test and form specifications, item writer guidelines, training for raters and assessors, and other relevant training materials. The last phase, assessment delivery, focused on the more operational aspects of assessment, including processes for administering tests, controlling item exposure, and providing feedback to stakeholders. This iterative process ensured that the test was, by design, aligned with CEFR's communicative competence goals, supported meaningful score interpretations, and provided a robust foundation for measuring language proficiency across CEFR levels.

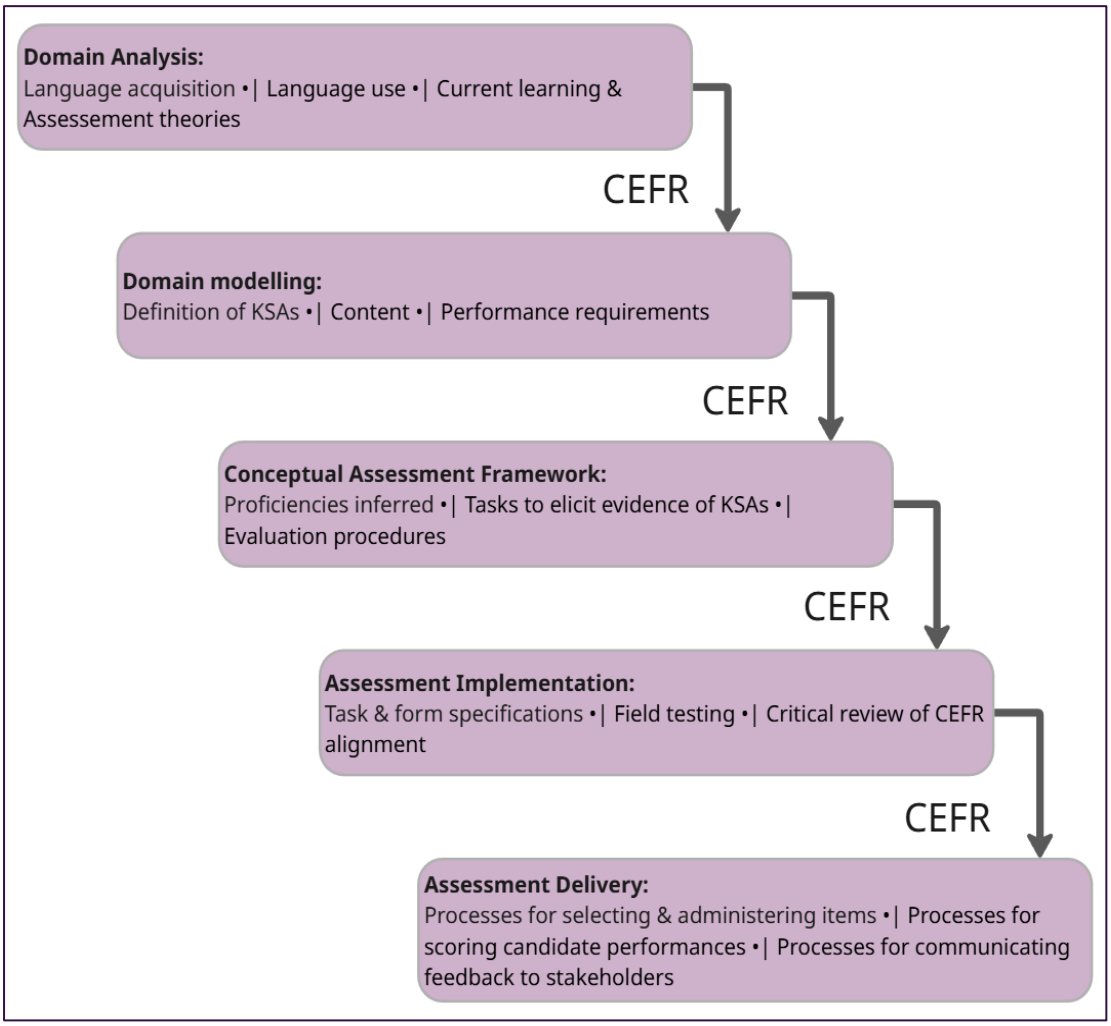


Figure 2.5: Developing ISE Digital within the ECD/PADDI framework

To ensure continued alignment with CEFR standards and maintain the integrity of the assessment framework, the outcomes of the final stage of the UATD approach (Figure 2.4), particularly the calculation of threshold regions and cut scores, inform and refine the Assessment Implementation stage iteratively. This redefinition underscores the significance of adhering to stringent criteria to ensure that items and tasks remain aligned with the targeted CEFR levels. Readability indices, such as the Gunning-Fog (FOG) and Automated Readability Index (ARI), as well as the Simple Measure of Gobbledygook (SMOG), which are derived from sophisticated tools like EDIA Papyrus, are meticulously employed to precisely calibrate the readability and linguistic characteristics of input texts, thereby maintaining the CEFR levelling standards. Moreover, psychometric analyses derived from this stage provide invaluable

insights into the types of questions item writers should focus on, enabling them to craft items that are most likely to align with targeted difficulty measures and ensure a robust alignment with the intended CEFR framework. The outcomes of this principled approach to item creation are reflected in the content analysis forms provided in the Manual (Council of Europe, 2009, Appendix A). The quantitative linking to the CEFR is reported by skill in Sections 5 to 8.

2.3 Standard Setting Methods

The literature on standard setting is vast and covers (at last count) more than 60 methods. This section focuses on the standard setting methods that were used to set CEFR-linked cut scores for ISE Digital.

2.3.1 Principled Cut Score Approach

The *Principled Cut Score* approach (Kanistra, 2023, forthcoming) was developed to address significant limitations in widely used standard setting methods such as the *Angoff*, *Bookmark*, and *ID Matching* methods, which often overlook or allow for misalignment between CEFR descriptors, test items, and panellist judgments (Lewis & Cook, 2020). Overlooking misaligned items during the cut score setting process compromises the validity of the resulting scores. Additionally, using conceptual frameworks such as the CEFR as PLDs presents challenges, as these frameworks do not provide a quantitative measure of the knowledge and skills candidates must demonstrate. As a result, even when panellists consistently align items with CEFR descriptors, this alignment does not necessarily ensure a reasonable or defensible cut score. The *Principled Cut Score* approach is based on an array of quantitative analyses and on the decision accuracy (*DA* (γ)) and consistency (*DC* (φ)) literature. In this approach, the test items must be calibrated using Rasch Measurement Theory (RMT) or Item Response Theory (IRT). Figure 2.6 graphically shows the six steps this approach employs.

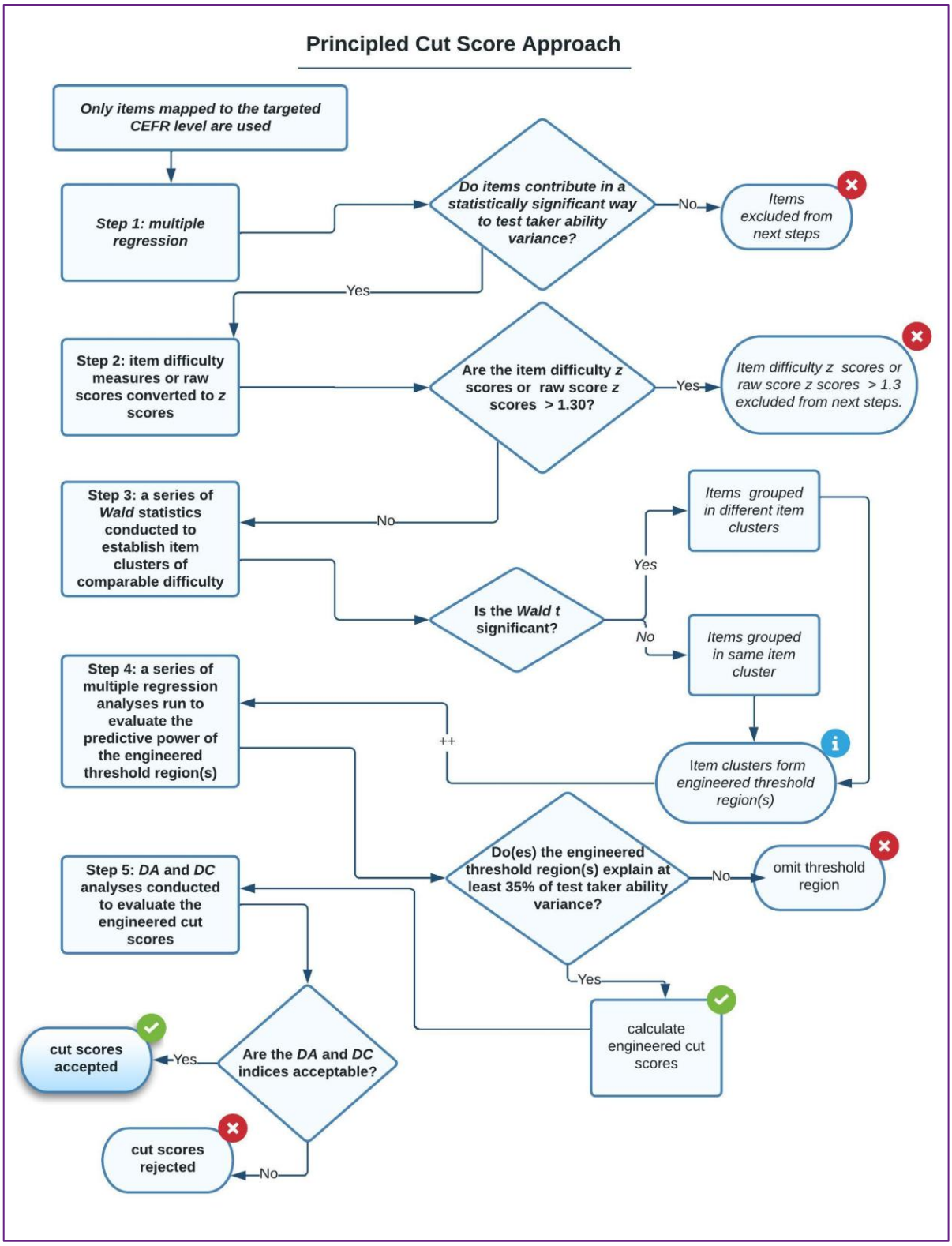


Figure 2.6: The Principled Cut Score approach (Kanistra, 2023, forthcoming)

As Figure 3.1 shows, cut scores can be calculated in five steps:

- **Step 1: Multiple regression analysis**: This step involves using multiple regression analyses (Pallant, 2016; Tabachnick & Fidell, 2014) to determine which items, of the ones mapped to the targeted CEFR level, significantly predict candidate ability. Items contributing uniquely and significantly to explaining candidate ability are carried forward to Step 2.
- **Step 2: Conversion of ability measures to z Scores**: Next, the item difficulty measures of the carried forward items are converted to z scores to examine how far

from the population mean potential cut scores are. This conversion helps identify cut scores that are appropriately located relative to the mean test-score distribution, as suggested by Subkoviak (1980, 1988).

- **Step 3: Item clustering using Wald statistics**: In this step, Wald statistics group items of comparable difficulty into clusters. Each cluster represents threshold regions where cut scores can be located.
- **Step 4: Multiple regression for predictive power of item clusters:** Multiple regression analyses are conducted for each item cluster identified in Step 3 to evaluate the cluster's ability to predict candidate ability. Clusters that explain a significant proportion of candidate ability in a statistically significant way determine the threshold regions on which cut scores are calculated.
- **Step 5: Evaluating calculated cut scores:** In the final step, cut scores are calculated using one of four methods: the minimum, maximum, mean, or median of the item difficulties within each threshold region. The accuracy and precision of these cut scores are then evaluated using Standard Error of judgments (SE_j), conditional standard error of measurement (*CSEM*), conditional reliability (*CREL*), decision consistency (*DC* (φ)), and decision accuracy (*DA* (γ)) indices.

2.3.2 The Item Descriptor Matching Method

The ID Matching method (Ferrara, Perie, & Johnson, 2008; Ferrara & Lewis, 2012) is a relatively new item mapping technique based on IRT. The conceptual roots of this method lie in NAEP's scale anchoring procedures and classification system, which evolved from a simple pass/fail model to more detailed levels, such as 'Below Basic', 'Basic', and 'Proficient'. The 2002 No Child Left Behind (NCLB) Act, which required states to document and communicate students' mastery of key knowledge, skills and abilities (KSAs) at different achievement levels, helped facilitate the adoption of this (among other new methods) over traditional approaches like Angoff, as they were better suited for complex assessments.

The question underpinning the ID Matching method is:

> "*Which performance level descriptor most closely matches the knowledge and skills required to respond successfully to this item (or score level for constructed-response items)"?*
>
> *(Ferrara, Perie, & Johnson, 2008, p. 12; Ferrara & Lewis, 2012, p. 262)*

Educators favour the *ID Matching* method due to its straightforward approach. Unlike other standard-setting methods, it does not require panellists to make probabilistic judgments or define a 'borderline' or a *'minimally competent candidate'* (MCC). Instead, panellists focus on identifying the KSAs required by test items and mapping them to predefined performance level descriptors (PLDs). This process aligns well with tasks familiar to educators after the implementation of the NCLB Act. The process of mapping items to achievement levels is less cognitively demanding than estimating the probability that a minimally competent candidate will get an item correct (Ferrara, Perie, & Johnson, 2008; Ferrara & Lewis, 2012).

Kanistra (forthcoming) also showed that the *ID Matching* method, applied in both face-to-face and virtual settings, is effective for standard setting tasks related to productive skills, such as writing. It demonstrates reliability and consistency across different environments, making it adaptable to various contexts, including synchronous virtual workshops. Additionally, the *ID Matching* method can be quite versatile, applicable to both task-centred and product-centred approaches. Unlike many standard setting methods traditionally used for productive skills (eg speaking and writing tasks), the *ID Matching* method uniquely bridges the gap between test-centred and product-centred methods. Harsch and Kanistra (2020) and Kanistra (forthcoming) have illustrated how this method complements the benchmarking process described in the CEFR Manual (Council of Europe, 2009). It incorporates an analysis of task demands, a step often missing in other productive skill methods. By ensuring the task aligns with CEFR levels before evaluating candidate performances, the method provides a stronger foundation for setting valid cut scores. Furthermore, the *ID Matching* method has shown robustness to individual differences among panellists, such as varying experience levels or direct affiliation with the organisation commissioning the standard setting study. This reduces the impact of potential biases and ensures that cut scores are not overly influenced by individual panellist idiosyncrasies (Kanistra, forthcoming).

In summary, the *ID Matching* method offers a structured, reliable, and adaptable approach to standard setting for productive skills. By incorporating both task and product evaluation, it addresses the complexities of productive skill assessments, aligns well with CEFR benchmarking practices, and ensures consistency and fairness across diverse settings.

2.4 Standard Setting Process

Following Harsch and Kanistra (2020) and Kanistra (forthcoming), standard-setting panellists first evaluated several speaking and writing tasks in terms of CEFR levels before benchmarking candidates' written and oral responses to those tasks. The following question guided the panellists' judgement task:

- Which CEFR level descriptor(s) most closely match(es) the knowledge, skills, abilities and/or cognitive processes required to produce an appropriate written/spoken response (A1, A2, etc.) to the speaking/writing task?
- Which CEFR level best reflects the knowledge, skills, and abilities demonstrated in the student's written/spoken response?

Panellists used the relevant CEFR scales and descriptors to evaluate the speaking and writing tasks. For oral and written responses, the Qualitative Features of Spoken Language and the Written Assessment Grid from the CEFR Companion Volume (Council of Europe, 2020) were applied. Following Kanistra (forthcoming), the following procedures were added either during the standard-setting workshops or in preparation for them:

1. **Preparation stage:** This was used to enhance the reliability and transparency of the standard setting workshops and to externally validate the linkage between tasks, criteria, and CEFR levels. To achieve this, the Trinity academic team used a modification of the *Dominant Profile* Method (Plake, Hambleton, & Jaeger, 1997) to map the ISE Digital speaking and writing assessment criteria to the CEFR. This activity enabled the team to select the CEFR scales and descriptors that better aligned with the examination construct, establish score profiles aligned with the CEFR levels (A1-C2), and estimate the expected cut scores.
2. **Range-finding:** Techniques described in the *Body of Work* method (Kingston & C. Tiemann, 2012) were used to reduce panellist fatigue. This step involved pre-identifying acceptable score ranges that aligned closely with the targeted CEFR level. Candidate performances that were clearly outside this range (eg far below the expected cut score) were excluded from the benchmarking process. This strategy reduced the mental load on panellists and ensured they focused on relevant scripts, improving the quality and consistency of their judgments.
3. **Pinpointing task:** This step is described in the Body of Work method (Kingston & Tiemann, 2012) and was incorporated into the panellist judgement task. It involved incorporating more scripts that received the same score as the expected cut scores, thus enabling panellists to validate and confirm their decisions. The pinpointing task ensures that cut scores are robust and defensible, allowing panellists to confirm their judgments systematically.
4. **Response ordering:** The candidates' written and spoken responses were carefully sequenced to account for contrast effects, as the order of presentation can influence panellist judgments due to the natural tendency to compare performances. To minimise this effect, the responses identified through the range-finding technique were arranged in ascending order, from lower to higher scores, with several tied responses (those receiving the same score) included in the sequence. These ties acted as a means of pinpointing tasks, enabling panellists to validate their decisions and mitigate any biases introduced by the response order, thereby ensuring greater consistency in panellist judgments.
5. **Orientation stage:** At the start of each workshop series for the speaking and writing modules, the module developers presented the construct of the speaking and writing module (as relevant) to the panellists. Additionally, the panellists were given a summary

of the construct, which they used as a reference alongside the CEFR descriptors when mapping the tasks to the CEFR.

6. **Test familiarity:** Before the standard-setting workshops, panellists were asked to complete the speaking and writing modules as candidates to better understand the cognitive and linguistic demands of the tasks. This firsthand experience provided deeper insights into the KSAs required, helping panellists assess task difficulty more accurately and reducing potential bias in cut score decisions.

The standard-setting workshops were conducted online via the Adobe Connect platform. The workshops included both synchronous and asynchronous sessions and spanned several days to enable panellists to complete the tasks. In line with the Manual (Council of Europe, 2009), Hambleton, Pitoniak, and Copella (2012), Pitoniak and Morgan (2012), Finch & French (2019), and Kanistra (forthcoming), the standard setting and benchmarking workshops comprised four stages:

- **Stage 1:** Introduction and Orientation
- **Stage 2:** Panellist familiarisation with the CEFR, test construct, and the speaking and writing modules of the ISE Digital examination
- **Stage 3:** Training in the standard setting method
- **Stage 4:** Standard setting and benchmarking of candidate speaking and writing responses.
- To assess the standard-setting procedures, formal and systematic data collection processes were implemented through evaluation questionnaires (Cizek, 2012), which were administered after Stages three and four. The questionnaires in Stage three were reviewed before Stage four, and relevant comments were provided to panellists prior to beginning the next stage. The standard setting panel included both internal panellists from Trinity and external language assessment experts. The two sub-panels remained separate throughout the workshops. Figure 2.7 illustrates the different stages of the workshops, indicating whether the activities were conducted synchronously or asynchronously. It should be noted that for all asynchronous activities, the facilitator remained on standby, logged into the virtual meeting room to respond promptly to panellist queries.

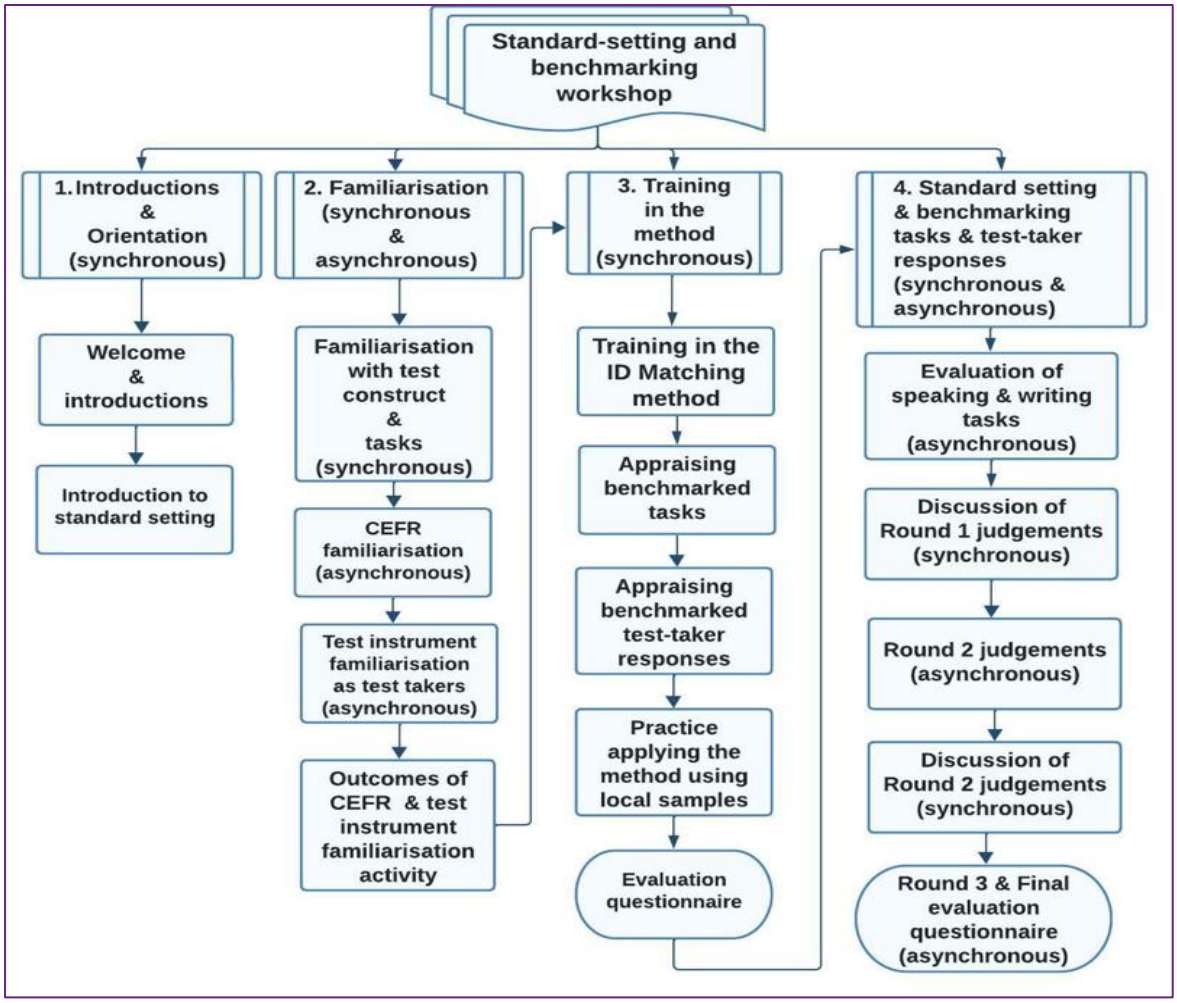


Figure 2.7: Overview of the speaking & writing workshops in the ID Matching method

2.5 Standard Setting Panellists

Standard setting panellists' judgements are central to the outcome of the calibration study, and it is widely acknowledged that panellist selection criteria are of utmost importance. Various standard setting researchers and practitioners explored the role of standard-setting panellists in the alignment studies and have suggested guidelines on the requirements for selecting a balanced and representative panel. (Berk, 1996; Cizek, 1996; Reckase, 2000; Kane, 2001; Hambleton, 2001; Raymond & Reid, 2001; Kaftandjieva, 2004; Hambleton and Pitoniak, 2006; and Cizek and Bunch, 2007).To date, the guidelines suggested by Raymond and Reid (2001, p. 130) remain the most comprehensive, and this study drew on these guidelines. Panellists were required to meet the following requirements:

- be subject matter experts
- be familiar with the level of the test-taking population
- collectively represent all relevant stakeholders
- have knowledge of the instruction (classroom or otherwise) to which test candidates are exposed
- appreciate the consequences of the standards

Additionally, panellists were required to be familiar with the CEFR and the level descriptors for each skill, which would expedite the overall ISE benchmarking process.

As might be expected, it is unlikely that all panellists will meet these requirements, particularly subject matter experts who represent a diverse constituency of stakeholders, including teachers on ISE preparation programmes, parents of candidates, and educational managers in various markets. To counteract differences in panellist expertise, Berk (1996, p. 222) suggests

that two panels could be identified, one comprising lay-person stakeholders and the other comprising subject matter experts. Each panel would contribute to different aspects of the cut-score setting process. The lay-person stakeholders would contribute at an initial stage, setting the expectations of different groups about the consequences of standard setting. Later in the standard setting process, they would offer their views on the plausibility of the proposed cut scores. The subject matter experts would fulfil all other stages of the benchmarking study. However, this approach still does not counter the logistical and practical difficulties in ascertaining comprehensive coverage of stakeholder representation.

Therefore, the panellists in this study were all subject-matter experts familiar with the level of the test-taking population. Table 2.1 summarises the basic information of the panellists who took part in the Writing and Speaking standard setting and benchmarking workshops.

Table 2.1: Overview of panellist status and expertise

Panellist ID	Status	Expertise
J01	External	Language assessment expert
J02	External	Language assessment expert, EFL teacher
J03	External	Language assessment expert, Item reviewer
J04	External	Language assessment expert
J05	External	Language assessment expert
J06	External	Language assessment expert
J07	External	Language assessment expert
J08	External	Item writer/ reviewer
J09	External	Language assessment expert
J10	External	Item writer/ reviewer
J11	Internal	Technical Content
J12	Internal	Technical Content
J13	Internal	Technical Content
J14	Internal	Examiner/ Examiner trainer
J15	Internal	Rater/ Senior rater

In accordance with the Manual (2009, p.42) and to ensure that they still represented as varied a group of stakeholders as possible, the panel comprised judges from both inside and outside the organisation (indicated in the table as *status*) and represented the different stages in language testing development as well as areas of expertise. As such, the group was drawn from Trinity's examiner panel, examiner trainers, academic consultants, and research staff. As recommended in the Manual, 15 panellists were invited (2009, p. 49); 5 of the panellists were internal experts representing key stages of the development process from item creation to assessing candidate written and spoken responses. Two panellists were active examiners, two were freelance item writers and reviewers, and three were active item reviewers as part of their wider roles at Trinity. The remaining eight were external experts. The internal and external panellists were kept separate during all phases of the speaking and writing standard-setting workshops.

3 Post-hoc Validation Methods

This section describes the types of analyses conducted to validate the ISE Digital CEFR alignment study outcomes. The indices described in this section were rigorously applied to evaluate classification decisions, underscoring the significance of methodological rigour and cut score analyses in CEFR alignment studies.

3.1 Framework for Evaluating Standard Setting Workshops

Several frameworks (Cizek & Earnest, 2016; Hambleton, Pitoniak, & Copella, 2012; Hambleton & Pitoniak, 2006; Kane, 1994) exist for evaluating standard-setting workshops. This study followed an adaptation by Kanistra (forthcoming) of Cizek and Earnest's framework (2016) to evaluate the CEFR alignment for all skills. Table 3.1 details this evaluation model.

Table 3.1: Framework for evaluating standard setting workshops (Kanistra, forthcoming)

Evaluation element	Description
Procedural	
Explicitness	○ The extent to which the standard setting purpose and process was clearly communicated to and understood by panellists
Practicability	○ The extent to which it was easy for the panellists to apply the standard setting method procedures. ○ The extent to which it was easy for the panellists to record their judgements.
Implementation	○ The extent to which the standard setting procedures were reasonable, and methodically implemented (familiarisation with the CEFR, test instrument, training in the method).
Feedback	○ The extent to which panellists reported to have confidence in their ability to apply the standard setting procedures, had confidence in their ratings and in their recommended cut scores.
Documentation	○ The extent to which the standard setting procedures are informed by the literature and are carefully documented. ○ The extent to which the data are carefully analysed from different perspectives.
Internal	
Intra-panellist consistency	○ The extent to which panellists' ratings are congruent with the empirical item difficulties or scores awarded. ○ The extent to which the CEFR item judgements are congruent with the rationalisation of the item judgements ○ The extent to which panellists' ratings are congruent between rounds. ○ The extent to which panellists' ratings are congruent with the severity/leniency they exhibit when appraising items.
Inter-panellist consistency	○ The extent to which panellists' ratings are consistent with each other. ○ The extent to which panellists are appraising items as a homogenous group and their ratings are comparable. ○ The extent to which panellists' ratings are independent and in accordance with the expectations of the Rasch model.
Consistency within the method	○ The extent to which the recommended cut scores are precise and do not negatively impact the reliability of the test instrument. ○ The extent to which two subgroups of panels differentiated by distinct characteristics (ie internals, externals) recommend consistent cut scores.
Decision consistency	○ The extent to which the recommended cut scores classify candidates as 'masters' and 'non-masters' consistently.
External	
Comparisons to other standard setting methods	○ The extent to which the cut scores from different methods are consistent and comparable. ○ The extent to which panellists participating in different standard setting workshops offer consistent judgements.
Comparisons to other information	○ The extent to which the pass rates from the recommended cut scores are in line with the pass rates of other test instruments at the same CEFR level.
Reasonableness of cut scores	○ The extent to which the recommended cut scores are reasonable. ○ The extent to which the recommended cut scores are in line with the panellists' judgemental task. ○ The extent to which the panellists relied on CEFR descriptors in the discussion stage.

For the listening and reading modules, internal validity was assessed through consistency within the method and decision consistency. The speaking and writing standard setting workshops were analysed for evidence of procedural, internal, and external validity. Table 3.2 illustrates the framework for evaluating the engineered cut scores and the speaking and writing standard-setting workshops.

3.2 Inter-panellist and Intra-panellist Consistency

The reliability, consistency, and agreement among judges in this benchmarking study were evaluated using Rasch Measurement Theory (RMT). The MFRM model has been used in standard setting and alignment studies to evaluate panellist rating consistency and congruence in their CEFR judgements (Eckes, 2009; Engelhard, 2009; Kanistra, forthcoming; Kollias, 2023; Papageorgiou, 2009; Kanistra & Kollias, 2024). Engelhard (2009, p. 314) defined the MFRM model that operationalises the conceptual model of standard setting and benchmarking studies as follows:

$$ln\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = \beta_n - \delta_i - \omega_j - \tau_k \qquad \text{Equation 1}$$

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 judgement of minimal competence required to pass for judge n ,

δ_i judgement of difficulty for an item i,

ω_j judgement of performance standards for round j, and

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

RMT enables the evaluation of intra-panellist and inter-panellist consistency at the individual and group levels. Inter-panellist consistency was evaluated using the following indices:

- Panellist severity measure;
- Most lenient (min) and most severe panellist's fair average (max);

The severity measures indicate how panellists scored the scripts on average, with positive logit values representing stricter ratings and negative values reflecting leniency. The fair average reports the expected raw score in the absence of severity or leniency, thus facilitating the evaluation of each panellist's impact on scoring consistency.

- The overall single-panellist rest of panellist (*SP/ROP*) (point-measure) correlation coefficient;
- Each panellist's SP*/ROP;*

Inter-panellist consistency in RMT can be evaluated through the single panellist/rest of panellists' point measure correlation (*SP/ROP*), which is both an individual and a group-level statistic. It is a metric similar to the Pearson product-moment correlation. It measures inter-panellist consistency by comparing a panellist's ranks and ratings to how the rest of the panellists collectively rank and score those same items. It checks whether a panellist's scoring aligns with the group's consensus. Values above 0.70 indicate strong alignment, while values below 0.30 suggest inconsistency (Myford & Wolfe, 2004a; Linacre, 2020). *SP/ROP* values near zero or negative indicate that the panellist's scoring is inconsistent with the group's consensus, potentially highlighting significant differences in judgment. Furthermore, FACETS software (Linacre 2024a) calculates the expected *SP/ROP* correlation values, serving as a benchmark for comparison with the observed data. When the observed *SP/ROP* aligns with the one predicted by the Rasch model, it corroborates the inter-panellist consistency.

- Overall exact agreement observed % and expected agreement%;
- Each panellist's exact agreement observed % expected agreement %

Inter-panellist agreement can be analysed using the *observed % agreement* and the *agreement % expected*, as calculated by *Facets*. These metrics operate at individual and group levels, assessing the degree to which panellists' ratings agree. The *observed % agreement* represents the proportion of instances where a panellist's CEFR evaluations exactly match those of another panellist. In contrast, *agreement % expected* reflects the proportion of exact matches anticipated if the panellists' judgments aligned perfectly with the Rasch model's predictions. For trained raters, the observed percentage is typically slightly higher than the expected percentage. When the *observed* and *expected % agreement* are closely aligned, it suggests that panellists are operating as independent experts, autonomously applying their judgment to the appraisal of scripts. However, a lower *observed % agreement* than the expected one may indicate insufficient training (Linacre, 2024). In contexts such as benchmarking or alignment studies, where panellists undergo extensive training to achieve a shared understanding of the CEFR descriptors, a slightly higher observed agreement over the expected one is to be expected and might even be desirable (Kanistra, forthcoming; Kanistra & Kollias, 2024). Within the RMT framework, observed agreement percentages exceeding 90% or those significantly higher than their expected values can signal potential issues. This is particularly relevant in scenarios where panellists feel compelled to conform to one another's judgments or are encouraged to act as mechanical scorers, adhering strictly to predefined principles without exercising their professional expertise. Such circumstances may undermine the panellists' autonomy and the application of their expert judgment (Linacre, 2024a).

- Overall Rasch *kappa;*
- Individual panellist Rasch *kappa*

Rasch kappa, a measure of agreement among panellists, is the Rasch version of Cohen's kappa. It indicates the extent to which panellists agree on the exact classification of items. An ideal value of Rasch kappa is close to 0, suggesting the panellists exhibit the right amount of agreement whilst maintaining their independence as experts. Values above 0 indicate more agreement, while values below 0 indicate disagreement. Mojtaba Taghvafard's research suggests that Rasch kappa values between -0.2 and +0.2 indicate expected agreement by the model. Values between |0.20| and |0.40| show slightly more agreement than expected. In contrast, values greater than or equal to |0.50| indicate very high agreement, suggesting that panellists are appraising items as rating machines. This scenario can be problematic, as it may indicate panellist dependence in a standard-setting context (Eckes, 2009). Rasch kappa is not directly reported in an MFRM analysis, but it can be calculated using Equation 2.

$$Rasch\ kappa = \frac{(Observed\% - Expected\%)}{(100 - Expected\%)}$$ Equation *2* *(Linacre, 2024)*

- Infit Mean-square (Infit *Mnsq*)
- Infit z standardised (Infit *Zstd*).

The Infit Mean-square (Infit *Mnsq*) and Infit z-standardised (Infit *Zstd*) indices serve as both individual and group-level statistics, with an expected value of 1 and a range extending from 0 to ±∞. These indices evaluate the degree to which observed ratings align with predictions generated by the Many-Facet Rasch Measurement (MFRM) model. Infit and outfit values near 1 indicate that observed ratings align well with model predictions. Values below 1 (overfit) suggest greater predictability than expected by the MFRM, while values exceeding 1 (misfit) indicate deviations that are less predictable and harder to explain (Myford & Wolfe, 2004a). Among these, misfit is generally more concerning than overfit, as it represents more substantial deviations from expected ratings. Linacre (2020) highlights that low Infit Mnsq values can signify high intra-rater reliability, as they reflect a panellist's consistent and predictable judgment patterns. Pollitt and Hutchinson (1987) emphasised the importance of interpreting infit and outfit indices within the specific analytical context. Accordingly,

acceptable ranges for these indices are often calculated as the Infit mean ± 2 standard deviations (SD). The Infit *Zstd* indices complement the Infit *Mnsq* by reporting the statistical significance of unexpectedness in the data. For small samples, such as the one analysed in this report, Infit *Mnsq* indices with *Zstd* values ≥ 2.00 are considered statistically significant, while values ≥ 2.6 are deemed highly significant (Bond & Fox, 2015; Engelhard, 2009, 2013; Linacre, 2002; Myford & Wolfe, 2004a; Wolfe & Smith, 2007; Yu, 2020).

3.3 Consistency Within the Method

Consistency within the method is another crucial aspect of the internal validity of a standard-setting workshop. It suggests that if a different panel of experts were convened to conduct another standard-setting workshop using the same or even a different method, they would be likely to achieve comparable outcomes. Considering the intricacies of standard setting and the potential for varying results across different studies, quantitative processes have become the standard for evaluating standard-setting practices. One way of investigating the accuracy and consistency of the standard setting method is by i) evaluating the cut scores in terms of their precision, accuracy, and reliability, and ii) evaluating the candidates' classification consistency and accuracy based on these cut scores.

One way to investigate the precision of the cut scores is by calculating the standard error of the mean of panellist judgements (SE_j) and comparing it to the standard error of measurement (*SEM*) of the test instrument (Council of Europe, 2009). In the ISE Digital-CEFR linking project, when a group of panellists convened, the standard error of the mean of panellist judgements (SE_j) was calculated under the Central Limit Theorem (CLT) using Equation 3. This equation estimates the standard error of the cut score using the population standard deviation of the panellist judgements (SD_j) divided by the square root of the number of panellists minus 1 (n-1):

$$SEj = \frac{SDj}{\sqrt{(n-1)}} \qquad \text{Equation 3}$$

When threshold regions and cut scores are set unanimously or engineered, the SD_j and SE_j equal 0. This occurs either due to the absence of variation in panellist judgments or the absence of a panellist group rather than indicating error-free cut scores (MacCann & Gordon, 2004). To address this issue, elements from the methodology of calculating a cut score in the *Bookmark* method were adopted (Cizek & Bunch, 2007). Specifically, the ability estimates (β_v) of candidates on the ISE Digital reading and listening modules, who had ability estimates within the engineered cut scores range, were used to calculate the standard deviations (SD_{jtt}), which were subsequently used in Equation 1 to calculate the error of the person mean which acts in this context similar to the error of panellist judgments (SE_{jtt}).

The literature varies with respect to what is acceptable in the relationship between the standard error of judgments (SE_j) and the *SEM* of the test. Cohen et al. (1999) suggest that an SE_j less than half the *SEM* of the test has a minimal impact on candidates' misclassifications. Jaeger (1991) recommends that the mean error of judgements should not be larger than a quarter of the *SEM* of the test, so that the impact of the additional error would not be greater than 3%. Kaftandjieva (2010) suggests an SE_j smaller than or equal to a third of the *SEM* of the test as a more practical standard, as it can be achieved with 15 panellists. This criterion was adhered to in this study.

Furthermore, the influence of the standard error of panellist judgments (SE_j) was examined in relation to the standard deviation of the candidate population. This was done because the *SEM* reaches its maximum when it equals the standard deviation of the observed scores (*SD* population). Consequently, the standard error of panellist judgments was considered appropriate if it was smaller than a third of the standard deviation of the candidate population (Kanistra, forthcoming; Sireci et al., 2008).

Additionally, the cut scores of the reading and listening modules (receptive skills) were evaluated using the conditional standard error of measurement (*CSEM*), which is the standard error of measurement (*SEM*) at the cut score point on the logit scale (Sireci et al., 2008). What is more, the accuracy of the location of the cut score was evaluated using the conditional reliability (*CREL*) of the recommended cut score. The *CREL* was calculated using Equation 4

(Nicewander, 2019, p. 15), where *I(X,θ)* is the score information function found in the test characteristic curve file (*TCCFILE*) provided by the software program Winsteps (*version 5.8.3*, Linacre, 2024).

$$\rho(X, X'|\theta = \frac{I(X,\theta)}{1+I(X,\theta)} \qquad \text{Equation 4}$$

Foreign language proficiency test scores are generally considered acceptable when they fall within the range of .80 – .90 (Nicewander, 2018, 2019). Therefore, a cut score is deemed appropriate when its *CREL* falls within the recommended range.

Cut scores are crucial in classifying test items and candidate performance into different CEFR levels (A1, B2, C1, etc.). Evaluating cut scores involves assessing the reliability and validity of classification decisions if the same candidates sat two parallel test administrations, using indices like classification consistency [*DC(φ)*] and decision accuracy [*DA(γ)*] to measure classification reliability and alignment with 'true' classifications when measurement error issues were factored in (Kaftandjieva, 2010; Kane, 1994; Deng & Hambleton, 2013; Lee, Hanson, & Brennan, 2002). Additional metrics, such as misclassification rates, false positive/negative rates, and Cohen's kappa (*κ*), assess the consistency and accuracy of these decisions. The location of cut scores, test length, and test reliability significantly affect classification indices, with scores near the test-score distribution mean typically showing lower decision accuracy and consistency (Subkoviak, 1988; Huynh, 1976, 1990). These indices are calculated via tools like *BB-CLASS v1.1* and *IRT-CLASS v2 using the* Livingston and Lewis (1995) CTT-based approach, denoted as LL, and Lee's (2008) IRT-based models, respectively.

Classification consistency [DC(φ)] and the kappa (κ) coefficient provide distinct insights. Notably, classification consistency [DC(φ)] reaches its peak at the extremes of the test score distribution—either very high or very low scores—because candidates at these extremes are more distinguishable, resulting in fewer classification errors. Conversely, it is lower near the centre of the test score distribution, where classification ambiguity is greater due to overlapping performance levels and narrower score differentials (Subkoviak, 1988; Huynh, 1976). Kappa (κ), on the other hand, reflects chance-corrected consistency and peaks at the test score distribution's centre rather than its extremes. It is influenced by test reliability, cut score placement, and score variability (Huynh, 1976, 1990). Chance consistency (pchance, φ_c) shows the proportion of consistent classifications expected by chance if the outcomes of the second administration were completely independent of the outcomes of the first administration.

3.4 Data Organisation

To facilitate quantitative analyses, when a panellist group convened, the panellist CEFR judgments were coded as shown in Table 3.2, ranging from 0.5 (Pre-A1) to 6 (C2). The plus levels (i.e., A1+, A2+, B1+, B2+) judges assigned were quantified as an average of the two adjacent scores. For example, as Table 3.2 shows, an A2 judgement was coded as a 2, and a B1 as a 3; thus, A2+ was coded as 2.5.

CEFR level	Assigned numeric value
C2	6
C1	5
B2+	4.5
B2	4
B1+	3.5
B1	3
A2+	2.5
A2	2
A1+	1.5
A1	1
Pre-A1	0.5

Table 3.2: Numeric value assigned to each CEFR level

4 Familiarisation Methodology and Outcomes

The purpose of the familiarisation activities was to encourage panellists to re-familiarise themselves with the CEFR scales and descriptors aligned with the construct of the speaking and writing modules. Following procedures developed by Kanistra (forthcoming), a number of non-scored activities and scored quizzes were created that prompted panellists to use top-down (refer to the overall descriptors to do the activities) and bottom-up techniques (refer to the key concepts and read the descriptor carefully). All activities were created in Trinity's learning management system (Totara, https://www.totara.com/).

The activities were grouped into three broad categories:

1. identification of CEFR scales relevant to the test construct
2. non-scored refamiliarisation activities
3. scored quizzes

The non-scored refamiliarisation activities preceded the scored quizzes. The refamiliarisation tasks prompted panellists to employ top-down approaches by allowing them to refer to the *Overall oral production*, *overall oral interaction, overall written production,* and *overall written interaction* scales, respectively, while attempting the quizzes. The descriptors within some of the quizzes were presented in ascending order of difficulty (easy to difficult), and panellists were asked to consider the question underpinning the *ID Matching* method:

> *"What makes each descriptor more difficult than the previous one?"*

These activities entailed asking the panellists to read each scale's key concepts carefully and either select the words in the descriptors that best reflected these key concepts or match the descriptor with the corresponding CEFR level. Figures 4.1 and 4.2 illustrate examples of activities that encourage panellists to adopt a top-down approach.

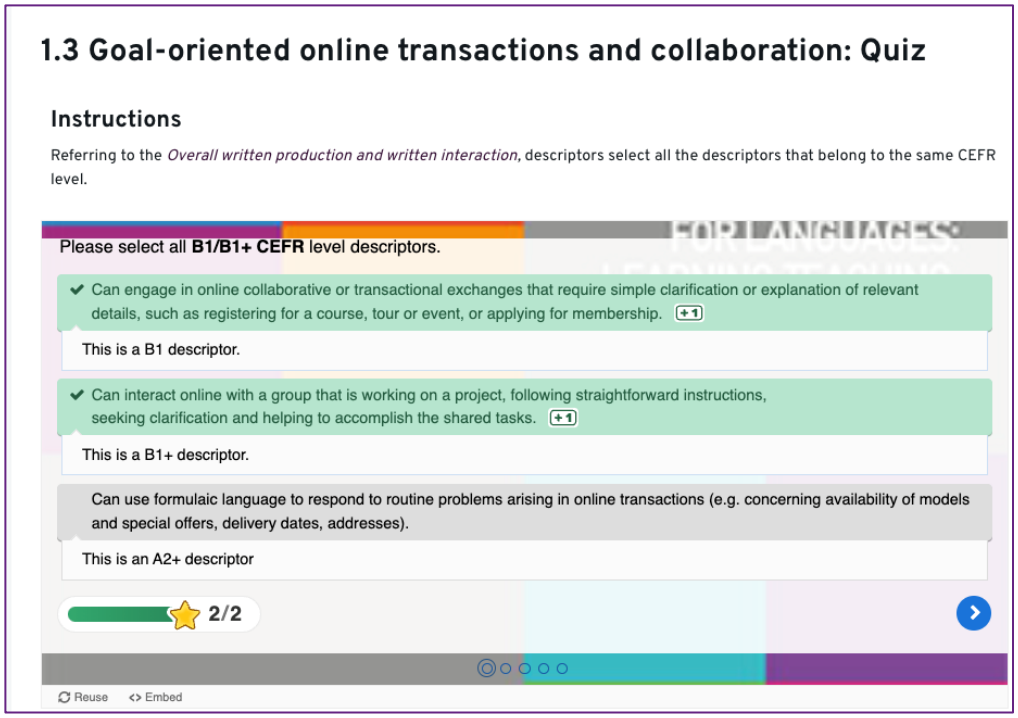


Figure 4.1: Example of top-down CEFR familiarisation activity for Speaking

1.1 Conversation: Task

Instructions

Open the *Overall oral production* and *Overall oral interaction* scales in a new window. Then read the descriptors and observe how language learners' proficiency progresses as the CEFR levels advance.

Please carefully read the key concepts operationalised in this scale before attempting the activity. This activity entails reading the different descriptors carefully and determining their CEFR level before turning the card to reveal the answer. The purpose of this activity is to facilitate a focused reading of the different CEFR-level descriptors.

Note: You will need to include plus (+) levels. CEFR descriptors range from A1 to C2. The picture in the flashcard is just for decoration.

Conversation

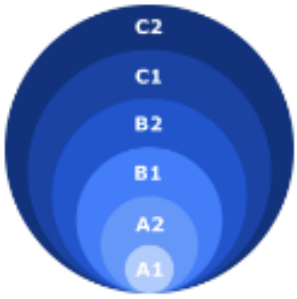
Conversation concerns interaction with a primarily social function: the establishment and maintenance of personal relationships.

Key concepts operationalised in the scale include:

- setting: from short exchanges, through maintaining a conversation and sustaining relationships, to flexible use for social purposes;
- topics: from personal news, through familiar topics of personal interest, to most general topics;
- language functions: from greetings, etc, through offers, invitations and permission, to degrees of emotion and allusive, joking usage.

Progression up the scale is represented by a movement from simple, factual conversation and social exchanges on familiar topics, to relatively long conversations on topics of general interest, to extended conversation and engagement demonstrating nuance and flexibility in the use of language.

Instructions
The descriptors are arranged in ascending order (from A1 to C2) and descriptors for plus levels are included.
Keep the following question in mind while completing this task:
What makes each descriptor more difficult than the previous one?



01. Can understand everyday expressions aimed at the satisfaction of simple needs of a concrete type, delivered directly to them in clear, slow and repeated language by a sympathetic interlocutor.

02. Can take part in a simple conversation of a basic factual nature on a predictable topic (e.g. their home country, family, school).

Turn

Figure 4.2: Example of top-down CEFR familiarisation activity for Writing

Other quizzes included the descriptors in jumbled order, and panellists were asked to order the descriptors in ascending order of difficulty. Panellists could refer to the key concepts operationalised in the relevant scale to help them with the ordering task. Once again, they were asked to evaluate their ordering by considering the question underpinning the *ID Matching* method: "What makes each descriptor more difficult than the previous one?". Such activities encouraged panellists to adopt bottom-up techniques; an example of such an activity can be seen in Figure 4.3. All the activities indirectly familiarised the panellists with the *ID Matching* method since the panellists were becoming accustomed to ordering descriptors in ascending order of difficulty and critically evaluating the difficulty of the knowledge skills and abilities (KSAs) described by each descriptor

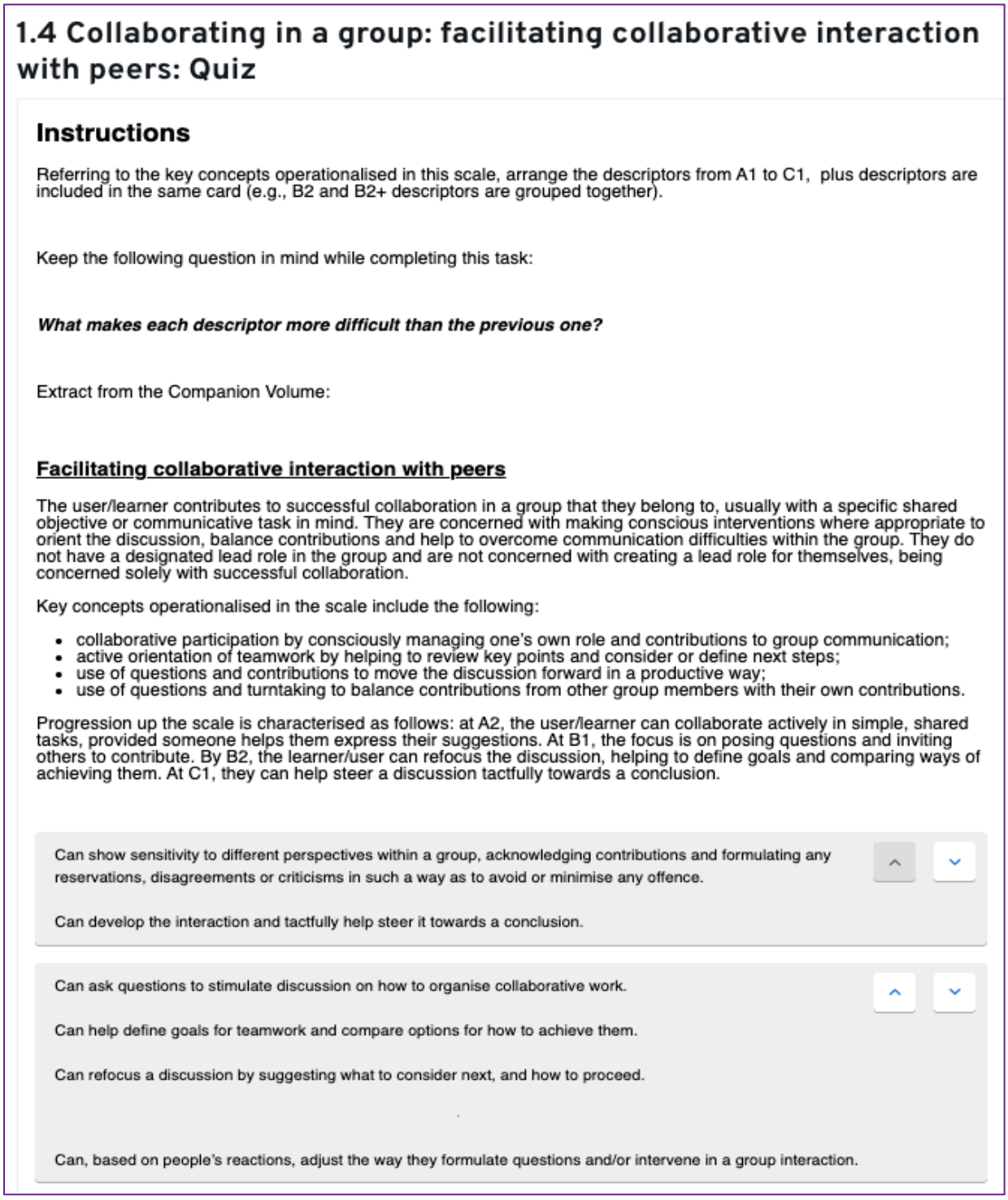


Figure 4.3: Example of bottom-up CEFR familiarisation activity for Writing

Panellists received immediate feedback on their performance thus allowing them to review their answers and critically evaluate their own understanding of the CEFR descriptors by reviewing their mistakes. For the recommended cut scores to be valid, panellists must be very familiar with the CEFR levels and must rank order CEFR descriptors appropriately. For this reason, the passing score for scored quizzes was informed by Cicchetti and Sparrow's guidelines (1981, as cited by Cicchetti, 1994) and set to 80%. Panellists received scoring feedback that would indicate whether their alignment to any of the CEFR scales was poor, weak, good, very good, or excellent (ie 0%-20% poor, 21% to 40% weak, between 41% to 60% good, 61% to 79% very good and between 80% and 100% excellent). Panellists were asked to redo a task until they achieved a score within the required range (80%-100 %) before being allowed to proceed to the next task. For the non-scored tasks, panellists could use the Totara chat function to discuss with each other the descriptors they found most challenging to identify or to order their level correctly. Each panellist could start their own discussion topic, and the rest of the panellists could make a maximum of one contribution to each topic.

Tables 4.1 and 4.2 present the results of the scored familiarisation activities for the Speaking and Writing modules, respectively. The platform was set to record the percentage of correct responses on the panellists' first attempt, as it was deemed desirable to evaluate their initial self-reported CEFR expertise level. Table 4.1 shows that the panellists exhibited the desirable CEFR familiarisation levels even before refamiliarising themselves with the pertinent scales. In the discussion forum, J05 explained their low score of 33% in Activity 2 (CEFR scales relevant for summarising a talk/conversation task), stating that they felt other scales (not given as an option) were more appropriate. This comment was addressed at length before the standard-setting workshop, where panellists could discuss the interplay between the test construct and the CEFR, as well as their experiences from taking the speaking module as candidates.

Overall, all panellists demonstrated the appropriate level of expertise in the Writing CEFR familiarisation activities. The *online conversation and discussion* descriptors posed challenges for J10, while the *Relaying specific information* descriptors posed challenges for a few panellists. Panellists on the discussion forum primarily attributed their difficulties to the *"thin lines between plus levels"* (J02). Most panellists actively participated in this thread, with J04 providing a concise summary of these challenges:

"In general, the plus levels are challenging because they combine elements from the levels just above and below. While I feel comfortable with the key terms of the six CEFR reference levels, I sometimes struggle when there's a blend between two levels."

However, the platform's setup ensured that all panellists achieved an acceptable percentage of correct answers before proceeding to the standard-setting tasks. These outcomes demonstrated that the panellists were thoroughly familiar with the relevant CEFR scales and descriptors, thus ensuring their readiness to participate in the standard-setting workshop.

.

Table 4.1: Panellist performance on the familiarisation activities – speaking module

Panellist ID	CEFR activities & scales: speaking module							
	1a. CEFR scales relevant to the *responding to questions* and *delivering a prepared talk* tasks	**1b. CEFR scales relevant to the *interacting* task**	**1.2 Sustained monologue: describing experience**	**1.3 Sustained monologue: giving information**	**1.4 Sustained monologue: putting a case**	**2. CEFR scales relevant for the *summarising a talk/conversation* task**	**2.1 Conversation**	**2.2 Obtaining goods and services**
J01	100	100	100	90	100	83	100	86
J02	100	100	100	80	100	100	100	100
J03	100	100	100	90	100	100	100	93
J04	100	100	100	80	100	100	100	86
J05	100	75	100	100	100	33	100	100
J06	100	100	100	80	100	100	100	86
J07	100	100	100	90	100	100	100	79
J08	100	100	100	100	100	100	100	100
J09	100	100	100	90	100	100	100	93
J10	100	100	100	90	100	100	100	86
J11	83	100	100	100	100	100	100	100
J12	100	100	100	100	100	100	100	100
J13	100	100	100	100	100	100	100	93
J14	100	100	100	90	100	100	100	100
J15	100	100	100	100	100	100	100	100

Table 4.2: Panellist outcomes of Writing Familiarisation activities

Panellist ID	CEFR activities & scales: writing module								
	1. **CEFR areas relevant to the *written online communication* task**	**1.2** **Online conversation & discussion**	**1.3** **Goal-oriented online transactions & collaboration**	**1.4** **Collaborating in a group: facilitating collaborative interaction with peers**	**1.5** **Collaborating in a group: Collaborating to construct meaning**	**2.** **CEFR areas relevant to the *writing from sources* task**	**2.2** **Relaying specific information**	**2.3** **Explaining data in writing**	**2.4** **Processing text in writing**
J01	86	91	100	100	100	80	85	100	88
J02	100	100	100	100	100	100	54	100	88
J03	100	86	70	100	100	100	85	100	82
J04	100	91	100	100	100	100	100	100	82
J05	100	91	100	100	100	100	69	100	88
J06	100	95	100	100	100	100	77	100	88
J07	100	82	100	100	100	100	77	80	71
J08	100	95	100	100	100	100	100	100	94
J09	100	95	100	100	100	100	85	100	88
J10	100	23	100	100	100	100	85	100	88
J11	100	100	100	100	100	100	100	100	100
J12	100	95	90	100	100	100	100	100	100
J13	100	95	100	100	100	100	85	100	88
J14	86	95	100	100	100	100	100	80	82
J15	100	86	100	100	100	100	85	80	94

5 Validating the Speaking Standard-Setting Workshop and Cut Scores

This section presents the results, adding validity evidence to the procedural, internal, and external aspects of the evaluation framework discussed in section 3.1 for the speaking module of the ISE Digital examination.

5.1 Psychometric Properties of the ISE Digital Speaking Module

To analyse the speaking module, a Many-Faceted Rasch Measurement (MFRM) analysis was conducted using *Facets v4.4.4* (Linacre, 2025). The MFRM analysis included 349 candidates and 81 task-level observations. Although 81 tasks were calibrated, it should be noted that each operational speaking test included four tasks (as per the specifications of the speaking module), with overlap across forms to ensure subset connectivity. The results of this analysis are summarised in Table 5.1.

Table 5.1: Rasch summary statistics for the ISE Digital speaking module

Index	**Real (N = 349)**
Number of tasks	81
Candidate mean measure (*SEm ; SD*)	0.09 (0.46; 2.997)
Test reliability	0.97
RMSE (CSEM)	0.48
Observed average (*SD*)	2.85 (0.91)
SEM	1.86
Fair average (*SD*)	2.80 (0.94)

The candidate mean measure was 0.09 logits with a standard error of mean (*SEm*) of 0.46 and a standard deviation (SD) of 2.99. The relatively large *SD* indicates considerable variability in candidate speaking ability, showing that the speaking tasks effectively distinguished a broad range of proficiency levels. Reliability was high (0.97), demonstrating strong separation of candidate abilities, with an RMSE (CSEM) of 0.48 logits and an SEM of 1.86, reflecting acceptable precision for a performance-based assessment. The close match between the observed mean score (2.85; SD = 0.91) and the fair average (2.80; SD = 0.94) suggests that examiners applied a comparable amount of severity and marked comparably across tasks and forms, aligning with the expectations of the Rasch model.

Overall, the indices demonstrate that the ISE Digital speaking module functions as a stable and reliable measure of oral proficiency, supporting its use in the CEFR standard-setting procedure described in the following sections.

5.2 Procedural Validity

The evaluation questionnaires were adapted from Cizek (2012, pp. 174-178). To align with the context of this study, some questions were modified. The surveys were administered after the *orientation and training-in-the-method* stages of the speaking standard-setting workshop.

5.2.1 Evaluating the *orientation* and *training-in-the-method stages*

The panellists were asked to rate the extent to which they agreed with the 14 survey statements. Figure 5.1 presents the survey statements and the analyses of this evaluation questionnaire. The bar graph illustrates the number of panellists who endorsed or opposed each statement, with the axis indicating the total number of panellists. Before moving on to the next workshop stage, the facilitator reviewed the survey responses and addressed any reported issues before initiating the standard-setting tasks.

The evaluation results for the *orientation* and *training-in-the-method* stages of the speaking standard-setting workshop highlight panellists' strong preparedness and overall satisfaction. A significant majority either '*strongly agreed*' or '*agreed*' that the orientation session provided a clear overview of the workshop's purpose (Q1) and effectively addressed questions about the CEFR alignment and the ISE Digital speaking exam (Q2). Similarly, participants '*strongly agreed*' or '*agreed'* that the facilitator helped them understand the standard-setting and benchmarking process (Q3). The timing and pace of the orientation and training sessions were deemed appropriate (Q4). The minor reservations expressed by three panellists (J06, J09, and J10) were primarily due to their heavy workload. They used this question to communicate their requirement for more time to rate the candidates' speaking performances during the asynchronous part of the workshop.

The CEFR familiarisation activities also received strong positive responses, with many panellists '*strongly agreeing*' or '*agreeing'* that these activities provided a focused reading of the CEFR descriptors (Q5) and refreshed their knowledge of these descriptors (Q6). Furthermore, the majority reported having a good understanding of the CEFR levels (e.g., A1, A2, B1, etc.) and descriptors for *oral production*, *sustained monologue*, and *oral interaction* (Q7 and Q8), with responses predominantly falling into the '*strongly agree*' and '*agree'* categories.

Additionally, taking the test as candidates helped panellists better understand the difficulty, content, and other aspects of the speaking component of the ISE Digital examination (Q9), as strong positive agreement was noted. Participants also '*strongly agreed'* or '*agreed'* that the training in the standard-setting method was clear (Q10) and that the practice activities effectively helped them apply this method (Q11). As a result, they developed a good understanding of their role in the CEFR alignment and benchmarking activities (Q13). They expressed confidence in applying the standard-setting method effectively (Q14). Ultimately, most participants '*strongly agreed'* or '*agreed'* that they felt prepared to begin the task (Q15).

These findings underscore the success of the orientation and training sessions in fostering a deep understanding of CEFR descriptors and the alignment and benchmarking process to be followed while building participants' confidence to undertake these tasks effectively.

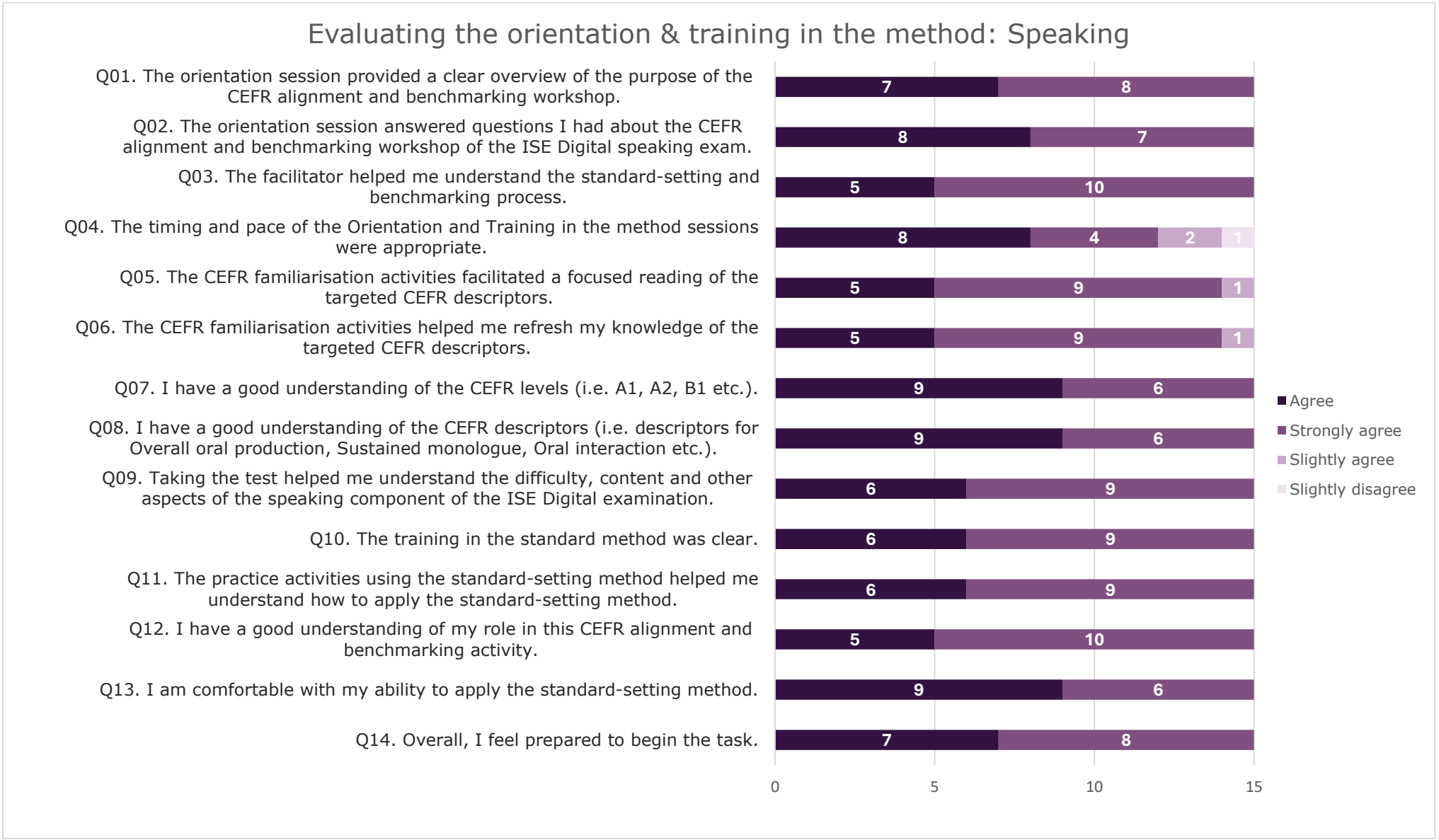


Figure 5.1: Evaluation of the orientation & training in the method stages – speaking

5.2.2 Evaluating the speaking *standard setting and benchmarking* workshop

Panellists were asked to rate their agreement with the eight statements included in the evaluation survey. The results, presented in Figure 5.2, show the distribution of panellist endorsements and oppositions for each statement. The bar graph visually represents the responses, with the axis indicating the total number of panellists. The last two questions of this survey served as Round 3, allowing the judges to reflect on and review the performances that were deemed representative of the different targeted CEFR levels. This enabled them to revise their judgments.

The survey results indicate overall positive feedback from the panellists regarding the speaking standard-setting workshop. Most panellists 'strongly agreed' or 'agreed' that the standard-setting procedures enabled them to map the ISE Digital speaking tasks and candidates' spoken responses to the targeted CEFR levels effectively (Q02 and Q03). The facilitator's role was highly valued, with the majority of panellists *strongly agreeing* or *agreeing* that the facilitator ensured that all panellists contributed to group discussions (Q04) and that no one person unfairly dominated the group (Q05). The panellists also expressed their confidence in their ratings, with a significant number 'strongly agreeing' or 'agreeing' that they felt confident in their ratings (Q01). Furthermore, panellists strongly agreed that they understood other panellists' ratings (Q06) and could effectively use those ratings to inform their judgments (Q07) when appropriate. Additionally, the final group-recommended CEFR classifications for the speaking exam were endorsed, with all panellists agreeing that they accurately represented the minimum levels of performance expected at the targeted CEFR levels (Q08 and Q09).

The final evaluation question required panellists to rank the factors influencing their judgments during the speaking standard-setting and benchmarking workshop in order of importance (see Table 5.1). These responses provided valuable insights into the decision-making processes and the factors the panellists prioritised in their decision-making process.

Table 5.2: Factors affecting panellists' judgements - speaking

Influential factors	Sum	Rank
Q9.1. My experience taking the test.	55	1
Q9.2. My own experiences with real students.	50	2
Q9.5. The group discussion.	46	3
Q9.3. The CEFR level descriptors & qualitative features of spoken language.	44	4
Q9.6. Other judges' ratings.	44	4
Q9.4. The candidates' oral responses.	34	5

When evaluating candidates' oral performances in the speaking standard-setting workshop, panellists identified their experience taking the test as the most influential factor (score: 55), followed by their own experiences with real students (score: 50). The group discussions that took place after the round 1 judgements and the CEFR level descriptors were deemed slightly more influential than the *qualitative features of spoken language* and the ratings from the other panellists (scores: 46 and 44 respectively), emphasising the role of collaboration with qualitative frameworks such as the CEFR in alignment and benchmarking studies.

The candidates' oral responses, which scored 34, were the least influential. While they were considered, they were used as the foundation for panellists' judgments, which were then informed by the CEFR, the panellists' collaboration with others, and their personal experience. These results demonstrate a balanced approach to standard-setting and benchmarking in the speaking domain, integrating practical experience, peer insights, and standardised criteria.

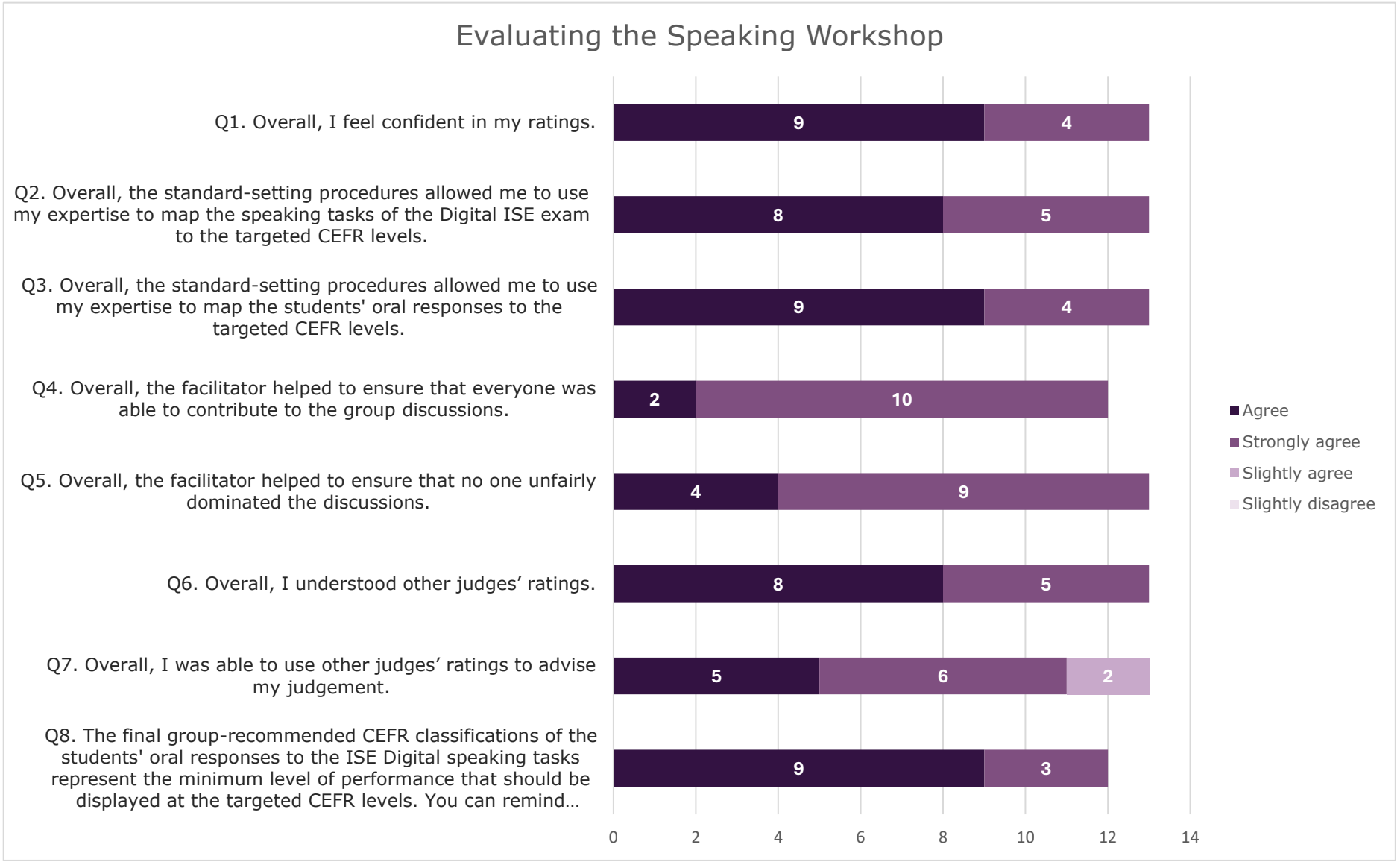


Figure 5.2 Evaluation of the standard setting and benchmarking stage – speaking

In conclusion, the evaluation of the *orientation* and *training-in-the-method* stages of the speaking standard-setting workshop reveals its overall effectiveness in equipping panellists with the confidence and essential knowledge required to make well-informed judgments. Panellists appreciated the workshop's clarity and inclusiveness, as well as the facilitator's role in promoting balanced discussions and collaboration. The *standard-setting* and *benchmarking* procedures implemented during the workshop streamlined the panellists' decision-making processes, ensuring their confidence in their ratings and the final recommended classifications. Consequently, it is reasonable to conclude that no systematic errors were introduced during the standard-setting process, which could have potentially invalidated the workshop's results.

5.3 Evaluating the Speaking Tasks

The development of speaking tasks for the ISE Digital examination adhered to the ECD and PADDI process described in Section 2.2 (Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006; Ferrara, Lai, & Nichols, 2016), aligning closely with the CEFR framework. Trinity's item creation procedures follow UATD principles (Kanistra, forthcoming), instructing item writers to design tasks that target the KSAs associated with specific CEFR levels, while ensuring that input materials meet the CEFR level requirements and readability indices. The outcomes of this principled approach to item creation are reflected in the content analysis forms provided in the Manual (Council of Europe, 2009, Appendix A).

The speaking Module of ISE Digital includes four task types: responding to questions (three questions, ascending in order of difficulty), delivering a prepared talk that includes an expansion question, interacting, and summarising a talk or conversation. Test content is tailored to different CEFR levels for all tasks. The expansion questions are designed to cater to different CEFR levels in delivering a prepared talk task. The module employs an adaptive format, directing candidates to one of three routes (A1-A2, B1-B2, or C1-C2) based on their performance in the routing section. Thus, the items are designed to span at least two adjacent CEFR levels, ensuring accessibility for lower-level candidates while allowing more proficient candidates to provide more complex responses.

Additionally, the speaking tasks increase in complexity, with *summarising a talk/conversation* being the most challenging, as this task requires mediation skills (drawing on listening and speaking skills). Panellists analysed the tasks' cognitive and linguistic demands, mapped them to appropriate CEFR levels, and justified their judgments by referencing specific CEFR scales (Harsch & Kanistra, 2020). Panellists were asked to evaluate three speaking tests, each aiming at the three routes (A1-A2, B1-B2, C1-C2). They were asked to reflect on the minimum proficiency level required to meet the linguistic and cognitive demands of the tasks successfully. However, some panellists indicated, through their judgements, whether tasks would allow candidates with higher proficiency levels to demonstrate their true writing and/or speaking abilities; as such, they sometimes recorded the higher end of the CEFR scale. This information was provided in the comments section, but unfortunately, these panellists did not specify the minimum CEFR proficiency level, so data could not be corrected.

Figure 5.3 shows the panellists' CEFR item ratings of the tasks and how they aligned to the CEFR. Table 5.3 explains the acronyms used.

Table 5.3: Acronyms used for the CEFR speaking scales

CEFR scale	Acronym
Overall Spoken Production	OOP
Overall Oral Interaction	OOI
Conversation	Conv.
Sustained Monologue	SM
Obtaining Goods and Services	OG & S
Processing Text (in Speech)	PTinS

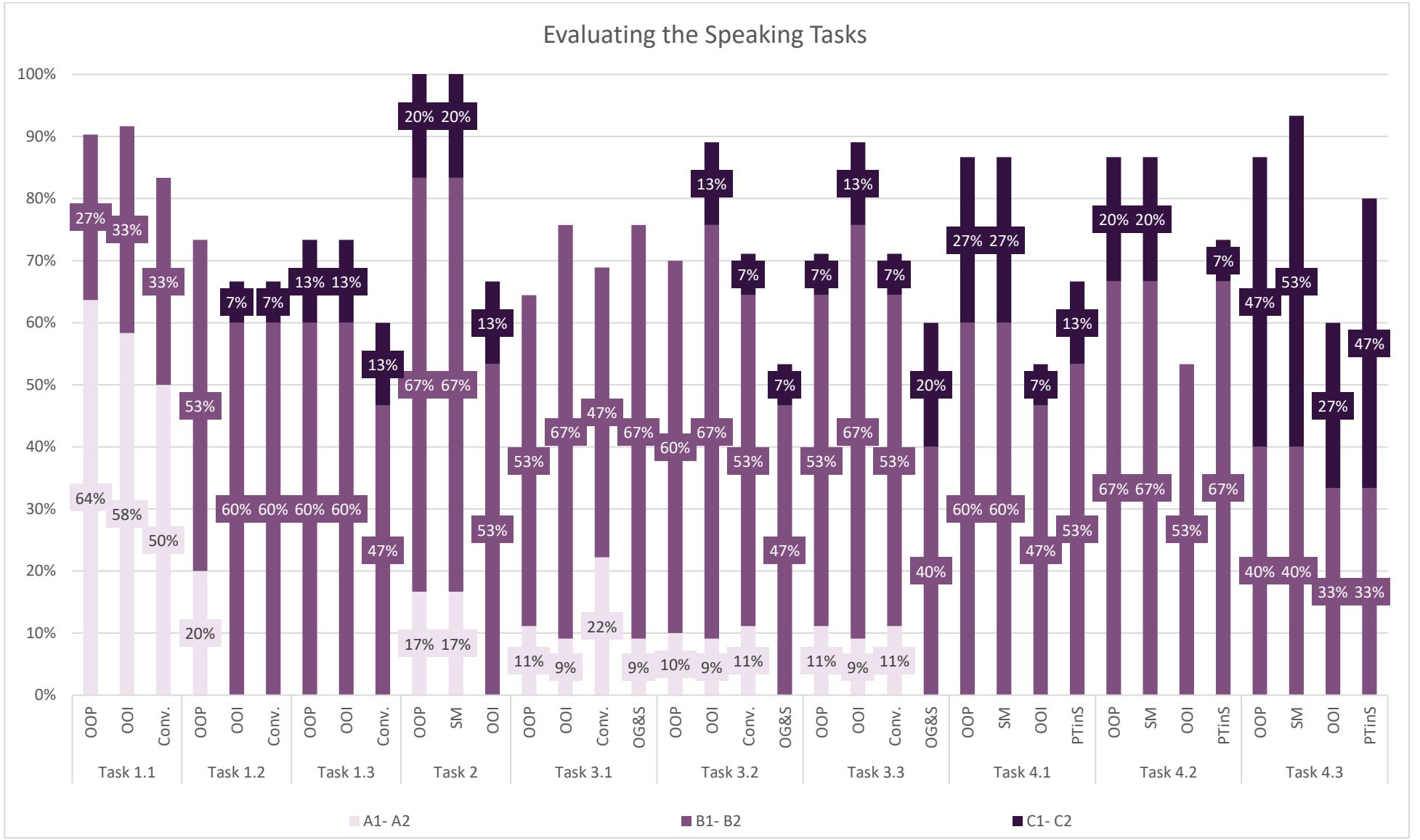


Figure 5.3: CEFR mapping of the speaking module tasks

The bar graph highlights the alignment of the speaking tasks in the ISE Digital examination with the intended CEFR levels, demonstrating a clear progression in cognitive and linguistic demands across tasks. All Task 1 items are designed to be more accessible, with Task 1.1 targeting mainly A1-A2 while ensuring accessibility to adjacent B1-B2 candidates, as shown by their majority alignment with A1-A2 CEFR descriptors (ranging from 50 to 64%) and notable overlap with B1-B2 (27% and 33%). Task 1.2, targeting B1-B2, maintains strong alignment with the target level (60% for most of the CEFR scales) while offering accessibility to both A1-A2 (20%) and C1-C2 (7%). Task 1.3, targeting mainly C1-C2 proficiency levels, maintains strong alignment with the adjacent B1-B2 route (ranging from 47% to 60%).

The tasks progressively increase in complexity. All tasks with a .2 extension primarily target B1-B2 (67%), while being accessible to C1-C2 (20%) and, to a lesser extent, A1-A2 (13%). Since the candidate leads in *delivering a prepared talk* task (represented in Figure 5.3 as Task 2), their performance significantly influenced the panellists' judgments. All the *interacting* items (described in Figure 5.3 as Task 3) are designed to be more challenging than the responding-to-questions items (represented in Figure 5.3 as Task 1). Task 3.1 was intended to be easier than Task 3.2, which, in turn, was designed to be less challenging than Task 3.3. This design was reflected in the panellists' judgments, as although Task 3.1 was mostly aligned with B1-B2 descriptors (47% to 67%), panellists also found that A1-A2 descriptors (9% to 22%) adequately reflected the task's demands. This indicated that Task 3.1 was, therefore, accessible to candidates with lower proficiency.

Though Task 3.2 was primarily aligned with the B1-B2 CEFR scales and descriptors, there was some overlap between the A1-A2 and C1-C2 levels and descriptors, suggesting that Task 3.2 could also be accessible to learners with lower and higher CEFR proficiency levels. A similar pattern was observed for Task 3.3, although a higher percentage of KSAs were mapped to the higher C1-C2 levels. Nevertheless, the panellists believed that B1-B2 candidates could handle the demands of the tasks exceptionally well, while A1-A2 candidates could manage them to a lesser extent. Similarly, Tasks 4.1 and 4.2 exhibited strong alignment with B1-B2 (ranging from 40% to 67%) while ensuring some accessibility to C1-C2 (ranging from 7% to 27%).

This alignment confirms that the *responding to questions* task is more accessible to lower proficiency candidates and that all tasks align with adaptive testing principles, targeting their primary CEFR levels while ensuring accessibility to adjacent levels. This design supports a fair progression within the adaptive testing framework and allows candidates to overcome potential misrouting issues.

This analysis of the ISE Digital speaking tasks demonstrates a well-structured alignment with the CEFR framework, ensuring that tasks meet their intended target levels while remaining accessible to adjacent levels. The *responding to questions* task, which has been designed to be more accessible, effectively bridges A1-A2 and B1-B2, providing an inclusive starting point for candidates. The progressive increase in complexity across subsequent tasks reflects careful calibration to the cognitive and linguistic demands of B1-B2 and C1-C2 levels, ensuring alignment with the adaptive nature of the test. By spanning adjacent CEFR levels and incorporating a balanced range of demands, the tasks uphold the principles of inclusivity, adaptability, and progression. This structured approach ensures that ISE Digital provides a robust and equitable assessment of speaking skills across a broad range of proficiency levels.

5.4 Inter- and Intra-Panellist Consistency

This section presents the analyses and results of two sources of internal validity evidence: 1) inter-panellist consistency and 2) intra-panellist consistency (Cizek & Earnest, 2016; Cizek, Hambleton, Pitoniak, & Copella, 2012; Hambleton & Pitoniak, 2006; Kane, 1994). Following Harsch and Kanistra (2020), panellists were asked to assess 16 candidate oral performances for the speaking tasks analysed in section 5.2, using the qualitative features of spoken language included in the CEFR CV (Council of Europe 2020, p. 183). This method generated 3,600 CEFR-level judgements per round for the oral performances of the four speaking tasks (16 oral performances × 5 criteria × 15 panellists × 3 tasks).

Inter- and intra-panellist consistency and reliability were evaluated within the RMT paradigm, allowing for a nuanced evaluation at both the individual and group levels (see Section 3.2 for a reminder of the key indices referred to in this section). A six-facet model was used to analyse the panellists' judgements on the speaking module: 1) candidate oral performances, 2) panellists, 3) task type, 4) panellist sub-groups (internals or externals), 5) round, and 6) criteria. Facets three to five were dummy facets used to facilitate various pairwise interactions, and as such, they did not affect the behaviour and measurement of the active facets. Tables 5.4 to 5.7 present the Rasch indices for panellist severity, inter- and intra-panellist consistency, and agreement for each round for each task. The first column indicates the Rasch index related to each measurement context, and columns two and three report the values for each index per round (ie the Round 1 and Round 2 judgement rounds). When interpreting the data in these tables, it is essential to note that higher values correspond to higher CEFR levels (e.g., A1 = 1, A1+ = 1.5, A2 = 2, A2+ = 2.5, and so on).

Table 5.4: Summary of panellist severity within RMT- speaking module (N=15)

Index	Speaking module	
	Round 1	**Round 2**
Average measure (*SD*)	-2.93 (0.94)	-2.91 (0.81)
Model *SE*	0.13	0.13
Measure min. (Model *SE*)	-4.51 (0.14)	-4.19 (0.13)
Measure max. (Model *SE*)	-1.03 (0.12)	-1.35 (0.12)
Fair average (min)	6.39	6.60
Fair average (max)	8.90	8.65

Overall, the mean measure of the panellists in both rounds (mean measure = -2.93 in R1; mean measure = -2.91 in R2) indicated that the panellists assigned relatively high CEFR judgements to most of the candidate oral performances. The panellists demonstrated high precision (*model SE* = 0.13 in Round 1 and Round 2) when evaluating candidate performances between the two rounds. A closer examination of panellist behaviour revealed that the spread measure between the most severe and the most lenient panellist decreased from 3.48 logits in Round 1 to 2.45 in Round 2, indicating that the discussions following the Round 1 judgements informed the ratings in Round 2. The effect of this spread on the raw judgements of the oral performances was 2.51 raw score points for Round 1 and 2.05 points for Round 2. This difference meant that the ratings of the most lenient panellist were approximately two CEFR levels higher than those of the most severe panellist, suggesting that not all panellist ratings were directly comparable. The MFRM model addressed these minor variations in the panellists' ratings, correcting any idiosyncratic behaviour exhibited by the panellists. This ensured that the behaviour of the panellists did not influence the final difficulty measures of the candidates' oral performances.

Table 5.5: Summary of inter-panellist consistency within RMT-speaking module (N=15)

Index	Speaking module	
	Round 1	**Round 2**
Overall *SP/ROP*	0.95	0.96
SP/ROP observed-(expected) minimum	0.92 (0.93)	0.93 (0.94)
SP/ROP observed-(expected) maximum	0.97 (0.96)	0.97 (0.97)
Overall Rasch *kappa*	0.02	0.04
Rasch *kappa* minimum	-0.01	-0.08
Rasch *kappa* maximum	0.07	0.10

Panellists demonstrated high inter-panellist consistency, as evidenced by the strong *SP/ROP* correlations (*SP/ROP* = 0.95 in Round 1; *SP/ROP* = 0.96 in Round 2). These values confirm that panellists consistently interpreted and applied the Qualitative Features of Spoken Language. Furthermore, the observed SP/ROP values closely matched the expected SP/ROP values, indicating that inter-panellist consistency aligned with the expectations of the Rasch model.

The Rasch kappa statistic offers an additional measure of agreement within the Rasch framework. In Round 1, Rasch kappa values varied from -0.01 to 0.07, while in Round 2, they ranged from -0.08 to 0.10. All these values fell within the acceptable range of -0.2 to +0.2, indicating that the panellists appraised candidate oral performances in accordance with the Rasch model's expectations while maintaining their independence as raters.

Table 5.6: Summary of inter-panellist agreement within RMT – speaking module (N=15)

Index	Speaking module	
	Round 1	Round 2
Overall exact observed % agreement (expected %)	43.2 (41.9%)	45.8% (43.3%)
exact observed % agreement (expected %) minimum	36.3% (31.7%)	37.8% (37.7%)
exact observed % agreement (expected %) maximum	46.8% (44.5%)	54.5% (47.7%)

Exact agreement among panellists was measured using the *exact observed % agreement* index. As expected, overall exact observed agreement increased following the discussion at the end of Round 1, rising from 43.2% (vs. 41.9% expected) in Round 1 to 45.8% (vs. 43.3% expected) in Round 2. These observed values were aligned closely with the expected values, reflecting the model's predictions. Furthermore, no panellist showed either observed or expected agreement above 80%, indicating that they acted as autonomous experts and exhibited a suitable level of agreement. This finding supports the credibility of their judgements.

Table 5.7: Summary of intra-panellist consistency within RMT (N=15)

Index	Speaking module	
	Round 1	Round 2
Mean Infit *Mnsq;* SD (*Zstd*)(Group)	1.02; 0.18 (0.10)	0.96; 0.18 (-0.20)
Minimum Infit *Mnsq* (*Zstd*)	0.68 (-2.70)	0.72 (-1.08)
Maximum Infit *Mnsq* (*Zstd*)	1.29 (1.90)	1.31 (1.80)

The detailed panellist measurement report is available in Appendix A and B. Table 5.7 shows that the mean Infit *Mnsq* values for the panellists remained near the ideal value of 1.00, varying between 1.02 and 0.96 across the two rounds. These outcomes demonstrate that the panellists maintained adequate intra-judge consistency throughout the Speaking standard-setting and benchmarking workshop, thereby supporting the internal validity of the resulting cut scores.

In line with Pollitt and Hutchinson (1987), the acceptable Infit range for Round 1 extended from 0.66 to 1.38, while for Round 2, it ranged from 0.60 to 1.32 (Infit mean ± 2SD). All panellists' Infit measures fell within these limits, which are deemed acceptable for trained panellists. Notably, the highest Infit values, which approached the upper boundary, were associated with *Zstd* values below ±2, suggesting these slight deviations had no significant impact on the reliability of the CEFR item judgments. These findings align with earlier evidence of internal consistency and further reinforce the credibility of the judges' evaluations.

In summary, the findings suggest that panellists' judgements were both consistent and reliable. The discussion at the end of Round 1 further aligned their assessments, ensuring that all judgements effectively contributed to recommending valid and reliable cut scores.

Consequently, the next set of analyses will focus on evaluating the consistency, accuracy, and precision of these recommended cut scores.

5.5 Consistency within the Method for the Speaking Module

As explained in Section 3.3, the consistency within the method for the speaking module was evaluated by following the processes and procedures outlined in the Manual (Council of Europe, 2009). The recommended cut scores for the speaking module were evaluated for their i) precision and accuracy, and ii) classification consistency and accuracy. As suggested by Kaftandjieva (2010), a dataset of 4,651 candidates was simulated based on the ability measures of the 394 candidates who had participated in test trialling using *Facets v4.4.4* (Linacre, 2025) to facilitate the in-depth analyses of the cut scores. A total of 81 different speaking tasks were used in the trialling exercises. Table 5.8 illustrates that the psychometric properties of the real and simulated data were remarkably similar.

Table 5.8: Psychometric characteristics of real & simulated candidate population – speaking

Index	Real (N = 349)	Simulated (N = 4,941)
Number of tasks	81	81
Candidate mean measure (*SEm; SD*)	0.09 (0.46; 2.997)	0.10 (0.52; 2.27)
Test reliability	0.97	0.95
RMSE (CSEM)	0.48	0.53
Observed average (*SD*)	2.85 (0.91)	2.91 (0.71)
SEM	1.86	1.91
Fair average (*SD*)	2.80 (0.94)	2.91 (0.70)

For the Speaking module, the panellists were not only asked to evaluate the cognitive demands of the speaking tasks but also to classify the candidates' spoken responses according to CEFR levels and identify those that best exemplified the targeted CEFR levels. Table 5.9 presents the results of the consistency within the method checks, based on the panellists' CEFR classification of the candidate spoken responses, focusing specifically on those responses they agreed best exemplified performance at levels A1 to C2.

Table 5.9: Evaluating the accuracy & precision of the speaking module cut scores (N = 4,941)

CEFR level	SE_j	SD_j	SE_j / SD_p	SE_j / *SEM*
A1	0.11	0.40	0.006	0.20
A2	0.11	0.42	0.007	0.21
B1	0.13	0.49	0.008	0.25
B2	0.12	0.43	0.007	0.22
C1	0.13	0.50	0.008	0.26
C2	0.12	0.46	0.007	0.23

The standard deviation of the panellist judgements (SD_j) and the standard error of the mean of their judgements (SE_j) were very small. As a result, the SE_j relative to the standard deviation of the population ($SE_j/SD_p \leq 0.33$; $SD_p = 16.4$) indicates that the classification error had minimal influence on CEFR level assignment. Importantly, this also implies that the classifications of the spoken performances used to inform the cut scores are robust. This is further supported by the fact that the SE_j of the classifications of the spoken performances was consistently lower than one-third of the conditional standard error of measurement (*CSEM*) for each cut score ($SE_j/CSEM \leq 0.33$), which satisfies the criterion proposed by Kaftandjieva

(2010). Taken together, these findings provide strong validity evidence of consistency within the method, supporting the use of the panellists' selected spoken performances as reliable representations of the CEFR levels for standard setting purposes. These findings provide validity evidence for the consistency within the method aspect of evaluating standard setting studies, and as such, the recommended cut scores can be further evaluated.

5.6 Decision Consistency and Accuracy

In this section, the decision consistency and accuracy of the recommended cut score are evaluated using two methods: the Livingston and Lewis (denoted as *LL*) (1995) CTT-based method and the IRT-based method by Lee (2008) using *BB-CLASS* v1.1 (Brennan, 2004) and *IRT-CLASS v2* (Lee & Kolen, 2008), respectively. The recommended cut scores were derived from the candidates' spoken responses that the panellists identified as being the best representation of the targeted CEFR levels. For the *LL* and Lee methods, the raw scores assigned to candidate responses were used. For the IRT-based method, the individual approach (P) was applied using candidate ability estimates (Lee, 2010).

The Lee method requires item parameters to be included in the program as well; thus, in the context of this study, the nine rating criteria were treated as items, and Samejima's normal ogive graded response model was used to calculate the *DA* (γ) and consistency *DC* (φ) indices for the recommended cut scores at each CEFR level. The unidimensionality assumption, an important aspect of this analysis, was met. Test takers' ability measures for the speaking module were obtained through an *MFRM* analysis, allowing measurement errors due to rater behaviour to be accounted for.

Table 5.10 presents the results of the evaluation of the recommended cut scores under the Livingston and Lewis, and Lee methods. The evaluation methods are listed in the first column, while the recommended cut scores are provided in the second column, expressed as raw weighted scores. The table reports decision accuracy [$DA(\gamma)$] and consistency [$DC(\varphi)$] in columns three and four, respectively, alongside the kappa coefficient in column five. The proportion of correct classifications by chance [*pchance* (φ_C)] is presented in column six, followed by the probability of misclassifications in column seven. The false positive and false negative rates are also provided in columns eight and nine.

Table 5.10: Evaluating the accuracy & precision of the speaking cut scores (N = 4,941)

Method	Speaking scaled score	*DA (γ)*	*DC (φ)*	*Kappa (κ)*	*pchance ($φ_c$)*	Probability of misclassification	False positive rate	False negative rate
CEFR Level A1								
LL	5	0.997	0.996	0.58	0.99	0.004	0.0005	0.003
Lee	5	0.96	0.94	0.83	0.66	0.06	0.03	0.02
CEFR Level A2								
LL	30	0.97	0.96	0.74	0.84	0.04	0.01	0.02
Lee	30	0.95	0.93	0.84	0.51	0.07	0.03	0.02
CEFR Level B1								
LL	55	0.94	0.92	0.80	0.57	0.09	0.03	0.03
Lee	55	0.95	0.93	0.82	0.54	0.07	0.02	0.03
CEFR Level B2								
LL	80	0.94	0.92	0.79	0.64	0.08	0.03	0.02
Lee	80	0.97	0.96	0.82	0.78	0.04	0.02	0.01
CEFR Level C1								
LL	105	0.97	0.99	0.81	0.75	0.05	0.02	0.01
Lee	105	0.99	0.99	0.79	0.93	0.01	0.001	0.002
CEFR Level C2								
LL	130	0.99	0.99	0.70	0.97	0.01	0.01	0.002
Lee	130	0.99	0.99	0.76	0.98	0.01	0.003	0.0004

All DA ($γ$) and DC ($φ$) measures exceeded the recommended minimum criterion of 0.85 (Subkoviak 1988) for certification examinations at each CEFR level across both CTT and IRT-based methods. This indicates that the classification of candidates into different CEFR levels is consistent and accurate. Similar to Lee (2010), Deng & Hambleton (2013), and Kanistra (forthcoming), the IRT-based method yielded higher *DA* ($γ$) and *DC* indices (including $φ$, $φc$, and $κ$ coefficients), particularly for the recommended cut scores that were further from and lower than the population mean. The $κ$ values were higher than or very close to the expected 0.60 in both the CTT and IRT frameworks. For most CEFR cut scores, except those positioned at the lower or maximum possible scores, the κ values were greater than or nearly equal to the pchance value ($φc$). Consistent with Subkoviak (1988), pchance ($φc$) increases when cut scores are set towards the lower or upper ends of the scale, which is expected because the least and most able candidates tend to perform similarly even in non-parallel tests. However, it should be noted that for all CEFR levels, $κ$ is exceptionally high, indicating that candidate classification largely depends on their performance in the speaking module of the ISE Digital exam.

In summary, the ISE Digital speaking module items were mapped to the CEFR in three phases: during the conceptualisation stage, during the item creation phase, and through standard setting using the ID Matching method. The responses of candidates were aligned with the CEFR via the Benchmarking approach as outlined in the Manual (Council of Europe, 2009). Consequently, the ISE Digital speaking Module aligns with the CEFR both qualitatively in terms of content and quantitatively via the Benchmarking approach, as reflected in the scores awarded to candidates' spoken responses.

6 Listening Cut Scores and Validity Evidence

The CEFR cut scores for the ISE Digital listening module were established using the *Principled Cut Score* approach (Kanistra, forthcoming). This approach (see Section 2.3 for more details) entails the following sequential steps:

- Establish the predictive power of each item
- Convert ability measures or raw scores to z-scores
- Establish item clusters
- Explore the predictive power of the threshold regions
- Locate the cut scores within the threshold regions

The estimated cut scores are then evaluated using post hoc checks. This section presents the results for each step and reviews the internal validity evidence for this method, focusing on its internal consistency and the classification consistency of the resulting cut scores.

6.1 Psychometric Properties of the ISE Digital Listening Module

The listening module was analysed using Rasch measurement, with 2,359 candidate responses calibrated in *Winsteps v5.8.3.0* (Linacre, 2024). Sixteen DIALANG listening items served as anchors, linking the scale to the CEFR through an established CEFR-aligned test. In total, 258 items (242 operational and 16 anchor items) were included in the calibration. A summary of the psychometric properties of the listening items included in the study is shown in Table 6.1.

Table 6.1: Rasch summary statistics for the ISE Digital listening module

Index	**Real (N = 2,359)**
Number of items	258
Item difficulty mean measure (*SEm ; SD*)	-0.56 (0.29; 1.29)
Candidate mean measure (*SEm ; SD*)	-1.10 (0.48; 1.18)
Test reliability	0.83
RMSE	0.49
Mean score (*SD*)	12.7 (6.90)
Score min - max	2 - 38

The item difficulty distribution (mean = −0.56 logits; SD = 1.29) indicates that the bank covers an appropriate span of difficulty for the ISE Digital population, ranging from accessible items for lower-level learners to more demanding items targeting higher proficiency. Candidate measures averaged -1.10 logits (SD = 1.18). The relatively large standard deviation (SD) reflects the range of candidate abilities in the pilot cohort.

Reliability was strong (0.83), indicating stable separation of candidate listening abilities, and the RMSE of 0.49 logits reflects adequate measurement precision for this type of receptive skills assessment. The observed score distribution (mean = 12.7; SD = 6.90; range = 2–38) shows that the module elicited a wide spread of responses.

Overall, the distribution of item difficulties, candidate measures, reliability and measurement error demonstrates that the listening module provides a stable and precise measure of receptive ability, supporting its use in the CEFR standard-setting procedure described in the following sections.

6.2 Establishing the Predictive Power of Each Item

This is the first step in the *Principled Cut Score* approach, which involves linear regression analysis to identify the items that significantly contribute to a candidate's ability. The linear regression analysis was performed on 258 calibrated items. Preliminary checks were conducted to ensure the dataset's suitability for multiple regression analysis, specifically verifying that there were no violations of the assumptions of normality, linearity, multicollinearity, and homoscedasticity. These assumptions are essential for the reliability and validity of the results (Pallant, 2016; Tabachnick & Fidell, 2014). According to the guidelines in Tabachnick and Fidell (2014), candidates with standardised residuals exceeding |3.3| (absolute value) were identified as outliers and were removed from the analysis. This resulted in a sample size of 2,351 candidates. The candidate ability measures (βv) were entered as the dependent variable in the multiple regression analysis.

The full dataset of 258 items explained 98.8% of the variance in ability measures in a statistically significant way (adjusted R^2 = .977, F = 343.220, *d.f.1 = 258, d.f.2 = 2092, p <.001, p* < .01). Of these, a subset of 133 items explained candidate ability variance in a statistically significant way (*Sig.* < .01*).* These 133 items were used in the second step of the *Principled Cut Score* approach (Kanistra, forthcoming).

6.3 Converting Item Difficulty Measures to z Scores

As a reminder, after the linear regression, the dataset comprised the ability measures of 2,351 candidates and the difficulty values of 133 items. In this step, the item difficulty measures are converted to z-scores. A z-score is a statistic that indicates the distance of an item's difficulty measure from the mean of the candidate population. A z-score of 0 indicates that the item's difficulty measure is at the population mean. Items with difficulty measures below the mean have a negative z-score, and those whose difficulty measures are above the mean have a positive z-score. Z-scores offer a quick insight into where a score (especially a proposed cut score) lies in relation to the overall candidate ability distribution. They also enable the identification of extreme values. Therefore, the objective of step 2 was to determine potential cut score locations relative to the population mean of the candidates' ability (βv = -1.10). This step is crucial because it prevents placing the cut scores at the extreme ends of the item difficulty scale, where the classification of test takers would largely depend on chance (Subkoviak 1980, 1988).

Since ISE Digital is a multilevel test targeting different CEFR levels, the z-score conversion must also be interpreted within this broader context. As explained in section 6.1, the anchoring was based on the DIALANG standard setting (Alderson 2005), with item difficulties derived from this procedure. This ensured that the means of each DIALANG-referenced CEFR level served as additional benchmarks. As a result, the z-score distances were examined not only relative to the overall population mean, but also against the expected mean performance at each CEFR level (derived from the DIALANG anchor items). In this way, the inspection of z scores supported the placement of cut scores at relevant points on the logit scale across the CEFR continuum, maximising classification consistency (DC, φ) and κ coefficients in line with Subkoviak's (1988) recommendations. Thus, as an additional check, the distances between potential cut scores and the DIALANG CEFR level means were also examined.

Table 6.2 presents the distance of the possible cut scores from the population mean ability measures. Columns 1 and 2 display the item IDs and their associated item difficulties (δ). Columns 3 to 9 present the associated z-scores and the context in which they are examined (mean candidate ability or CEFR level).

Table 6.2: Cut score position relative to the population and DIALANG Listening means

Item ID	Item difficulty (*δ*)	z-score candidate ability mean	z-score DIALANG A1	z-score DIALANG A2	z-score DIALANG B1	z-score DIALANG B2	z-score DIALANG C1	z-score DIALANG C2
Item 2	-4.43	-2.83	0.20	-0.70	-1.61	-2.68	-2.80	-2.99
Item 3	-3.574	-2.10	0.51	-0.39	-1.31	-2.37	-2.49	-2.69
Item 4	-3.522	-2.06	0.53	-0.37	-1.29	-2.35	-2.48	-2.67
Item 8	-3.13	-1.72	0.67	-0.23	-1.15	-2.21	-2.33	-2.53
Item 9	-2.97	-1.59	0.73	-0.18	-1.09	-2.15	-2.28	-2.47
Item 12	-2.94	-1.56	0.74	-0.16	-1.08	-2.14	-2.27	-2.46
Item 13	-2.929	-1.55	0.74	-0.16	-1.07	-2.14	-2.26	-2.45
Item 16	-2.913	-1.54	0.75	-0.15	-1.07	-2.13	-2.26	-2.45
Item 18	-2.913	-1.54	0.75	-0.15	-1.07	-2.13	-2.26	-2.45
Item 19	-2.904	-1.53	0.75	-0.15	-1.06	-2.13	-2.25	-2.44
Item 20	-2.8	-1.44	0.79	-0.11	-1.03	-2.09	-2.22	-2.41
Item 22	-2.782	-1.43	0.80	-0.11	-1.02	-2.08	-2.21	-2.40
Item 23	-2.779	-1.42	0.80	-0.11	-1.02	-2.08	-2.21	-2.40
Item 25	-2.583	-1.26	0.87	-0.04	-0.95	-2.01	-2.14	-2.33
Item 27	-2.538	-1.22	0.89	-0.02	-0.93	-1.99	-2.12	-2.31
Item 29	-2.49	-1.18	0.90	0.00	-0.92	-1.98	-2.10	-2.30
Item 30	-2.476	-1.17	0.91	0.00	-0.91	-1.97	-2.10	-2.29
Item 31	-2.314	-1.03	0.97	0.06	-0.85	-1.91	-2.04	-2.23
Item 32	-2.307	-1.02	0.97	0.06	-0.85	-1.91	-2.04	-2.23
Item 36	-2.196	-0.93	1.01	0.10	-0.81	-1.87	-2.00	-2.19
Item 37	-2.148	-0.89	1.03	0.12	-0.79	-1.85	-1.98	-2.17
Item 39	-2.101	-0.85	1.04	0.14	-0.78	-1.84	-1.96	-2.16
Item 40	-2.087	-0.84	1.05	0.14	-0.77	-1.83	-1.96	-2.15
Item 41	-2.046	-0.80	1.06	0.16	-0.76	-1.82	-1.94	-2.14
Item 45	-2.002	-0.77	1.08	0.17	-0.74	-1.80	-1.93	-2.12
Item 55	-1.958	-0.73	1.09	0.19	-0.72	-1.79	-1.91	-2.10
Item 57	-1.874	-0.66	1.12	0.22	-0.69	-1.75	-1.88	-2.07
Item 58	-1.874	-0.66	1.12	0.22	-0.69	-1.75	-1.88	-2.07
Item 59	-1.833	-0.62	1.14	0.23	-0.68	-1.74	-1.87	-2.06
Item 61	-1.83	-0.62	1.14	0.24	-0.68	-1.74	-1.87	-2.06

Item ID	Item difficulty (*δ*)	z-score candidate ability mean	z-score DIALANG A1	z-score DIALANG A2	z-score DIALANG B1	z-score DIALANG B2	z-score DIALANG C1	z-score DIALANG C2
Item 62	-1.816	-0.61	1.15	0.24	-0.67	-1.73	-1.86	-2.05
Item 63	-1.792	-0.59	1.15	0.25	-0.66	-1.73	-1.85	-2.04
Item 65	-1.792	-0.59	1.15	0.25	-0.66	-1.73	-1.85	-2.04
Item 66	-1.792	-0.59	1.15	0.25	-0.66	-1.73	-1.85	-2.04
Item 67	-1.792	-0.59	1.15	0.25	-0.66	-1.73	-1.85	-2.04
Item 70	-1.788	-0.58	1.16	0.25	-0.66	-1.72	-1.85	-2.04
Item 71	-1.686	-0.50	1.19	0.29	-0.63	-1.69	-1.81	-2.01
Item 73	-1.562	-0.39	1.24	0.33	-0.58	-1.64	-1.77	-1.96
Item 74	-1.548	-0.38	1.24	0.34	-0.58	-1.64	-1.76	-1.96
Item 75	-1.46	-0.31	1.27	0.37	-0.54	-1.61	-1.73	-1.92
Item 78	-1.445	-0.29	1.28	0.37	-0.54	-1.60	-1.73	-1.92
Item 80	-1.444	-0.29	1.28	0.37	-0.54	-1.60	-1.73	-1.92
Item 82	-1.427	-0.28	1.29	0.38	-0.53	-1.59	-1.72	-1.91
Item 83	-1.416	-0.27	1.29	0.38	-0.53	-1.59	-1.72	-1.91
Item 84	-1.377	-0.24	1.30	0.40	-0.51	-1.58	-1.70	-1.89
Item 85	-1.356	-0.22	1.31	0.41	-0.51	-1.57	-1.70	-1.89
Item 86	-1.354	-0.22	1.31	0.41	-0.51	-1.57	-1.69	-1.89
Item 88	-1.354	-0.22	1.31	0.41	-0.51	-1.57	-1.69	-1.89
Item 89	-1.349	-0.21	1.31	0.41	-0.50	-1.57	-1.69	-1.88
Item 90	-1.328	-0.19	1.32	0.42	-0.50	-1.56	-1.69	-1.88
Item 92	-1.292	-0.16	1.33	0.43	-0.48	-1.55	-1.67	-1.86
Item 93	-1.268	-0.14	1.34	0.44	-0.48	-1.54	-1.66	-1.85
Item 96	-1.194	-0.08	1.37	0.46	-0.45	-1.51	-1.64	-1.83
Item 98	-1.18	-0.07	1.37	0.47	-0.44	-1.50	-1.63	-1.82
Item 99	-1.177	-0.07	1.38	0.47	-0.44	-1.50	-1.63	-1.82
Item 101	-1.173	-0.06	1.38	0.47	-0.44	-1.50	-1.63	-1.82
Item 104	-1.173	-0.06	1.38	0.47	-0.44	-1.50	-1.63	-1.82
Item 105	-1.132	-0.03	1.39	0.49	-0.43	-1.49	-1.61	-1.81
Item 108	-1.043	0.05	1.42	0.52	-0.39	-1.46	-1.58	-1.77
Item 109	-1.043	0.05	1.42	0.52	-0.39	-1.46	-1.58	-1.77
Item 111	-1.036	0.05	1.43	0.52	-0.39	-1.45	-1.58	-1.77
Item 113	-0.996	0.09	1.44	0.54	-0.38	-1.44	-1.57	-1.76
Item 117	-0.996	0.09	1.44	0.54	-0.38	-1.44	-1.57	-1.76
Item 123	-0.95	0.13	1.46	0.55	-0.36	-1.42	-1.55	-1.74
Item 124	-0.86	0.20	1.49	0.58	-0.33	-1.39	-1.52	-1.71
Item 128	-0.86	0.20	1.49	0.58	-0.33	-1.39	-1.52	-1.71

Item ID	Item difficulty (*δ*)	z-score candidate ability mean	z-score DIALANG A1	z-score DIALANG A2	z-score DIALANG B1	z-score DIALANG B2	z-score DIALANG C1	z-score DIALANG C2
Item 129	-0.818	0.24	1.51	0.60	-0.31	-1.37	-1.50	-1.69
Item 130	-0.777	0.27	1.52	0.61	-0.30	-1.36	-1.49	-1.68
Item 131	-0.77	0.28	1.52	0.62	-0.30	-1.36	-1.48	-1.68
Item 133	-0.756	0.29	1.53	0.62	-0.29	-1.35	-1.48	-1.67
Item 134	-0.756	0.29	1.53	0.62	-0.29	-1.35	-1.48	-1.67
Item 135	-0.717	0.32	1.54	0.64	-0.28	-1.34	-1.47	-1.66
Item 136	-0.696	0.34	1.55	0.64	-0.27	-1.33	-1.46	-1.65
Item 137	-0.687	0.35	1.55	0.65	-0.27	-1.33	-1.45	-1.65
Item 138	-0.646	0.38	1.57	0.66	-0.25	-1.31	-1.44	-1.63
Item 139	-0.613	0.41	1.58	0.67	-0.24	-1.30	-1.43	-1.62
Item 141	-0.594	0.43	1.59	0.68	-0.23	-1.29	-1.42	-1.61
Item 143	-0.531	0.48	1.61	0.70	-0.21	-1.27	-1.40	-1.59
Item 144	-0.471	0.53	1.63	0.73	-0.19	-1.25	-1.38	-1.57
Item 145	-0.411	0.58	1.65	0.75	-0.17	-1.23	-1.36	-1.55
Item 156	-0.395	0.60	1.66	0.75	-0.16	-1.22	-1.35	-1.54
Item 157	-0.379	0.61	1.66	0.76	-0.16	-1.22	-1.34	-1.53
Item 160	-0.369	0.62	1.67	0.76	-0.15	-1.21	-1.34	-1.53
Item 162	-0.358	0.63	1.67	0.77	-0.15	-1.21	-1.34	-1.53
Item 163	-0.31	0.67	1.69	0.78	-0.13	-1.19	-1.32	-1.51
Item 165	-0.267	0.71	1.70	0.80	-0.11	-1.18	-1.30	-1.49
Item 169	-0.261	0.71	1.71	0.80	-0.11	-1.17	-1.30	-1.49
Item 171	-0.261	0.71	1.71	0.80	-0.11	-1.17	-1.30	-1.49
Item 173	-0.096	0.85	1.77	0.86	-0.05	-1.11	-1.24	-1.43
Item 175	-0.019	0.92	1.79	0.89	-0.03	-1.09	-1.21	-1.40
Item 176	-0.012	0.92	1.80	0.89	-0.02	-1.08	-1.21	-1.40
Item 178	0.079	1.00	1.83	0.92	0.01	-1.05	-1.18	-1.37
Item 180	0.095	1.01	1.83	0.93	0.02	-1.05	-1.17	-1.36
Item 182	0.113	1.03	1.84	0.94	0.02	-1.04	-1.17	-1.36
Item 186	0.113	1.03	1.84	0.94	0.02	-1.04	-1.17	-1.36
Item 187	0.192	1.09	1.87	0.96	0.05	-1.01	-1.14	-1.33
Item 189	0.2	1.10	1.87	0.97	0.05	-1.01	-1.14	-1.33
Item 193	0.222	1.12	1.88	0.97	0.06	-1.00	-1.13	-1.32
Item 195	0.271	1.16	1.90	0.99	0.08	-0.98	-1.11	-1.30
Item 196	0.271	1.16	1.90	0.99	0.08	-0.98	-1.11	-1.30
Item 198	0.286	1.17	1.90	1.00	0.08	-0.98	-1.10	-1.29
Item 199	0.298	1.18	1.91	1.00	0.09	-0.97	-1.10	-1.29

Item ID	Item difficulty (*δ*)	z-score candidate ability mean	z-score DIALANG A1	z-score DIALANG A2	z-score DIALANG B1	z-score DIALANG B2	z-score DIALANG C1	z-score DIALANG C2
Item 200	0.335	1.22	1.92	1.02	0.10	-0.96	-1.09	-1.28
Item 201	0.349	1.23	1.93	1.02	0.11	-0.95	-1.08	-1.27
Item 202	0.364	1.24	1.93	1.03	0.11	-0.95	-1.08	-1.27
Item 204	0.418	1.29	1.95	1.05	0.13	-0.93	-1.06	-1.25
Item 205	0.479	1.34	1.97	1.07	0.15	-0.91	-1.03	-1.23
Item 206	0.488	1.35	1.98	1.07	0.16	-0.90	-1.03	-1.22
Item 207	0.612	1.45	2.02	1.12	0.20	-0.86	-0.99	-1.18
Item 208	0.612	1.45	2.02	1.12	0.20	-0.86	-0.99	-1.18
Item 209	0.744	1.56	2.07	1.16	0.25	-0.81	-0.94	-1.13
Item 210	0.744	1.56	2.07	1.16	0.25	-0.81	-0.94	-1.13
Item 211	0.808	1.62	2.09	1.19	0.27	-0.79	-0.92	-1.11
Item 213	0.841	1.65	2.10	1.20	0.28	-0.78	-0.90	-1.10
Item 214	0.876	1.68	2.12	1.21	0.30	-0.76	-0.89	-1.08
Item 215	0.884	1.68	2.12	1.21	0.30	-0.76	-0.89	-1.08
Item 217	0.898	1.69	2.12	1.22	0.30	-0.76	-0.88	-1.07
Item 222	0.925	1.72	2.13	1.23	0.31	-0.75	-0.87	-1.06
Item 223	1.033	1.81	2.17	1.27	0.35	-0.71	-0.83	-1.03
Item 231	1.07	1.84	2.19	1.28	0.37	-0.69	-0.82	-1.01
Item 233	1.149	1.91	2.21	1.31	0.40	-0.67	-0.79	-0.98
Item 237	1.209	1.96	2.24	1.33	0.42	-0.64	-0.77	-0.96
Item 238	1.209	1.96	2.24	1.33	0.42	-0.64	-0.77	-0.96
Item 239	1.209	1.96	2.24	1.33	0.42	-0.64	-0.77	-0.96
Item 241	1.331	2.06	2.28	1.37	0.46	-0.60	-0.73	-0.92
Item 242	1.55	2.25	2.36	1.45	0.54	-0.52	-0.65	-0.84
Item 247	1.74	2.41	2.43	1.52	0.61	-0.45	-0.58	-0.77
Item 248	2.379	2.95	2.66	1.75	0.84	-0.22	-0.35	-0.54
Item 250	2.57	3.11	2.73	1.82	0.91	-0.15	-0.28	-0.47
Item 252	2.57	3.11	2.73	1.82	0.91	-0.15	-0.28	-0.47
Item 254	2.96	3.44	2.87	1.96	1.05	-0.01	-0.14	-0.33
Item 257	3.29	3.72	2.99	2.08	1.17	0.11	-0.02	-0.21
Item 258	3.456	3.86	3.05	2.14	1.23	0.17	0.04	-0.15

The population-based z scores were calculated using the overall mean ability measure (βv = -1.10) and its corresponding standard deviation (SD = 1.18) through Equation 5. For the CEFR-referenced z scores, the mean and standard deviation of the relevant CEFR anchor items (from the DIALANG standard setting) were used.

$$z = \frac{(c-M)}{SD} \qquad \textit{Equation 5}$$

Where *c* is the cut score location, *M* is the population mean, and *SD* is the standard deviation of the test-ability measures.

According to Subkoviak (1988), κ coefficients in the range of .58 to .70 and decision consistency (DC, φ) values of .86 or higher are recommended benchmarks for high-stakes certification assessments. The z-scores observed in this study (Table 6.2), ranging from |0.00| to |3.86| (highlighted cells in Table 6.2), fall within acceptable limits when interpreted through Subkoviak's reference tables (1988, pp. 49). Assuming a test reliability above 0.80, these z-scores correspond to κ values between .58 and .71 and φ values between .86 and .98. This suggests that, in principle, the indices targeted in this study can be obtained at levels appropriate for a high-stakes test. Nonetheless, it remains advisable to avoid placing cut scores at the extremes of the item-difficulty distribution to maintain interpretive stability.

6.4 Establishing Item Clusters

In the ID Matching method, the threshold region consists of items arranged in a consecutive ascending order of difficulty. To maintain this property in the engineered threshold region, a series of Wald t-tests were conducted to identify clusters of items with similar difficulty, comparing the first item to the second, then to the third, and so on. To reduce the risk of false rejections of the null hypothesis due to the high number of comparisons (Cohen, 1988, 1992), the significance level was set at 0.01. Table 6.3 presents the results of the Wald t-test analyses. The compared items are listed in columns 1 and 2, while column 3 (Cluster) indicates whether the items are grouped in the same cluster. Column 4 (Wald t) reports the Wald statistic for each comparison, followed by the degrees of freedom (*d.f.*) in column 5 for each interaction, and column 6 (Significance, two-tailed prob.) states whether the Wald statistic was statistically significant. This analysis returned 11 clusters to take forward to the next analytic step.

Table 6.3: Using Wald statistics to establish item clusters for the listening module

Item ID	Item Id compared	Cluster	Wald t	d.f.	Significance (two-tailed prob.)[2]
DIALANG A1+	Item 3	End of Cluster 1	-3.04	924	0.00
Item 03	Item 4	Cluster 2	-0.19	524	0.85
Item 03	DIALANG A2	Cluster 2	-1.79	926	0.07
Item 03	Item 9	Cluster 2	-2.29	524	0.02
Item 03	Item 12	Cluster 2	-1.26	290	0.21
Item 03	Item 13	Cluster 2	-2.46	524	0.01
Item 03	Item 16	Cluster 2	-0.76	264	0.45
Item 03	Item 18	Cluster 2	-0.76	264	0.45
Item 03	Item 19	Cluster 2	-2.51	486	0.01
Item 03	Item 20	Cluster 2	-1.34	280	0.18
Item 03	Item 22	End of Cluster 2	-2.70	488	0.01
Item 22	Item 23	Cluster 3	-0.01	299	0.99
Item 23	Item 25	Cluster 3	-0.78	488	0.43
Item 23	Item 27	Cluster 3	-0.78	340	0.43
Item 23	DIALANG A2	Cluster 3	-1.35	926	0.18
Item 23	Item 30	Cluster 3	-1.22	488	0.22
Item 23	Item 31	Cluster 3	-0.95	282	0.34

[2] Statistically significant at the $p \leq .01$

Item ID	Item Id compared	Cluster	Wald t	d.f.	Significance (two-tailed prob.)[2]
Item 23	Item 32	Cluster 3	-1.55	340	0.12
Item 23	Item 36	Cluster 3	-1.93	340	0.05
Item 23	Item 37	End of Cluster 3	-2.65	550	0.01
Item 37	Item 39	Cluster 4	-0.24	548	0.81
Item 37	Item 40	Cluster 4	-0.22	400	0.82
Item 37	Item 41	Cluster 4	-0.50	548	0.62
Item 37	Item 45	Cluster 4	-0.71	548	0.48
Item 37	Item 55	Cluster 4	-0.93	548	0.35
Item 37	Item 57	Cluster 4	-1.02	400	0.31
Item 37	Item 58	Cluster 4	-1.02	400	0.31
Item 37	Item 59	Cluster 4	-1.67	640	0.10
Item 37	DIALANG A2+	Cluster 4	-1.92	1018	0.06
Item 37	Item 62	Cluster 4	-1.09	390	0.28
Item 37	Item 63	Cluster 4	-0.43	312	0.67
Item 37	Item 65	Cluster 4	-0.43	312	0.67
Item 37	Item 66	Cluster 4	-0.43	312	0.67
Item 37	Item 67	Cluster 4	-0.43	312	0.67
Item 37	Item 70	Cluster 4	-0.51	326	0.61
Item 37	Item 71	End of Cluster 4	-3.11	2601	0.00
Item 71	Item 73	Cluster 5	-0.38	2343	0.71
Item 71	Item 74	Cluster 5	-0.99	2577	0.32
Item 71	Item 75	Cluster 5	-0.97	2391	0.33
Item 71	Item 78	Cluster 5	-1.72	2577	0.09
Item 71	Item 80	Cluster 5	-1.84	2631	0.07
Item 71	Item 82	Cluster 5	-1.75	2539	0.08
Item 71	Item 83	Cluster 5	-1.07	2381	0.29
Item 71	Item 84	Cluster 5	-2.08	2539	0.04
Item 71	Item 85	Cluster 5	-1.03	2343	0.30
Item 71	Item 86	Cluster 5	-0.91	2333	0.36
Item 71	Item 88	Cluster 5	-0.91	2333	0.36
Item 71	Item 89	Cluster 5	-2.49	2619	0.02
Item 71	Item 90	Cluster 5	-1.79	2435	0.07
Item 71	Item 92	End of Cluster 5	-4.02	3003	0.00
Item 92	Item 93	Cluster 6	-0.15	988	0.88
Item 92	Item 96	Cluster 6	-0.60	950	0.55
Item 92	Item 98	Cluster 6	-0.68	952	0.50
Item 92	Item 99	Cluster 6	-0.37	761	0.71

Item ID	Item Id compared	Cluster	Wald t	d.f.	Significance (two-tailed prob.)[2]
Item 92	Item 101	Cluster 6	-0.46	786	0.64
Item 92	Item 104	Cluster 6	-0.46	786	0.64
Item 92	Item 105	Cluster 6	-0.95	950	0.34
Item 92	Item 108	Cluster 6	-1.01	802	0.31
Item 92	Item 109	Cluster 6	-1.01	802	0.31
Item 92	Item 111	Cluster 6	-1.23	846	0.22
Item 92	Item 113	Cluster 6	-0.94	761	0.35
Item 92	Item 117	Cluster 6	-0.94	761	0.35
Item 92	Item 123	Cluster 6	-0.55	728	0.58
Item 92	Item 124	Cluster 6	-1.22	744	0.22
Item 92	Item 128	Cluster 6	-1.22	744	0.22
Item 92	Item 129	Cluster 6	-1.82	786	0.07
Item 92	Item 130	End of Cluster 6	-2.92	950	0.00
Item 130	Item 131	Cluster 7	-0.02	290	0.98
Item 130	Item 133	Cluster 7	-0.10	570	0.92
Item 130	Item 134	Cluster 7	-0.10	488	0.92
Item 130	Item 135	Cluster 7	-0.37	2533	0.71
Item 130	Item 136	Cluster 7	-0.28	322	0.78
Item 130	Item 137	Cluster 7	-0.12	250	0.90
Item 130	Item 138	Cluster 7	-0.81	2581	0.42
Item 130	Item 139	Cluster 7	-0.68	382	0.50
Item 130	Item 141	Cluster 7	-1.13	2569	0.26
Item 130	Item 143	Cluster 7	-1.17	486	0.24
Item 130	Item 144	Cluster 7	-1.90	2587	0.06
Item 130	Item 145	Cluster 7	-1.32	328	0.19
Item 130	Item 156	Cluster 7	-1.02	280	0.31
Item 130	Item 157	Cluster 7	-1.13	290	0.26
Item 130	Item 160	Cluster 7	-2.52	2531	0.02
Item 130	Item 162	Cluster 7	-2.05	576	0.04
Item 130	Item 163	Cluster 7	-1.54	322	0.12
Item 130	Item 165	Cluster 7	-1.52	294	0.13
Item 130	Item 169	Cluster 7	-0.88	264	0.38
Item 130	Item 171	Cluster 7	-0.88	264	0.38
Item 130	Item 173	End of Cluster 7	-3.24	488	0.00
Item 173	Item 175	Cluster 8	-0.39	556	0.70
Item 173	Item 176	Cluster 8	-0.23	299	0.82
Item 173	Item 178	Cluster 8	-0.46	282	0.65

Item ID	Item Id compared	Cluster	Wald t	d.f.	Significance (two-tailed prob.)[2]
Item 173	Item 180	Cluster 8	-0.57	296	0.57
Item 173	Item 182	Cluster 8	-0.86	384	0.39
Item 173	Item 186	Cluster 8	-0.86	384	0.39
Item 173	Item 187	Cluster 8	-1.18	384	0.24
Item 173	DIALANG B1	Cluster 8	-1.71	956	0.09
Item 173	Item 193	Cluster 8	-1.56	552	0.12
Item 173	Item 195	Cluster 8	-0.98	292	0.33
Item 173	Item 196	Cluster 8	-0.98	292	0.33
Item 173	Item 198	Cluster 8	-1.11	296	0.27
Item 173	Item 199	Cluster 8	-1.20	324	0.23
Item 173	Item 200	Cluster 8	-0.79	266	0.43
Item 173	Item 201	Cluster 8	-1.14	299	0.25
Item 173	Item 202	Cluster 8	-1.62	330	0.11
Item 173	Item 204	Cluster 8	-0.62	252	0.54
Item 173	Item 205	Cluster 8	-1.69	324	0.09
Item 173	Item 206	Cluster 8	-1.65	296	0.10
Item 173	Item 207	Cluster 8	-1.73	282	0.08
Item 173	Item 208	Cluster 8	-1.73	282	0.08
Item 173	Item 209	End of Cluster 8	-2.82	330	0.01
Item 209	Item 210	Cluster 9	0.00	170	1.00
Item 209	Item 211	Cluster 9	-0.14	132	0.89
Item 209	Item 213	Cluster 9	-0.31	380	0.75
Item 209	Item 214	Cluster 9	-0.23	106	0.82
Item 209	Item 215	Cluster 9	-0.37	170	0.71
Item 209	Item 217	Cluster 9	-0.45	224	0.66
Item 209	Item 222	Cluster 9	-0.37	122	0.71
Item 209	Item 223	Cluster 9	-0.76	170	0.45
Item 209	Item 231	Cluster 9	-1.03	330	0.30
Item 209	Item 233	Cluster 9	-0.77	139	0.44
Item 209	Item 237	Cluster 9	-1.01	136	0.31
Item 209	Item 238	Cluster 9	-1.01	136	0.31
Item 209	Item 239	Cluster 9	-1.01	136	0.31
Item 209	Item 241	Cluster 9	-1.83	380	0.07
Item 209	Item 242	Cluster 9	-1.47	132	0.14
Item 209	DIALANG B1+	End of Cluster 9	-3.37	776	0.00
DIALANG B1+	Item 248	Cluster 10	-1.41	776	0.16
DIALANG B1+	Item 250	Cluster 10	-0.44	698	0.66

Item ID	Item Id compared	Cluster	Wald t	d.f.	Significance (two-tailed prob.)[2]
DIALANG B1+	Item 252	Cluster 10	-0.44	698	0.66
DIALANG B1+	DIALANG B2+/C1	End of Cluster 10	-5.09	1402	0.00
DIALANG B2+/C1	DIALANG C1	Cluster 11	-1.13	1420	0.26
DIALANG B2+/C1	Item 258	Cluster 11	-0.62	732	0.53

* Statistically significant at the $p \leq .01$.

6.5 Exploring the Predictive Power of the Threshold Regions

The predictive power of the eleven clusters was evaluated through ten separate multiple-regression analyses. Table 6.4 presents the results of these regression analyses. Column 1 indicates the item cluster, while column 2 reports the *R*-value, which measures the strength of the relationship between the item cluster and candidate ability. Column 3 presents the *R*-squared (R^2) value, which quantifies the proportion of the variance of the dependent variable explained by the cluster. Column 4 presents the standard error of the estimate (*SE*), which represents the standard error of the predicted candidate ability measures derived from the regression model. A lower *SE* indicates a more accurate prediction from the model. Columns 5 to 8 assess the statistical significance of the results, with the corresponding effect size (f^2) displayed in column 9. The last column presents whether the cluster meets the evaluation criteria for inclusion in the next step.

Table 6.4: Evaluating the predictive power of the item clusters (N=2,351)

	Summary of the regression models								
				Change Statistics					
Clusters	***R***	***R²***	**Std. Error of the Estimate**	***F* Change**	***d.f.1***	***d.f.2***	***Sig. F* Change**	***f²***	***outcome***
Cluster 1	.204	0.041	1.136	101.483	1	2349	<.001	0.04	Fail
Cluster 2	.702	0.493	0.828	227.233	10	2340	<.001	0.97	Pass
Cluster 3	.758	0.575	0.758	361.527	9	2341	<.001	1.35	Pass
Cluster 4	.852	0.725	0.610	385.446	16	2334	<.001	2.64	Pass
Cluster 5	.859	0.737	0.596	468.561	14	2336	<.001	2.81	Pass
Cluster 6	.896	0.802	0.517	557.255	17	2333	<.001	4.06	Pass
Cluster 7	.918	0.842	0.463	590.618	21	2329	<.001	5.33	Pass
Cluster 8	.895	0.802	0.519	428.006	22	2328	<.001	4.04	Pass
Cluster 9	.833	0.693	0.645	329.538	16	2334	<.001	2.26	Pass
Cluster 10	.521[a]	0.272	0.991	219.010	4	2346	<.001	0.37	Pass
Cluster 11	.391	0.153	1.069	140.867	3	2347	<.001	0.18	Fail

For an item cluster to form a threshold region, it should explain a statistically significant amount of candidate ability in a substantive way ($p<0.01$, $f^2 \geq 0.35$) and demonstrate a correlation of at least 0.50 ($R>0.50$, $p<0.01$, $R^2>0.26$). The effect size of R^2 is calculated using Cohen's (1988, 1992) formula, shown in equation 6:

$$f^2 = \frac{R^2}{1-R^2} \qquad \textit{Equation 6}$$

As shown in the last column, Clusters 1 and 11 did not meet the criteria and could not be included in the next step of the analysis, which involves locating the cut scores within the threshold regions.

6.6 Locating the Cut Scores within the Threshold Regions

Cut scores determined using the *Principled Cut Score* approach follow the same calculation methodology as the *ID Matching* method. Therefore, cut scores can be calculated by using one of the following methods: (1) the minimum, (2) the maximum, (3) the mean, or (4) the median of the item difficulties within the established threshold regions. Alternatively, cut

scores can be placed before, after, or at the mean between the last item of a threshold region and the first item of the subsequent region.

The cut scores in the ISE Digital listening module were determined based on the position of the DIALANG anchor items in the different item clusters. For instance, the C2 cut score was calculated by considering the item difficulties of items beyond the DIALANG C1 anchor item to ensure accurate differentiation between proficiency levels. The CEFR item difficulty scale, derived from a Rasch analysis, is proportional (ranging from -4 to 4). Therefore, cut scores obtained using a data-based scalar approach should be adjusted to ensure that each level occupies a proportional amount of space and that no CEFR level differs in width by an arbitrary amount. The ISE Digital listening item difficulty scale was also derived from a Rasch analysis anchored to the six CEFR levels as established through the standard setting of the DIALANG project, ensuring that items and candidates are placed on the same scale. Thus, the ISE Digital listening cut scores advanced by at least one logit as illustrated in Table 6.5) as a one-logit difference in a proportional scale can be equivalent to approximately a year of instruction (Linacre, 2022) in certain academic contexts or more.

Table 6.5: A summary of the listening module cut scores per CEFR level

Cluster	CEFR level	Measure
1	A1	-2.50
3	A2	-1.44
5	B1	-0.36
6	B2	0.79
9 (beginning)	C1	2.06
9 (end)	C2	3.24

6.7 Evaluating Cut Scores: Consistency Within the Method

As briefly explained in section 3.3, the method's consistency was evaluated by following the processes and procedures outlined in the Manual (Council of Europe, 2009). Hence, the engineered cut scores were evaluated for their i) precision, accuracy, and reliability and ii) classification consistency and accuracy. Following Kaftandjieva (2010), a dataset of 5,000 candidates was simulated based on the ability measures of the 2,359 candidates (the entire population) who had taken part in test trialling, using *Winsteps v5.8.3.0* (Linacre, 2024) to facilitate in-depth analyses of the cut scores. When the data set was analysed, one candidate was excluded because their data were unmeasurable. The psychometric properties of the real and simulated data were very close (see Table 6.6).

Table 6.6: Psychometric characteristics of real and simulated candidate population

Index	Real (N = 2,359)	Simulated (N = 4,999)
Number of items	258	258
Item difficulty mean measure (*SEm; SD*)	-0.56 (0.29; 1.29)	-0.41 (0.21; 1.37)
Candidate mean measure (*SEm; SD*)	-1.10 (0.48; 1.18)	-1.06 (0.50; 1.34)

Test reliability	0.83	0.85
RMSE	0.49	0.52
Mean score (*SD*)	12.7 (6.90)	12.4 (7.20)
Score min - max	2 - 38	0 - 40

Table 6.7 presents the results of the consistency within the method, as evaluated by the cut scores calculated from the clusters that met the criteria in Section 6.6 (above).

Table 6.7: Evaluating the accuracy & precision of the listening module cut scores (N = 4,999)

Cluster	CEFR level	Measure	SE_{jtt}	SD_{jtt}	*CSEM*	SE_{jtt} / SD_p	SE_{jtt} / *CSEM*	*CREL*
1	A1	-2.50	0.010	0.29	0.18	0.01	0.06	0.97
3	A2	-1.44	0.011	0.32	0.15	0.01	0.07	0.98
5	B1	-0.36	0.015	0.31	0.14	0.01	0.11	0.98
6	B2	0.79	0.031	0.3	0.16	0.03	0.19	0.97
9 (beginning)	C1	2.06	0.037	0.24	0.22	0.03	0.17	0.95
9 (end)	C2	3.24	0.197	0.39	0.34	0.17	0.59	0.90

The standard deviation of the test takers' measures (SD_{jtt}) and the standard error of their mean (SE_{jtt}) were both very small. As a result, the *SEjtt* of the calculated cut scores was less than one-third of the population standard deviation for each CEFR group ($SE_{jtt}/SD_p \leq 0.33$, $SD_p = 1.18$), indicating that the cut score errors are unlikely to affect the reliability of the ISE Digital listening module. This is further supported by the fact that SE_{jtt} was also below one-third of the conditional standard error of measurement (*CSEM*) for each cut score ($SE_{jtt}/CSEM \leq .33$), meeting the criterion proposed by Kaftandjieva (2010). Additionally, the CREL of each cut score ranged was higher than the .80 minimum recommended criterion for English language examinations (Nicewander, 2018, 2019). The *CREL* reached its optimal value when the cut score measures were closer to the population mean ability measure (-1.10). Overall, the error associated with each calculated cut score was small, adding only a small amount to candidate ability measures.

6.8 Evaluating Cut Scores: Decision Consistency

The calculated cut scores were further evaluated in terms of their classification accuracy [DA(γ)] and consistency [*DC*(*φ*)] using two methodologies: the Lee IRT-based method (Lee, 2008) with *IRT-CLASS v2* (Lee & Kolen, 2008) and Rudner (2001, 2005) IRT-based methods with the *cacIRT R package, v1.4* (Lathrop, 2015), respectively. Both evaluation methods employed the individual approach (P), which incorporated item parameters, candidate ability measures, and their standard errors (Lee, 2010).

Table 6.8 summarises the results from the two evaluation methods, illustrating the classification consistency and accuracy of the engineered cut scores for the ISE Digital listening module. The evaluation methods are listed in the first column, while the recommended cut scores are provided in the second column, expressed as ability measures (*βv*) in logits, with their respective scaled scores in brackets. The table reports decision accuracy [*DA*(*γ*)] and consistency [*DC*(*φ*)] in columns three and four, respectively, alongside the kappa coefficient in column five. The proportion of correct classifications by chance [pchance (*φc*)] is presented in column six, followed by the probability of misclassifications in column seven. The false positive

and false negative rates are also provided in columns eight and nine. It is important to note that some cells are blank because the *CacIRT* R package, v1.4 (Lathrop, 2015), does not calculate all indices.

Table 6.8: Evaluating the accuracy & precision of the listening cut scores (N = 4,999)

Method	Listening ability (βv) in logits (scaled score)	*DA (γ)*	*DC (φ)*	*Kappa (κ)*	*pchance (φ_c)*	Probability of misclassification	False positive rate	False negative rate
CEFR Level A1								
LL	-2.50 (5)	0.96	0.97	0.79	0.82	0.04	0.004	0.03
Rudner	-2.50 (5)	0.97	0.95			0.03		
CEFR Level A2								
LL	-1.44 (30)	0.94	0.95	0.88	0.51	0.06	0.05	0.005
Rudner	-1.44 (30)	0.94	0.92			0.06		
CEFR Level B1								
LL	-0.36 (55)	0.96	0.97	0.88	0.64	0.04	0.03	0.007
Rudner	-0.36 (55)	0.95	0.93			0.05		
CEFR Level B2								
LL	0.79 (80)	0.99	0.99	0.92	0.88	0.001	0.007	0.001
Rudner	0.79 (80)	0.99	0.99			0.01		
CEFR Level C1								
LL	2.06 (105)	0.99	0.99	0.85	0.96	0.005	0.003	0.001
Rudner	2.06 (105)	0.99	0.99			0.01		
CEFR Level C2								
LL	3.24 (130)	1.00	1.00	0.66	0.99	0.001	0.001	<0.0002
Rudner	3.24 (130)	1.00	1.00			0.001		

All *DA (γ)* and *DC (φ)* measures were higher than the recommended minimum criterion of .85 (Subkoviak, 1988) for certification examinations for each one of the CEFR levels and across all levels. Additionally, κ values were higher than the expected 0.60. Additionally, for the A2, B1, and B2 CEFR cut scores, the κ values were higher than pchance (φ_c). That the cut scores at the edges of the CEFR continuum fall below pchance (φ_c) is not surprising, because pchance (φ_c) typically increases when cut scores are placed towards the lower or upper end of the candidate ability measure range (Subkoviak, 1988). This is because the least and most able candidates perform similarly even in tests that are not parallel. It should be noted, however, that for all CEFR levels, κ is exceptionally high, indicating that candidate classification was determined by their performance on the ISE Digital listening module.

In summary, the ISE Digital listening module items were mapped to the CEFR in three ways: first, during the module's conceptualisation stage(Griffiths, 2023); second, during the item creation stage; and third, through standard setting using a *Principled Cut Score* approach (Kanistra, forthcoming) that incorporates elements from Philip's (2012) *Benchmark* standard setting method and North and Jones' (2009) data-driven scalar approach. Therefore, the ISE Digital listening module is aligned with the CEFR both qualitatively, in terms of content, and quantitatively, through the DIALANG CEFR scale.

7 Reading Cut Scores and Validity Evidence

Like the ISE Digital listening module, the CEFR cut scores for the reading module were established using the *Principled Cut Score* approach (Kanistra, forthcoming). The key steps in this approach (see Section 2.3 for more details) were listed at the start of Section 6 and were applied to both the listening and reading modules, as were the post hoc evaluation and validation checks. This section presents the reading module results for each step and reviews the internal validity evidence for this method, focusing on method consistency and classification decision reliability.

7.1 Psychometric Properties of the ISE Digital Reading Module

The reading module was analysed using Rasch measurement, with 565 contributing to the calibration of 247 items, including 15 DIALANG anchor items using *Winsteps v5.8.3.0* (Linacre, 2024). These anchor items ensured that the reading module scale was aligned to the CEFR proficiency continuum through an established instrument. A summary of this analysis outlining the psychometric characteristics of the reading items is shown in Table 7.1.

Table 7.1: Rasch summary statistics for the ISE Digital reading module

Index	Real (N = 565)
Number of items	247
Item difficulty mean measure (*SEm; SD*)	-0.79 (0.11;1.76)
Candidate mean measure (*SEm; SD*)	-0.42 (0.06; 1.42)
Test reliability	0.84
RMSE	0.63
Mean score (*SD*)	15.9 (6.90)
Score min - max	2 - 34

The item difficulty distribution (mean = –0.79 logits; *SD* = 1.76) indicates that the bank spans an appropriate range of difficulty for the ISE Digital cohort. Candidate measures averaged at 0.42 logits with a rather large standard deviation (*SD* = 1.42), reflecting a broad spread of candidate reading ability within the population.

Reliability was acceptable for a receptive skills assessment (0.84), and the RMSE of 0.63 logits indicates adequate measurement precision for interpreting reading proficiency on a Rasch scale. The observed score distribution (mean = 15.9; *SD* = 6.90; range = 2–34) suggests that the module captured a meaningful range of response patterns across candidates.

Overall, the distribution of item difficulties, candidate measures, reliability and measurement error indicates that the reading module provides a coherent and sufficiently precise measure of reading ability to support the CEFR standard setting procedure presented in the following sections.

7.2 Establishing the Predictive Power of Each Item

This is the first step in the *Principled Cut Score* approach, which involves linear regression to identify items that significantly explain candidate ability. The analysis was conducted on 247 items, including 15 DIALANG reading items (serving as CEFR-calibrated anchor items) and 232 from the ISE Digital reading module. Preliminary checks were performed to confirm the dataset's suitability for multiple regression, including exploring whether the assumptions of normality, linearity, multicollinearity, and homoscedasticity were not violated. These assumptions are crucial for ensuring the reliability and validity of the results (Pallant, 2016; Tabachnick & Fidell, 2014). Following guidelines in Tabachnick and Fidell (2014), candidates

with standardised residuals exceeding |3.3| were identified as outliers and removed, resulting in a sample of 539 candidates. The candidate ability measures (βv) were used as the dependent variable in the multiple regression analysis.

The full dataset of 247 items explained 98.8% of the variance in ability measures in a statistically significant way (adjusted R^2 = .995, F = 238.96, *d.f.1 = 247, d.f.2 = 291, p <.00, p* < .01). Of these, a subset of 31 items explained candidate ability variance in a statistically significant way (*Sig.*<0.01*).* These 31 items were used in the second step of the *Principled Cut Score* approach (Kanistra, forthcoming).

7.3 Converting Item Difficulty Measures to z-scores

As a reminder, after performing the linear regression, the dataset included performance data for 539 candidates, each with a candidate ability measure, and 31 items with associated difficulty measures. In this step, item difficulty measures are converted to z-scores. As explained in Section 6.2, z-scores provide a quick sense of where a score (especially a proposed cut score) lies relative to the overall ability of the candidate population. They also help identify extreme values. Therefore, the goal of converting difficulty measures to z-scores is to determine potential cut score locations relative to the population mean (βv = -0.423). This is essential because it prevents placing cut scores at the extreme ends of the item difficulty scale, where candidate classification would largely depend on chance (Subkoviak 1980, 1988).

Since ISE Digital is a multilevel test targeting different CEFR levels, the z-score conversion should also be interpreted within this broader context. As a reminder, the anchoring was based on the DIALANG standard setting (Alderson, 2005), with item difficulties derived from that procedure. This ensured that the means of each DIALANG-referenced CEFR level served as CEFR benchmarks. As a result, the z-score distances were examined not only relative to the overall population mean, but also against the expected mean performance at each CEFR level (derived from the DIALANG anchor items). In this way, the inspection of z scores supported the placement of cut scores at relevant points on the logit scale, maximising classification consistency (DC, φ) and κ coefficients in line with Subkoviak's (1988) recommendations. Thus, as an additional check, the distances between potential cut scores and the DIALANG CEFR level means were also examined and are reported in Table 7.2.

Table 7.2 presents the distance of the possible cut scores from the population mean ability measures. Columns 1 and 2 display the item IDs and their associated item difficulties (δ). Columns 3 to 9 present the associated z-scores and the context in which they are examined (mean candidate ability or CEFR level).

Table 7.2: Cut score position relative to population and DIALANG Reading means

Item ID	Item difficulty (δ)	z-score candidate ability mean	z-score DIALANG A1	z-score DIALANG A2	z-score DIALANG B1	z-score DIALANG B2	z-score DIALANG C1	z-score DIALANG C2
Item 09	-3.30	-2.03	0.39	-0.46	-1.07	-1.87	-2.42	-2.50
Item 10	-3.17	-2.87	0.45	-0.41	-1.02	-1.81	-2.36	-2.45
Item 12	-3.14	-2.84	0.46	-0.40	-1.01	-1.80	-2.35	-2.44
Item 13	-2.80	-2.51	0.60	-0.26	-0.87	-1.66	-2.21	-2.30
Item 28	-1.85	-1.55	0.99	0.13	-0.48	-1.28	-1.82	-1.91
Item 34	-1.05	-0.75	1.31	0.45	-0.16	-0.95	-1.50	-1.59
Item 46	-0.97	-0.68	1.34	0.48	-0.12	-0.92	-1.47	-1.56
Item 51	-0.62	-0.32	1.48	0.63	0.02	-0.78	-1.33	-1.41
Item 73	-0.46	-0.16	1.55	0.69	0.08	-0.71	-1.26	-1.35
Item 82	-0.17	0.13	1.67	0.81	0.20	-0.59	-1.14	-1.23

Item ID	Item difficulty (*δ*)	z-score candidate ability mean	z-score DIALANG A1	z-score DIALANG A2	z-score DIALANG B1	z-score DIALANG B2	z-score DIALANG C1	z-score DIALANG C2
Item 91	-0.04	0.26	1.72	0.87	0.26	-0.54	-1.09	-1.17
Item 92	0.86	1.16	2.09	1.23	0.62	-0.17	-0.72	-0.81
Item 93	0.86	1.16	2.09	1.23	0.62	-0.17	-0.72	-0.81
Item 104	1.01	1.31	2.15	1.29	0.68	-0.11	-0.66	-0.75
DIALANG_B2/B2+	1.12	1.42	2.19	1.34	0.73	-0.07	-0.62	-0.70
DIALANG_B2/B2+	1.12	1.42	2.19	1.34	0.73	-0.07	-0.62	-0.70
Item 114	1.15	1.45	2.20	1.35	0.74	-0.06	-0.61	-0.69
Item 148	1.44	1.74	2.32	1.47	0.86	0.06	-0.49	-0.57
Item 150	1.57	1.86	2.37	1.52	0.91	0.11	-0.44	-0.52
Item 159	1.60	1.89	2.39	1.53	0.92	0.13	-0.42	-0.51
Item 183	1.70	2.00	2.43	1.57	0.96	0.17	-0.38	-0.47
Item 200	1.91	2.21	2.52	1.66	1.05	0.26	-0.29	-0.38
DIALANG_B2+	2.03	2.33	2.56	1.71	1.10	0.30	-0.25	-0.33
Item 205	2.33	2.63	2.69	1.83	1.22	0.43	-0.12	-0.21
Item 206	2.47	2.77	2.74	1.88	1.28	0.48	-0.07	-0.16
Item 207	2.51	2.81	2.76	1.90	1.29	0.50	-0.05	-0.14
Item 223	2.61	2.91	2.80	1.94	1.33	0.54	-0.01	-0.10
DIALANG_C1	2.64	2.93	2.81	1.95	1.34	0.55	0.00	-0.09
Item 230	2.64	2.94	2.81	1.95	1.35	0.55	0.00	-0.09
Item 231	2.74	3.04	2.85	2.00	1.39	0.59	0.04	-0.04
Item 238	2.95	3.25	2.94	2.08	1.47	0.68	0.13	0.04
Item 245	3.75	4.04	3.26	2.40	1.80	1.00	0.45	0.36

The population-based z scores were calculated using the overall mean ability measure (βv = -0.267) and its corresponding standard deviation (SD = 1.323) through Equation 6.

$$Z = \frac{(c-M)}{SD} \qquad \textit{Equation 7}$$

Where c is the cut score location, M is the population mean, and SD is the standard deviation of the test-ability measures. For the z-scores comparing the item difficulties in the context of each CEFR level, the mean and standard deviation of the relevant CEFR anchor items (from the DIALANG standard setting) were used in equation 6.

Subkoviak (1988) recommends that, for high-stakes certification testing, κ coefficients should range from 0.58 to 0.70, with decision consistency (*DC, φ*) values of at least 0.86. The z-scores reported in Table 7.2 ranged from |0.6| to |4.04| (highlighted cells in Table 7.2) and fall within acceptable limits when interpreted through Subkoviak's reference tables (1988, pp. 49). In other words, assuming a test reliability above 0.80, these z-scores correspond to κ values between .58 and .71 and φ values between .86 and .98. This suggests that, in principle, the indices targeted in this study can be obtained at levels appropriate for a high-stakes test. Nonetheless, it remains advisable to avoid placing cut scores at the extremes of the item-difficulty distribution to maintain interpretive stability.

7.4 Establishing Item Clusters

In the ID Matching method, the threshold region comprises items arranged in a consecutive ascending order of difficulty. To preserve this property in the engineered threshold region, a series of Wald *t-tests* were conducted to identify clusters of items with comparable difficulty, comparing the first item to the second, then to the third, and so on. The significance level was set at .01 to reduce the risk of falsely rejecting the null hypothesis due to the high number of comparisons (Cohen 1988, 1992). Table 8 presents the results of the Wald statistics analyses. The items being compared are listed in columns 1 and 2, with column 3 indicating whether they would be grouped into the same cluster. Column 4 reports the Wald statistic for each comparison along with its degrees of freedom in column 5. Column 6 indicates whether the Wald statistic is statistically significant.

Table 7.3: Using Wald statistics to establish item clusters for the reading module

Item ID	Item ID compared	Cluster	Wald t (*d.f.* 1076)	Significance (two-tailed prob.)[3]
Item 09	Item 10	Cluster 1	-0.57	0.57
Item 09	Item 12	Cluster 1	-0.69	0.49
Item 09	Item 13	Cluster 1	-2.26	0.02
Item 09	Item 28	End of Cluster 1	-7.20	0.00
Item 28	Item 34	End of Cluster 2	-5.15	0.00
Item 34	Item 46	Cluster 3	-0.51	0.61
Item 34	Item 51	End of Cluster 3	-2.95	0.00
Item 51	Item 73	Cluster 4	-1.14	0.25
Item 51	Item 82	End of Cluster 4	-3.19	0.00
Item 82	Item 91	Cluster 5	-0.92	0.36
Item 82	Item 92	End of Cluster 5	-6.97	0.00
Item 92	Item 93	Cluster 6	0.00	1.00
Item 92	Item 104	Cluster 6	-0.99	0.32
Item 92	DIALANG_B2/B2+	Cluster 6	-1.68	0.09
Item 92	DIALANG_B2/B2+	Cluster 6	-1.68	0.09
Item 92	Item 114	Cluster 6	-1.86	0.06
Item 92	Item 148	End of Cluster 6	-3.62	0.00
Item 148	Item 150	Cluster 7	-0.76	0.45
Item 148	Item 159	Cluster 7	-0.93	0.45
Item 148	Item 183	Cluster 7	-1.54	0.12
Item 148	Item 200	End of Cluster 7	-2.70	0.01
Item 200	DIALANG_6B2+	Cluster 8	-0.64	0.52
Item 200	Item 205	Cluster 8	-2.14	0.03
Item 200	Item 206	End of Cluster 8	-2.72	0.01

[3] Statistically significant at the $p \leq .01$

Item ID	Item ID compared	Cluster	Wald t (*d.f.* 1076)	Significance (two-tailed prob.)[3]
Item 206	Item 207	Cluster 9	-0.22	0.83
Item 206	Item 223	Cluster 9	-0.66	0.51
Item 206	DIALANG_C1	Cluster 9	-0.77	0.44
Item 206	Item 230	Cluster 9	-0.77	0.44
Item 206	Item 231	Cluster 9	-1.23	0.22
Item 206	Item 238	Cluster 9	-2.07	0.04
Item 206	Item 245	End of Cluster 9	-4.51	0.00
Item 245		Cluster 10		

7.5 Exploring the Predictive Power of Threshold Regions

An item cluster can form a threshold region only if it explains a substantive amount of candidate ability (R>0.50, p <.01, R^2>0.26) in a statistically significant way (p< 0.01, f^2≥ 0.35). The predictive power of the ten clusters was evaluated by conducting seven separate multiple-regression analyses.

Table 7.4: Evaluating the predictive power of the reading item clusters (N=539)

Summary of the regression models									
				Change Statistics					
Clusters	***R***	***R²***	**Std. Error of the Estimate**	***F* Change**	***d.f.1***	***d.f.2***	***Sig. F* Change**	***f²***	***outcome***
Cluster 1	.676[a]	0.456	0.979973	112.096	4	534	<.001	0.84	Pass
Cluster 2	.339a	0.115	1.247131	69.576	1	537	<.001	0.13	Fail
Cluster 3	.575[a]	0.331	1.085482	132.345	2	536	<.001	0.49	Pass
Cluster 4	.535[a]	0.287	1.120572	107.664	2	536	<.001	0.40	Pass
Cluster 5	.614[a]	0.377	1.047195	162.154	2	536	<.001	0.61	Pass
Cluster 6	.718[a]	0.516	0.92692	94.343	6	532	<.001	1.07	Pass
Cluster 7	.666[a]	0.443	0.992084	106.136	4	534	<.001	0.80	Pass
Cluster 8	.572[a]	0.328	1.088896	86.892	3	535	<.001	0.49	Pass
Cluster 9	.784[a]	0.615	0.826891	121.256	7	531	<.001	1.60	Pass
Cluster 10	.454[a]	0.206	1.180928	139.492	1	537	<.001	0.26	Fail

Table 7.4 presents the results of these regression analyses. Column 1 indicates the item cluster, while column 2 reports the *R*-value, which measures the strength of the relationship between the item cluster and candidate ability. Column 3 presents the *R*-squared (R^2) value, which quantifies the proportion of the variance of the dependent variable explained by the cluster. Column 4 presents the standard error of the estimate (*SE*), which represents the standard error of the predicted candidate ability measures derived from the regression model. A lower *SE* indicates a more accurate model prediction. Columns 5 to 8 evaluate the statistical significance of the results, with the corresponding effect size (f^2) displayed in column 9. The last column presents whether the cluster meets the evaluation criteria for inclusion in the next step. Applying the same criteria used for the listening module (see Section 6.4), Clusters 2 and 10 did not statistically significantly predict candidate ability, indicating that cut scores could not be set within these clusters.

7.6 Locating the Cut Scores Within the Threshold Regions

As stated in Section 7.6, cut scores determined using the *Principled Cut Score* approach follow the same calculation methodology as the *ID Matching* method. Therefore, cut scores can be calculated by using one of the following methods: (1) the minimum, (2) the maximum, (3) the mean, or (4) the median of the item difficulties within the established threshold regions. Alternatively, cut scores can be placed before, after, or at the mean between the last item of a threshold region and the first item of the subsequent region.

The cut scores in the ISE Digital reading module were determined based on the position of the DIALANG anchor items in the different item clusters. For instance, the C2 cut score was calculated by considering the item difficulties of items beyond the DIALANG C1 anchor item to ensure accurate differentiation between proficiency levels. The CEFR item difficulty scale, derived from a Rasch analysis, is proportional (ranging from -4 to 4). Therefore, cut scores obtained using a data-based scalar approach should be adjusted to ensure that each level occupies a proportional amount of space and that no CEFR level differs in width by an arbitrary amount. The ISE Digital reading item difficulty scale was also derived from a Rasch analysis, anchored to the six CEFR levels established through the DIALANG standard setting study, ensuring that items and candidates were placed on this common scale. Thus, the ISE Digital reading cut scores advanced by at least one logit (see Table 7.5) as a one-logit difference in a proportional scale can be equivalent to approximately a year of instruction (Linacre, 2022) in certain academic contexts or more.

Table 7.5: A summary of the listening module cut scores per CEFR level

Cluster	CEFR level	Measure
1	A1	-2.33
3	A2	-1.05
5	B1	-0.04
6	B2	1.10
9 (beginning)	C1	2.18
9 (end)	C2	3.35

7.7 Evaluating Cut Scores: Consistency Within the Method

As briefly explained in section 3.3, the consistency within the method was evaluated by following the processes and procedures outlined in the Manual (Council of Europe, 2009). Hence, the calculated cut scores were evaluated for their i) precision, accuracy, and reliability and ii) classification consistency and accuracy. Following Kaftandjieva (2010), a dataset of 5,000 candidates was simulated based on the ability measures of the 565 candidates who had taken part in test trialling, using *Winsteps v5.8.3.0* (Linacre, 2024) to facilitate the in-depth analyses of the cut cores. When the data set was analysed, 59 test takers were excluded because they were unmeasurable. The psychometric properties of the real and simulated data were very close (see Table 7.6).

Table 7.6: Psychometric characteristics of real & simulated candidate population - reading

Index	Real (N = 565)	Simulated (N = 4,941)
Number of items	247	247
Item difficulty mean measure (*SEm; SD*)	-0.79 (0.11;1.76)	-0.74 (0.12; 1.93)
Candidate mean measure (*SEm; SD*)	-0.42 (0.06; 1.42)	-0.43 (0.02; 1.56)
Test reliability	0.84	0.85
RMSE	0.63	0.62
Mean score (*SD*)	15.9 (6.90)	15.9 (7.20)
Score min - max	2 - 34	0 - 34

Table 7.7 presents the results of the method's consistency, as evaluated by the cut scores calculated from the clusters that met the criteria in section 7.6 above.

Table 7.7: Evaluating the accuracy & precision of the reading module cut scores (N = 4,941)

Cluster	CEFR level	Measure	SE_{jtt}	SD_{jtt}	CSEM	SE_{jtt} / SD_p	SE_{jtt} / CSEM	CREL
1	A1	-2.33	0.01	0.36	0.23	0.01	0.04	0.95
3	A2	-1.05	0.01	0.28	0.15	0.01	0.07	0.98
5	B1	-0.04	0.01	0.32	0.13	0.01	0.08	0.98
6	B2	1.10	0.01	0.28	0.15	0.01	0.07	0.98
9 (beginning)	C1	2.18	0.03	0.34	0.22	0.02	0.14	0.95
9 (end)	C2	3.35	0.11	0.52	0.38	0.07	0.29	0.87

The standard deviation of the test takers' measures (SD_{jtt}) and the standard error of their mean (SE_{jtt}) were both very small. As a result, the *SEjtt* of the calculated cut scores was less than one-third of the population standard deviation for each CEFR group (SE_{jtt} / $SD_p \leq 0.33$; $SD_p = 1.56$), indicating that the cut score errors are unlikely to affect the reliability of the ISE Digital listening module. This is further supported by the fact that SEjtt was also below one-third of the conditional standard error of measurement (*CSEM*) for each cut score (SE_{jtt} / *CSEM* ≤ .33), meeting the criterion proposed by Kaftandjieva (2010).

Additionally, the CREL of each cut score ranged was higher than the .80 minimum recommended criterion for English language examinations (Nicewander, 2018, 2019). The *CREL* reached its optimal value when the cut-score measures were closer to the population mean ability measure (-0.43). Overall, the error associated with each calculated cut score was small, thus adding only a small amount of error to candidate ability measures.

7.8 Evaluating Cut Scores: Decision Consistency

The calculated cut scores were further evaluated in terms of their classification accuracy [DA(γ)] and consistency [*DC*(*φ*)] using two methodologies: the Lee IRT-based method (Lee, 2008) with *IRT-CLASS v2* (Lee & Kolen, 2008) and Rudner (2001, 2005) IRT-based methods with *cacIRT R package, v1.4* (Lathrop, 2015), respectively. Both evaluation methods employed the individual approach (P), which incorporated item parameters, candidate ability measures and their standard errors (Lee, 2010).

Table 7.8 summarises the results from the two evaluation methods, illustrating the classification consistency and accuracy of the calculated cut scores for the ISE Digital reading module. The evaluation methods are listed in the first column, while the recommended cut scores are provided in the second column, expressed as ability measures (β_v) in logits, with the scaled scores reported in brackets. The table reports decision accuracy [*DA*(*γ*)] and consistency [*DC*(*φ*)] in columns three and four, respectively, alongside the kappa coefficient in column five. The proportion of correct classifications by chance [*pchance* (φ_C)] is presented in column six, followed by the probability of misclassifications in column seven. The false-positive and false-negative rates are also provided in columns eight and nine. *The CacIRT R package (v1.4*; Lathrop, 2015) does not calculate all indices.

Table 7.8: Evaluating the DA and DC of the reading cut scores (N = 4,941).

Method	Reading ability (β_v) in logits (scaled score)	*DA (γ)*	*DC (φ)*	*Kappa (κ)*	*pchance (φ_c)*	Probability of misclassification	False positive rate	False negative rate
CEFR Level A1								
LL	-2.33 (5)	0.99	0.98	0.83	0.90	0.02	0.004	0.01
Rudner	-2.33 (5)	0.98	0.97			0.03		
CEFR Level A2								
LL	-1.05 (30)	0.97	0.96	0.90	0.58	0.04	0.02	0.01
Rudner	-1.05 (30)	0.97	0.96			0.04		
CEFR Level B1								
LL	-0.04 (55)	0.96	0.95	0.90	0.51	0.05	0.03	0.01
Rudner	-0.04 (55)	0.97	0.96			0.04		
CEFR Level B2								
LL	1.10 (80)	0.96	0.96	0.85	0.72	0.04	0.03	0.005
Rudner	1.10 (80)	0.97	0.96			0.04		
CEFR Level C1								
LL	2.18 (105)	0.99	0.98	0.76	0.94	0.01	0.01	0.001
Rudner	2.18 (105)	0.99	0.99			0.01		
CEFR Level C2								
LL	3.35 (130)	1.00	1.00	0.80	0.99	0.002	0.001	0.0002
Rudner	3.35 (130)	0.99	0.99			0.01		

For each one of the CEFR levels, all *DA (γ)* and *DC (φ)* measures were higher than the recommended minimum criterion of 0.85 (Subkoviak 1988) for certification examinations. Additionally, κ values were higher than the expected 0.60, higher than pchance (φ_c). That the cut scores at the edges of the CEFR continuum fall below pchance (φ_c) is not surprising, because pchance (φ_c) typically increases when cut scores are placed towards the lower or upper end of the candidate ability measure range (Subkoviak, 1988). This is because the least and most able candidates perform similarly even in tests that are not parallel. It should be noted, however, that for all CEFR levels, κ is exceptionally high, indicating that candidate classification was determined by their performance on the ISE Digital reading module.

Summing up, the ISE Digital reading module items were mapped to the CEFR in three ways: first, during the module's conceptualisation stage (Griffiths, 2023); second, during the item creation stage; and third, through the standard setting employing a *Principled Cut Score* approach (Kanistra, forthcoming). Therefore, the ISE Digital reading module is aligned to the CEFR both qualitatively, in terms of content, and quantitatively through the DIALANG CEFR scale.

8 Validating the Writing Standard-Setting Workshop and Cut Scores

This section presents the results for the ISE Digital writing module, applying the evaluation framework presented in Section 3.1 as well as validity evidence for the defensibility of the cut scores.

8.1 Psychometric Properties of the ISE Digital Writing Module

The Writing module was analysed using Many-Faceted Rasch Measurement (MFRM) in *Facets v4.4.4* (Linacre, 2025), with 490 candidates and 33 task-level observations included in the calibration. The task sets reflect the structure of the speaking module, with tasks drawn from multiple operational forms to ensure sufficient connectivity for stable estimation. The results of this analysis are summarised in Table 8.1.

Table 8.1: Rasch summary statistics for the ISE Digital writing module

Index	Real (N = 490)
Number of tasks	33
Candidate mean measure (*SEm; SD*)	0.62 (0.45; 2.10)
Test reliability	0.94
RMSE (CSEM)	0.51
SEM	2.33
Observed average (*SD*)	3.13 (0.97)
Fair average (*SD*)	3.41 (0.99)

The candidate mean measure was 0.62 logits (*SEm* = 0.45; *SD* = 2.10), indicating a wide range of writing proficiency within the sample. Reliability was high (0.94), indicating a strong, highly reliable separation of candidate abilities. The *RMSE* (*CSEM*) of 0.51 logits and the standard error of measurement (*SEM)* of 2.33 indicate acceptable measurement precision for a performance-based writing assessment. The observed average score (3.13; SD = 0.97) and the fair average (3.41; SD = 0.99) were very close, suggesting that examiners demonstrated comparable levels of severity across tasks in line with the expectations of the Rasch model.

Overall, these indices show that the ISE Digital writing module functions as a coherent and reliable measure of writing proficiency and provides adequate technical support for the standard-setting procedures presented in the following sections.

8.2 Procedural Validity

The evaluation questionnaires were adapted from Cizek (2012, pp. 174-178). To align with the context of this study, some questions were modified. The surveys were administered after the *orientation* and *training in the method* phases of the writing standard setting workshop.

8.2.1 Evaluating the orientation and training in the method stages

The panellists were asked to rate the extent to which they agreed with the 14 statements in the survey. Figure 8.1 presents the survey statements and the analyses of this evaluation questionnaire. The bar graph displays the number of panellists who endorsed or opposed each statement, while the axis indicates the total number of panellists. Before proceeding to the next workshop stage, the facilitator reviewed the survey responses and addressed any reported issues before starting the standard-setting tasks.

The evaluation results for the *orientation and training in the method* stages of the writing standard setting workshop demonstrate strong positive feedback. Panellists '*agreed*' or '*strongly agreed*' that the orientation provided a clear overview, addressed questions effectively, and facilitated understanding of the standard-setting process (Q01–Q03). The timing and pace of the sessions were widely endorsed as appropriate (Q04). The CEFR familiarisation activities were effective in enhancing the understanding of CEFR levels and descriptors whilst refreshing their prior knowledge (Q05–Q08).

Panellists also noted that the test-taking experience clarified the difficulty and structure of the ISE Digital writing module of the exam (Q09). The *training in the method* was clear, and the practice activities effectively supported the application of the standard setting method (Q10–Q11). Most panellists expressed confidence in their role and ability to apply the method (Q12–Q13) and felt adequately prepared to begin the standard-setting tasks (Q14). The minor reservations were limited to two panellists (J13 and J09) and did not significantly detract from overall confidence. J13 expressed his reservation because s/he was not a trained examiner but acknowledged that "*the training has been useful in explaining a method that can be applied, and I am confident that I can apply this method to match tasks and output to CEFR levels.*" J09 expressed cautious confidence, noting the challenge of standard setting tasks and student performances in a language other than the one used in the examinations in their own context.

These findings underscore the success of the orientation and training sessions in fostering a deep understanding of CEFR descriptors and the alignment and benchmarking process to be followed while building participants' confidence to undertake these tasks effectively.

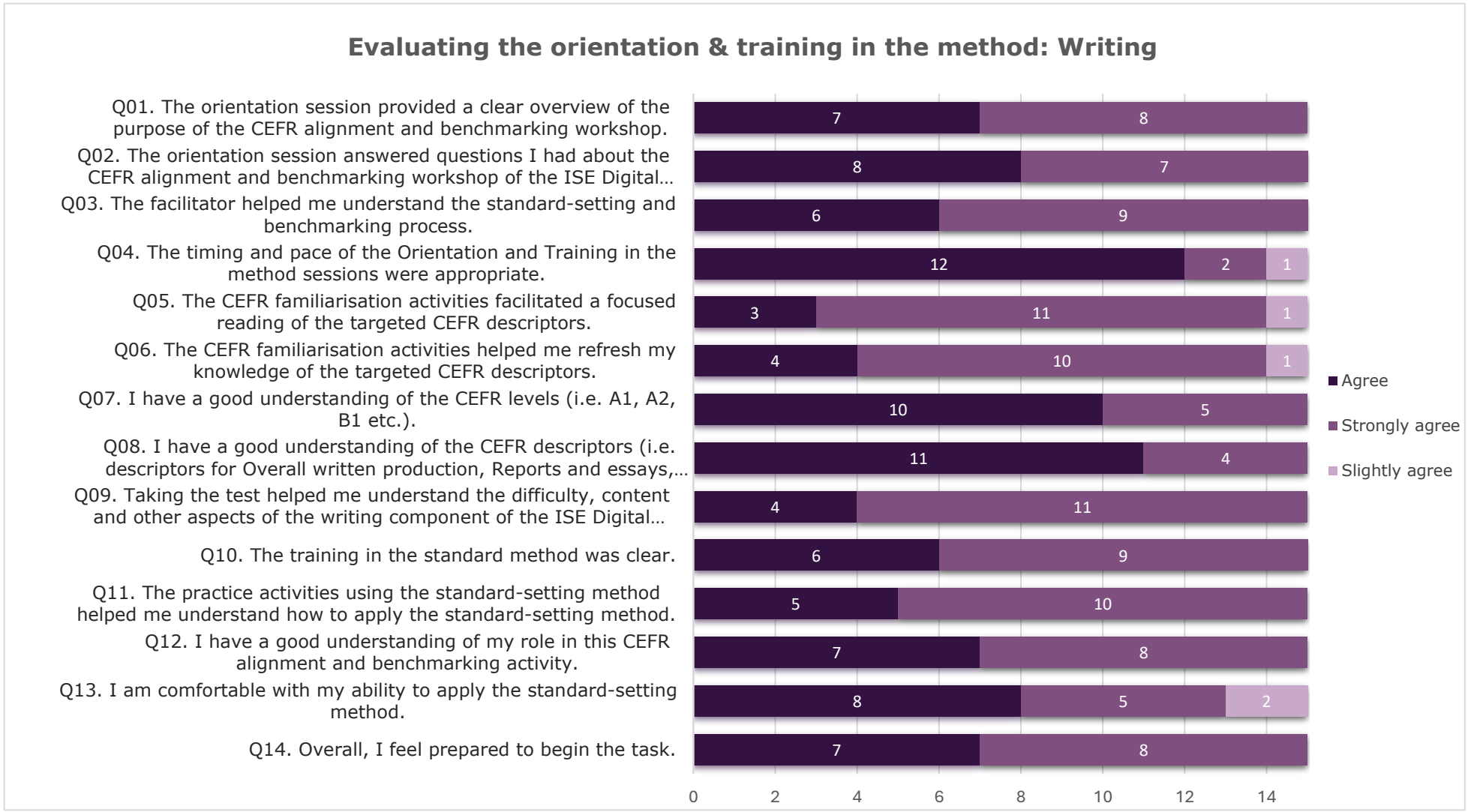


Figure 8.1: Evaluation of the orientation & training in the method stages – writing

8.2.2 Evaluating the writing standard setting and benchmarking workshop

Panellists were asked to rate their agreement with nine statements in the evaluation survey. The results, depicted in Figure 8.2, show the distribution of panellist endorsements and oppositions for each statement. The bar graph illustrates the responses, with the axis indicating the total number of panellists. The last two questions of this survey served as a Round 3, as they enabled the judges to reflect on and review the performances deemed representative of the different targeted CEFR levels and to change their judgements.

The survey results indicate overall positive feedback from the panellists regarding the writing workshop. Most panellists "strongly agree" or "agree" that the standard-setting procedures enabled them to effectively map writing tasks and responses to the targeted CEFR levels (Q02, Q03). The facilitator's role was highly appreciated, as it ensured inclusive and balanced discussions (Q04, Q05). Panellists also felt confident in their ratings (Q01) and found other panellists' ratings helpful in informing their judgments (Q06, Q07). Additionally, the group-recommended CEFR classifications for Tasks 1 and 2 were widely endorsed as reflective of the minimum performance levels for the targeted CEFR standards (Q08, Q09). Only two panellists modified some of their ratings during this final stage. These changes were reflected in the analysis of the standard-setting data.

Table 8.2: Factors affecting panellists' judgements – writing

Influential factors	Sum	Rank
Q10.4. The students' written responses.	58	1
Q10.3. The CEFR level descriptors & Written Assessment Grid.	56	2
Q10.5. The group discussion.	53	3
Q10.1. My experience taking the test.	48	4
Q10.6. Other judges' ratings.	45	5
Q10.2. My own experiences with real students.	34	6

The final question of the evaluation questionnaire required panellists to arrange the factors that influenced their judgments during the writing standard setting and benchmarking workshop in order of importance.

When evaluating candidates' written performances, the panellists primarily focused on the written responses, which received the highest attention score (58). The CEFR level descriptors and *Written Assessment Criteria Grid* ranked second (56), demonstrating their importance as a reference framework for evaluation, which is highly desirable in CEFR alignment studies. Group discussions (53) were also critical, suggesting the value of the between-rounds discussion in standard-setting, fostering collective judgment and collaboration. Panellists paid moderate attention to their experience taking the test (48) and to other judges' ratings (45), suggesting a balance between personal insights and peer evaluations. The least focus was placed on their own experiences with real students (34), indicating that while relevant, it was not as influential as the other factors during the evaluation process.

To summarise, the analysis of the two evaluation surveys indicates that the panellists were confident in their understanding and application of the standard-setting method and process. The data further show that the training and familiarisation activities effectively clarified the CEFR descriptors. The online standard-setting workshop promoted effective collaboration and group discussion. The panellists prioritised the direct assessment of the candidates' written responses, guided by the CEFR descriptors and the *Written Assessment Criteria Grid*, over personal experiences and peer ratings. The opinions and comments of the participants do not suggest any errors in the implementation of the standard setting method for the Writing module of the ISE Digital examinations. Therefore, it is reasonable to conclude that no systematic errors were introduced during the standard-setting process, which could have potentially invalidated the workshop results.

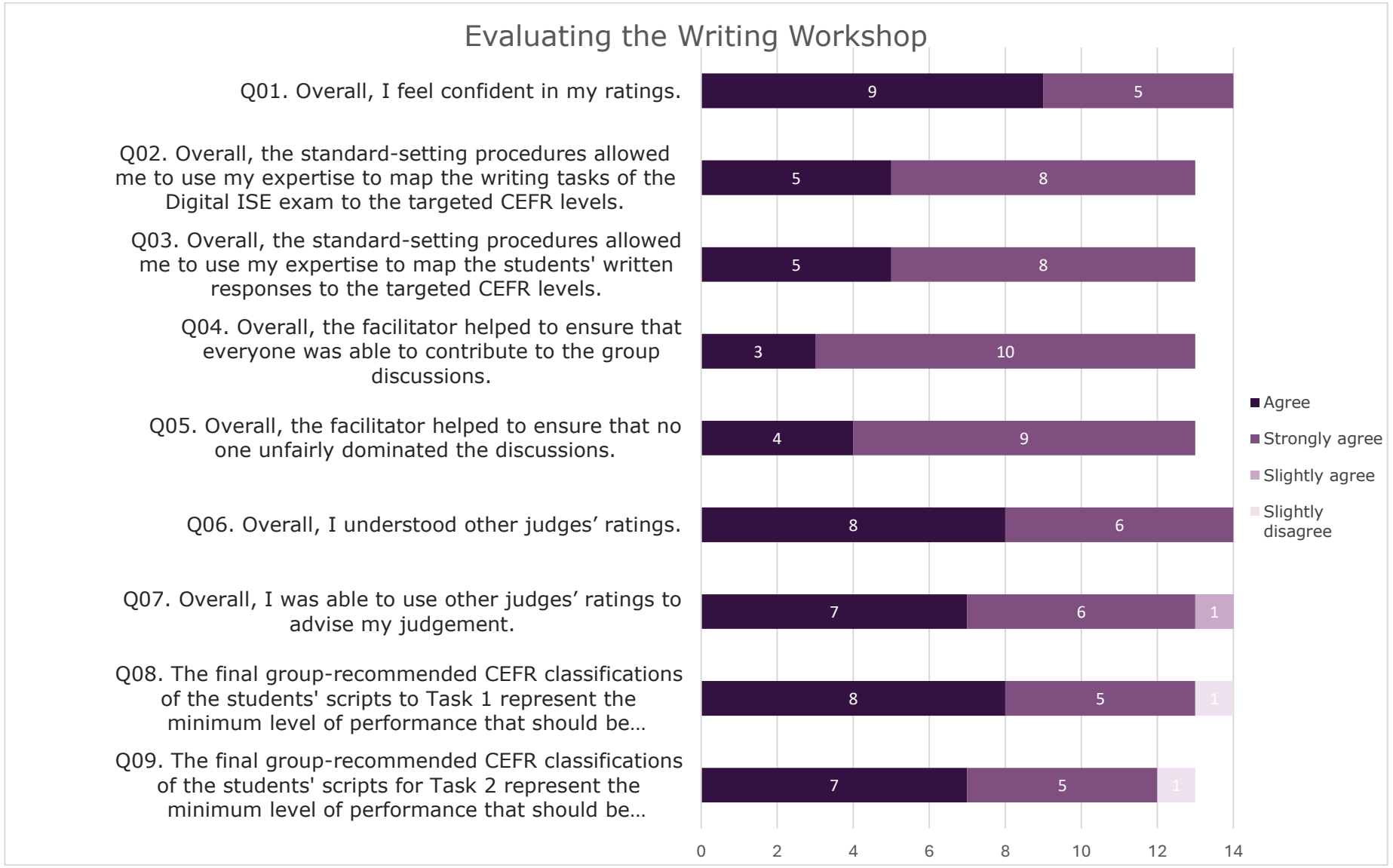


Figure 8.2: Evaluating the writing workshop

8.3 Evaluating the Writing Tasks

As explained in Section 2.2, Trinity's test development team drew extensively from the relevant theories about writing and writing proficiency. The design process was also aligned with the PADDI process (Ferrara, Lai, & Nichols, 2016) and closely referenced the CEFR framework. Trinity's item creation procedures follow UATD principles (Kanistra, forthcoming), instructing item writers to design tasks targeting the KSAs associated with specific CEFR levels while ensuring input materials meet the CEFR level requirements and readability indices.

The *written online communication* tasks are written to target A2-C1 CEFR levels, while remaining accessible to candidates of all proficiencies. The *writing from sources* tasks target B1(+) to C2 CEFR levels, while remaining accessible for B1 candidates. The outcomes of this principled approach to item creation are reflected in the content analysis forms provided in the Manual (Council of Europe, 2009, Appendix A). Panellists were asked to evaluate five *written online communication* tasks and three *writing from sources* tasks. The panellists were first asked to analyse the tasks' cognitive and linguistic demands and then map them to the CEFR level(s) A1-A2, B1-B2, or C1-C2, reflecting the routes the candidates would follow in the adaptive setting of the test. Panellists were asked to rationalise their judgements by selecting the specific CEFR scales (Table 8.3) that best operationalised such demands (Harsch & Kanistra, 2020).

Figures 8.3 (*written online communication*) and 8.4 (*writing from sources*) show the panellists' CEFR item ratings of the tasks and how they aligned to the CEFR. Table 8.2 explains the acronyms used.

Table 8.3: Acronyms used for the CEFR writing scales

CEFR scale	Acronym
Overall Written Production	OWP
Overall Written Interaction	OWI
Correspondence	Corres.
Online Conversation and Discussion	OC&D
Facilitating Collaborative Interaction with Peers	FCIwP
Collaborating to Construct Meaning	CtCM
Goal-oriented online transactions and collaboration	GOOT&C
Overall Mediation	OM
Written Reports and Essays	RE
Relaying Specific Information	RSI
Explaining Data (in Writing)	ED in W
Processing Text (in Writing)	PT in W

8.3.1 Written online communication tasks

Tasks 1 and 2 dealt with very concrete topics and were written mainly to target A2-B1 levels. Panellists' ratings (Figure 8.3) and comments validated Trinity's intended alignment. For example, J07 stated:

> *"To me, this task can be placed at the A2-B1 band. It can be responded to with a good A2 language sufficiently, but it will be too difficult for the A1 level. B1 can comfortably respond to this task, giving an adequate response where, whereas it may be limiting for a B2 speaker to reflect B2-level language features (although by forcing the limits of the task, a response at the B2 level can be provided). That's why I placed this task at B level rather than A level. I would place it in the A2-B1 band if there was one."*

Another panellist (J09), to give another example, expressed similar views for the same tasks:

> *"The prompt is, in my opinion, more of an A2 task, but the points you have to answer (if and how to improve) elevate the overall requirements of the task to B1."*

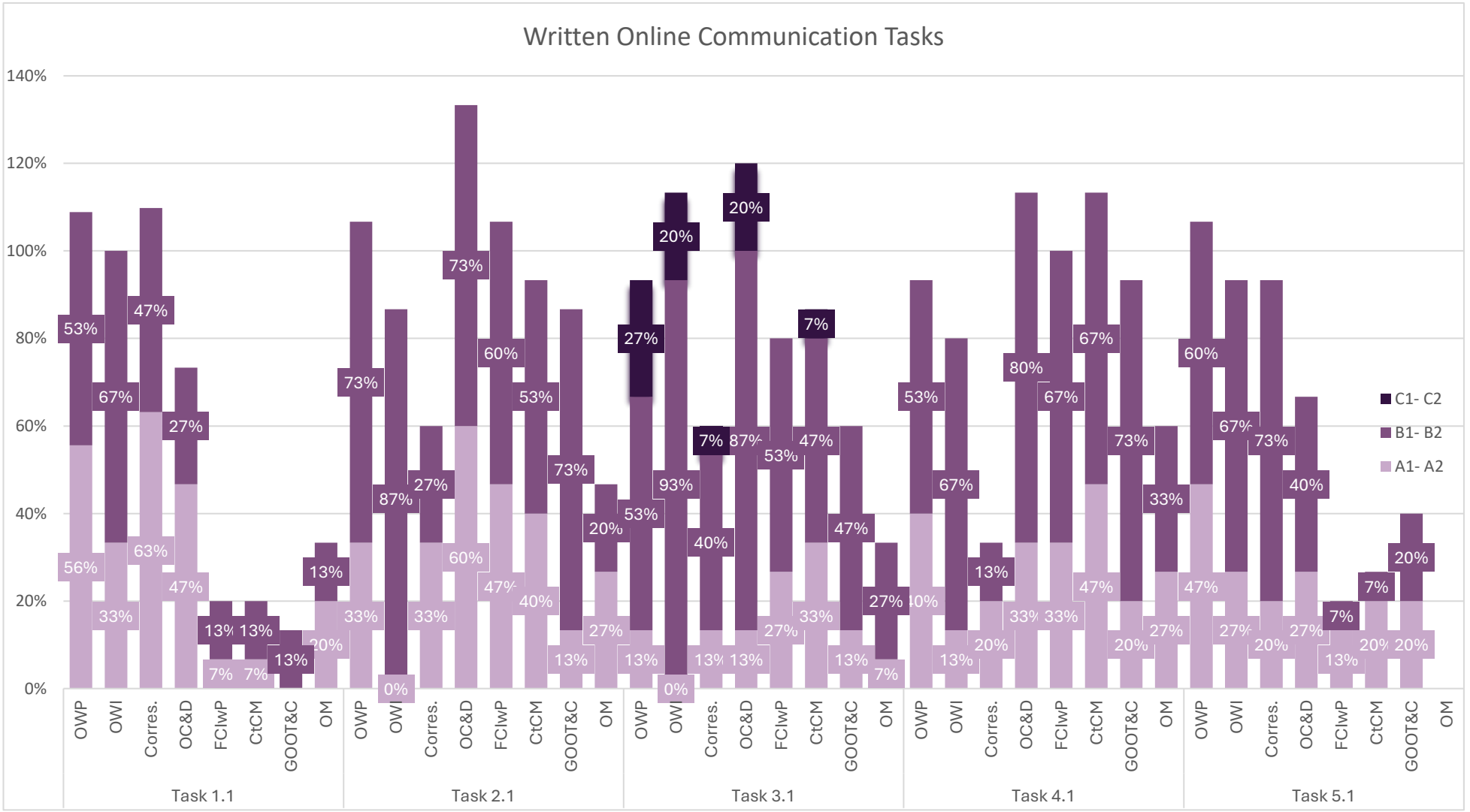


Figure 8.3: CEFR mapping of written online communication tasks

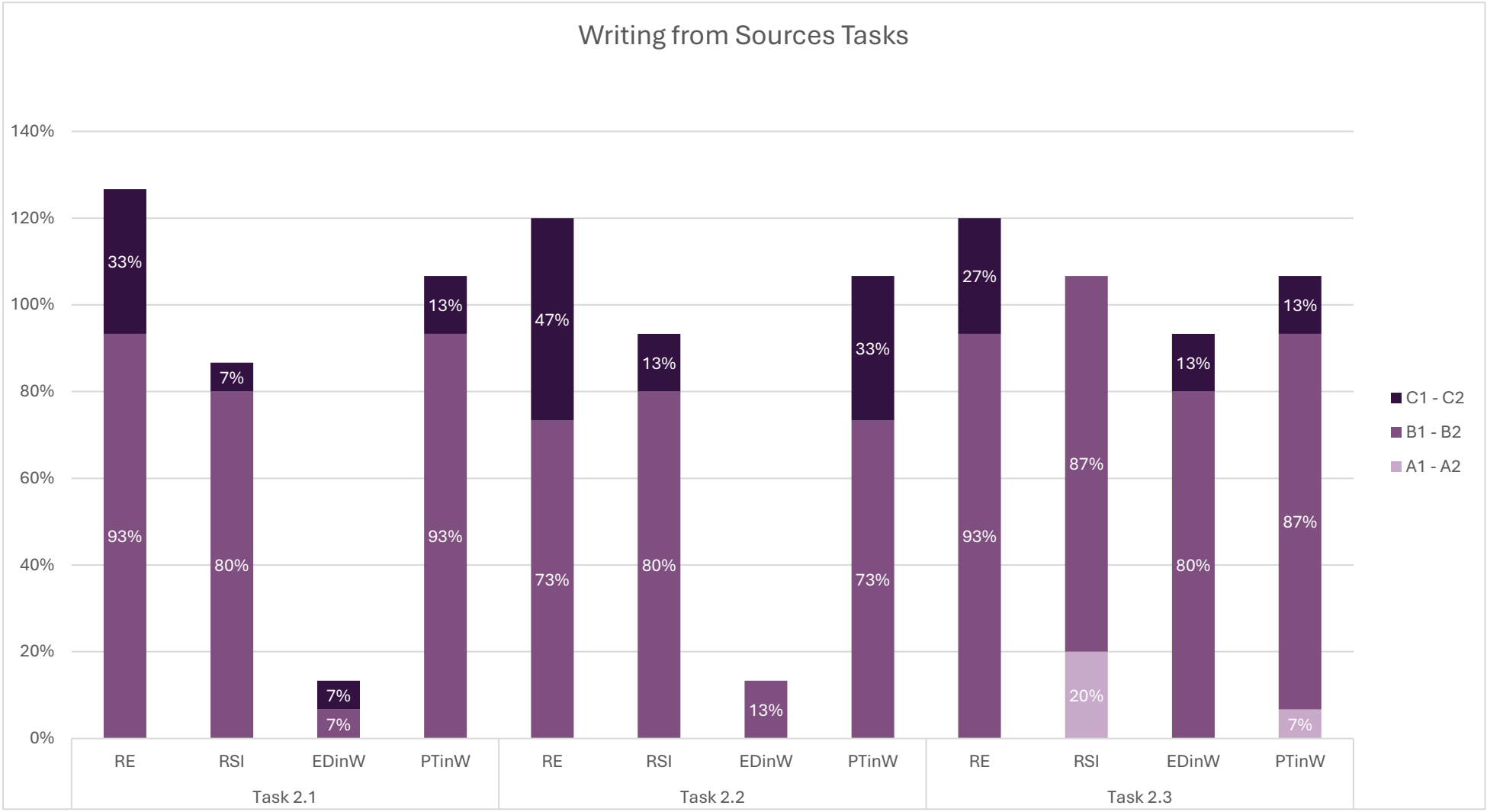


Figure 8.4: CEFR mapping of Writing from Sources tasks

Task 3 was the only task written for the higher proficiency levels C1-C2. Panellists validated Trinity's mapping by reflecting it in their ratings and their comments. For example, J09 rationalised their judgement by saying *"the topic is abstract, there are elements of C-level requirements in it (adaptation of register and style)"*.

Tasks 4 and 5 were designed to assess B1 and B2 CEFR-level KSAs, primarily while ensuring accessibility for A2-level candidates. Once again, panellists validated Trinity's intended CEFR mapping by selecting more B1-B2 CEFR scales to pinpoint the KSAs the candidates needed to have to produce an appropriate response for these tasks. The same panellists also selected CEFR scales for the lower-adjacent CEFR levels to demonstrate that the same tasks were accessible to lower-proficiency candidates. As explained in Section 5, panellists varied in how they approached this task, which introduced unavoidable muddiness to the data that could not be corrected.

8.3.2 Writing from sources tasks

A similar pattern was observed for the *writing from sources* tasks (Figure 8.4). These tasks were designed to target the KSAs illustrated by the B2 to C2 CEFR levels, but they also needed to be accessible to B1 candidates. For this reason, the item writers were instructed to craft the sources accompanying those tasks in language generally accessible to B1-level students. The demands of the tasks themselves, though, should not prohibit C1 and C2 candidates from producing appropriate responses for their ability level. Panellists' ratings and comments confirm Trinity's mapping as both B1-B2 and C1-C2 scales and descriptors were chosen to illustrate the mapping of the tasks. For Task 1, J01 explained:

> *"The topic of this task seemed to point at C1-C2, especially in R&E and also the fact that the rubric insists on a "formal" report which does not appear in B1-B2."*

J07 stated the following about Task 2:

> *"This is a more challenging task than the previous one as the relation of the texts with the task is not so direct and obvious. The information needs to be interpreted in line with the task demands. …A B2 level learner can attempt and successfully complete this task though less fluently than a C1 level learner".*

Overall, this stage allowed panellists to critically evaluate Trinity's mapping (performed during the test design and subsequent content creation process). Panellists' CEFR ratings of the *written online communication* tasks and *writing from sources* tasks aligned with Trinity's intended mapping, thus adding external validity evidence to Trinity's alignment process initiated at the test design and content creation stages.

8.4 Inter- and Intra-Panellist Consistency

This section presents the analyses and the results of two sources of internal validity evidence: 1) inter-panellist and 2) intra-panellist consistency (Cizek & Earnest, 2016; Cizek, Hambleton, Pitoniak, & Copella, 2012; Hambleton & Pitoniak, 2006; Kane, 1994). Following Harsch and Kanistra (2020), panellists were asked to assess 15 candidate written scripts for the six *written online communication* (WOC) tasks analysed in section 8.3 and 13 candidate scripts for the three *writing from sources* tasks (WfS), using the *Written Assessment Criteria Grid* included in the Manual (Council of Europe 2009, p. 188). This approach resulted in collecting 900 CEFR-level judgements per round for WOC scripts (15 scripts × 4 criteria × 15 panellists) and 975 CEFR-level judgements per round for WS scripts (13 scripts × 5 criteria × 15 panellists).

Inter and intra-panellist consistency and reliability were evaluated within the RMT paradigm, which allowed their nuanced evaluation at individual and group levels. A six-facet model was used to analyse the panellists' judgements on the Writing module: 1) candidate written scripts, 2) panellists, 3) task type, 4) panellist sub-groups (internals or externals), 5) round, and 6) criteria. Facets three to five were dummy facets used to facilitate various pairwise interactions; as such, they did not affect the behaviour or measurement of the active facets. Tables 8.4 to 8.8 present the Rasch indices for panellist severity, inter- and intra-panellist consistency, and agreement for each round. The first column indicates the Rasch index related to each measurement context, while columns two to four report the values for each index for each task type and round. When interpreting the data in these tables, it is essential to note that higher

values correspond to higher CEFR levels (e.g., A1 = 1, A1+ = 1.5, A2 = 2, A2+ = 2.5, and so on).

Table 8.4: Summary of panellist severity within RMT (N=15)

Index	WOC task		WS task	
	Round 1	Round 2	Round 1	Round 2
Average measure (*SD*)	-0.06 (1.19)	0.13 (0.91)	-2.17 (1.18)	-2.11 (1.02)
Model *SE*	0.24	0.24	0.27	0.27
Measure min. (Model *SE*)	-2.24 (0.28)	-1.67 (0.27)	-4.70 (0.28)	-4.42 (0.26)
Measure max. (Model *SE*)	2.17 (0.24)	1.54 (0.24)	-0.21 (0.25)	-0.47 (0.26)
Fair average (min)	5.27	5.68	6.48	6.61
Fair average (max)	7.55	7.27	8.32	8.22

Overall, the panellists' mean measure for the WOC task in both rounds (mean measure = -0.06 in R1; mean measure = 0.13 in R2) indicated that the panellists assigned slightly low CEFR judgements when rating candidate written scripts. Such rating behaviour was aligned with the test construct. For the WfS candidate scripts, the panellists' lower mean measure (mean measure = -2.17 in R1; mean measure = -2.11 in R2) implied that their ratings reflected higher CEFR levels. The panellists exhibited high precision (*model SE* = 0.24, WOC; *model SE* = 0.27, WfS) when appraising candidate scripts between rounds and task types. This corroborates the test construct, as the WOC task asks candidates to produce shorter, arguably linguistically simpler responses, whilst the WfS task requires longer, more complex responses. Examining panellist behaviour in more detail, it was observed that the spread measure between the most severe and the most lenient panellist dropped between rounds from 4.41 logits to 3.21 for WOC and from 4.91 to 4.89 for WfS, revealing that the discussion that took place after the round 1 judgements informed panellists' ratings in round 2. The impact of this spread on the judgements of the written scripts was 1.59 raw-score points for WOC and 1.61 for WfS. Such a difference meant that the ratings of the most lenient panellists were only half a CEFR level higher than those of the most severe judge, indicating that the panellists were well aligned in their judgements. The MFRM model eliminated these minor variations in the panellists' ratings, correcting for any idiosyncratic behaviour. This ensured that panellist behaviour did not affect the final script difficulty measures.

Table 8.5 presents a summary of inter-panellist consistency within RMT. As denoted by the high *SP/ROP* values correlation (*SP/ROP* = 0.96, WOC, Round 2; *SP/ROP = 0.93,* WfS, Round 2), panellists exhibited a high inter-panellist consistency. This added evidence of inter-panellist consistency, corroborating the fact that the panellists were interpreting and applying the *Written Assessment Criteria Grid* in a similar fashion. Additionally, the observed *SP/ROP* were very close to the expected *SP/ROP,* thus corroborating that inter-panellist consistency was aligned with the expectations of the Rasch model.

Table 8.5: Summary of inter-panellist consistency within RMT-writing (N=15)

Index	WOC task		WfS task	
	Round 1	Round 2	Round 1	Round 2
Overall *SP/ROP*	0.95	0.96	0.92	0.93
SP/ROP observed-(expected) minimum	0.87 (0.94)	0.93 (0.94)	0.88 (0.88)	0.90 (0.90)
SP/ROP observed-(expected) maximum	0.97 (0.95)	0.97 (0.95)	0.95 (0.92)	0.96 (0.93)
Overall Rasch *kappa*	-0.02	0.02	0.02	0.06
Rasch *kappa minimum*	-0.08	-0.08	-0.08	-0.11
Rasch *kappa maximum*	0.08	0.11	0.09	0.21

The Rasch kappa statistic offers an additional measure of agreement within the Rasch framework. For the WOC task, the Rasch kappa ranged from -0.02 in Round 1 to 0.02 in Round 2. For the WfS task, the Rasch kappa ranged from 0.02 to 0.06. Overall, the panellists demonstrated the appropriate level of agreement both at the individual and group. They appraised candidate scripts in line with the expectations of the Rasch model while maintaining their independence as panellists and experts.

Exact agreement among panellists was measured by the *exact observed % agreement* index (Table 8.6). As expected, the overall *exact observed % agreement* increased after the discussion at the end of Round 1 (36% (34.9%) WOC; 46.8% (43.2%) WfS, albeit close to the expected one (within ± 5%), in line with the model's expectations. However, as shown by the minimum *exact observed % agreement,* at least one panellist had agreement indices lower than those expected by the model. Still, once again, these lower values were not substantially lower than the expected ones (within ±7 %), implying that panellists acted as autonomous experts and exhibited the appropriate level of agreement, thus adding validity evidence to the credibility of their judgements.

Table 8.6: Summary of inter-panellist agreement within RMT- writing (N=15)

Index	WOC task		WS task	
	Round 1	Round 2	Round 1	Round 2
Overall exact observed % agreement (expected %)	29.6% (31%)	36% (34.9%)	41% (39.8%)	46.8% (43.2%)
exact observed % agreement (expected %) minimum	19.9% (20.1%)	26.6% (30.5%)	20.2% (26.2%)	20.1% (27.9%)
exact observed % agreement (expected %) maximum	39.8% (34.5%)	47.5% (37.9%)	47% (42.4%)	58.2% (47.1%)

The detailed panellist measurement reports are available in Appendices C to F. Table 8.6 shows that the mean Infit *Mnsq* values for the panellists remained near the ideal value of 1.00, ranging from 0.84 to 1.06 across tasks and rounds. These outcomes demonstrate that the panellists maintained adequate intra-judge consistency throughout the writing module standard-setting and benchmarking workshop, thereby supporting the internal validity of the resulting cut scores.

In line with Pollitt and Hutchinson (1987), the acceptable Infit range (Infit mean ± 2SD) for the WOC task was 0.39 to 1.43 in Round 1 and 0.34 to 1.34 in Round 2. For the WfS task, the acceptable range was 0.45 to 1.65 in Round 1 and from 0.25 to 1.57 in Round 2. All panellists' infit measures fell within these limits, which are considered appropriate for trained panellists. These findings are consistent with earlier evidence of internal consistency and further reinforce the credibility of the panellists' evaluations.

Table 8.7: Summary of intra-panellist consistency within RMT-writing (N=15)

Index	WOC task		WS task	
	Round 1	Round 2	Round 1	Round 2
Mean Infit *Mnsq;* SD (*Zstd*)(Group)	0.91; 0.26 (-0.40)	0.84; 0.25 (-0.70)	1.05; 0.30 (0.20)	0.91; 0.33 (-0.50)
Minimum Infit *Mnsq* (*Zstd*)	0.53 (-2.10)	0.37 (-3.02)	0.54 (-1.08)	0.37 (-2.07)
Maximum Infit *Mnsq* (*Zstd*)	1.40 (1.30)	1.32 (1.10)	1.63 (2.01)	1.40 (1.30)

In summary, these results indicate that panellist judgements were consistent and reliable. The end of the Round 1 discussion made panellists more consistent in their judgements. This implies that all panellist judgements contributed effectively to the recommendation of a reliable and valid cut score. Therefore, the next set of analyses will focus on the decision consistency, accuracy, and precision of the panellists' recommended cut scores.

8.5 Consistency Within the Method for the Writing Module

As explained in Section 3.3, the consistency within the method for the writing module was evaluated by following the processes and procedures outlined in the Manual (Council of Europe, 2009). The recommended cut scores for the writing module were evaluated for their i) precision and accuracy, and ii) classification consistency and accuracy. As recommended by Kaftandjieva (2010), a dataset of 4,524 candidates was simulated based on the ability measures of the 490 candidates who had participated in test trialling, using *Facets v4.4.4* (Linacre, 2025) to facilitate the in-depth analyses of the cut cores. Table 8.8 shows that the psychometric properties of the real and simulated data were very close.

Table 8.8: Psychometric characteristics of real & simulated candidate population - writing

Index	Real (N = 349)	Simulated (N = 5,013)
Number of tasks	33	33
Candidate mean measure (*SEm ; SD*)	0.62 (0.45; 2.10)	0.57 (0.54; 2.40)
Test reliability	0.94	0.94
RMSE (CSEM)	0.51	0.59
SEM	2.33	2.57
Observed average (*SD*)	3.13 (0.97)	3.22 (1.17)
Fair average (*SD*)	3.41 (0.99)	3.26 (1.16)

For the Writing module, the panellists were not only asked to evaluate the cognitive demands of the writing task but also to classify the candidates' written responses according to CEFR levels and identify those that most accurately represented the targeted CEFR levels. Table 8.9 presents the results of the consistency checks within the method, based on the panellists' CEFR classifications of the candidate scripts, focusing specifically on those scripts they agreed best exemplified performance at levels A1 to C2.

Table 8.9: Evaluating the accuracy and precision of the writing cut scores (N = 5,014)

CEFR level	SE_j	SD_j	SE_j / SD_p	SE_j / *SEM*
A1	0.14	0.51	0.013	0.05
A2	0.09	0.32	0.008	0.03
B1	0.13	0.48	0.012	0.05
B2	0.10	0.39	0.010	0.04
C1	0.11	0.40	0.010	0.04
C2	0.14	0.53	0.013	0.05

The standard deviation of the panellist judgements (SD_j) and the standard error of the mean of their judgements (SE_j) were very small. As a result, the SE_j relative to the population's standard deviation (SE_j/ $SD_p \leq .33$; $SD_p = 12.6$) indicates that classification errors had minimal impact on CEFR level assignment. Importantly, this also suggests that the classifications of the written scripts used to determine the cut scores are robust. This is further supported by the fact that the SE_j of the script classifications was consistently less than one-third of the conditional standard error of measurement (*CSEM*) for each cut score (SE_j/*CSEM* ≤ 0.33), meeting the criterion proposed by Kaftandjieva (2010).

Overall, these findings provide strong evidence of consistency within the method, endorsing the use of the panellists' selected scripts as reliable representations of the CEFR levels for standard setting purposes. These results offer validity evidence for the consistency aspect of the method used in standard setting studies, and therefore, the recommended cut scores can be subjected to further evaluation.

8.6 Decision Consistency and Accuracy

In this section, the decision consistency and accuracy of the recommended cut score are evaluated using two approaches: the Livingston and Lewis (denoted as *LL*) (1995) CTT-based method and the IRT-based method by Lee (2008) using *BB-CLASS* v1.1 (Brennan, 2004) and *IRT-CLASS v2* (Lee & Kolen, 2008), respectively. The recommended cut scores were determined from the candidates' scripts that the panellists identified as best representing the targeted CEFR levels. For the Livingston and Lewis, as well as the Lee method, the raw scores assigned to the candidate responses were used. For the IRT-based method, the individual approach (P) was applied using candidate ability estimates (Lee, 2010).

The Lee method requires item parameters to be included in the program as well; thus, in the context of this study, the seven rating criteria were treated as items, and Samejima's normal ogive graded response model was used to calculate the *DA* (γ) and consistency *DC* (φ) indices at each CEFR level, recommended cut scores. The unidimensionality assumption, an important aspect of this analysis, was met. Candidates' ability measures and scores for the writing module were obtained through an *MFRM* analysis, allowing measurement errors due to rater behaviour to be accounted for.

Table 8.10 presents the results of the evaluation of the recommended cut scores under the Livingston and Lewis, and Lee methods. The evaluation methods are listed in the first column, while the recommended cut scores are provided in the second column, expressed as raw scores. The table reports decision accuracy [*DA*(γ)] and consistency [*DC*(φ)] in columns three and four, respectively, alongside the kappa coefficient in column five. The proportion of correct classifications by chance [*pchance* (φ_C)] is presented in column six, followed by the probability of misclassifications in column seven. The false-positive and false-negative rates are also provided in columns eight and nine.

Table 8.10: Evaluating the DA and DC of calculated cut scores (N = 5,013).

Method	Writing scaled score	*DA* (*γ*)	*DC* (*φ*)	*Kappa* (*κ*)	*pchance* (φ_c)	Probability of misclassification	False positive rate	False negative rate
CEFR Level A1								
LL	5	0.99	0.99	0.64	0.97	0.01	0.006	0.001
Lee	5	0.97	0.96	0.81	0.80	0.04	0.02	0.01
CEFR Level A2								
LL	30	0.95	0.93	0.74	0.73	0.07	0.02	0.03
Lee	30	0.96	0.93	0.83	0.62	0.06	0.02	0.01
CEFR Level B1								
LL	55	0.94	0.91	0.79	0.57	0.05	0.03	0.03
Lee	55	0.95	0.92	0.85	0.50	0.08	0.03	0.03
CEFR Level B2								
LL	80	0.93	0.90	0.80	0.50	0.10	0.04	0.03
Lee	80	0.95	0.93	0.84	0.56	0.07	0.04	0.01
CEFR Level C1								
LL	105	0.95	0.92	0.75	0.70	0.08	0.03	0.02
Lee	105	0.97	0.96	0.82	0.76	0.04	0.03	0.01
CEFR Level C2								
LL	130	0.97	0.97	0.50	0.94	0.03	0.03	0.003
Lee	130	0.98	0.98	0.76	0.90	0.02	0.02	0.01

All *DA* (*γ*) and *DC* (*φ*) measures exceeded the recommended minimum criterion of 0.85 (Subkoviak 1988) for certification examinations at each CEFR level across both CTT and IRT-based methods. This shows that the classification of candidates into various CEFR levels is consistent and precise. Similar to Lee (2010), Deng & Hambleton (2013), and Kanistra (forthcoming), the IRT-based method yielded higher *DA* (*γ*) and *DC* indices, including *φ*, *φc*, and *κ* coefficients. The *κ* values exceeded the expected 0.60 in the CTT paradigm and went beyond 0.76 in the IRT paradigm. For most CEFR cut scores, apart from C2, where the cut score is very close to the maximum weighted raw score of 47, the *κ* values were either greater than or nearly equal to pchance (*φc*). As stated by Subkoviak (1988), pchance (*φc*) increases when cut scores are placed towards the lower or upper ends of the scale, which is expected because the least and most able candidates tend to perform similarly even on non-parallel tests. It is also worth noting that, across all CEFR levels, *κ* values are notably high, indicating that candidate classification largely relies on their performance on the Writing module of the ISE Digital exam.

In summary, the ISE Digital writing module tasks were mapped to the CEFR in three phases: during the conceptualisation stage, during the item creation phase, and through standard setting using the *ID Matching* method. Candidates' written scripts were mapped to the CEFR using the *Benchmarking* approach as described in the Manual (Council of Europe, 2009). Therefore, the ISE Digital writing module is aligned with the CEFR both qualitatively, in terms of content, and quantitatively, through the *Benchmarking* approach reflected by the scores given to the candidates' scripts.

9 Conclusion

This study aligned ISE Digital to the CEFR through three complementary stages:

- During test design, where CEFR-aligned constructs, tasks, and evidentiary models were embedded from the outset;
- During item writing and piloting, where tasks and items were reviewed, refined, and evaluated for CEFR alignment, and
- Through standard setting, using multiple quantitative and qualitative methods to establish defensible cut scores.

Consequently, the qualification is aligned to the CEFR both qualitatively, through content and task design, and quantitatively, through empirically supported standard setting procedures. The standard setting process itself incorporated several innovative features:

- Multiple, context-relevant methods were used (cf. Kaftandjieva, 2010), involving different panels and teams to support triangulation of cut score recommendations.
- For the listening and reading, a *Principled Cut Score* approach (Kanistra, forthcoming) was implemented, using CEFR-linked DIALANG items as anchors to support embedded standard setting.
- For the speaking and writing modules, online standard-setting tools enabled a flexible and rigorous workshop design that supported high-quality decision-making (cf. Kollias, 2023; Kanistra, forthcoming). Additionally, the panellists were asked to evaluate critically the speaking and writing tasks, as well as the candidate performances, thereby adding external validity evidence to Trinity's item development process.

Survey results further confirmed the quality of the standard setting procedures and their execution. Most panellists *strongly agreed* or *agreed* that the standard setting process enabled them to map tasks and performances accurately to the targeted CEFR levels. Panellists also expressed confidence in their ratings and reported that access to other panellists' judgments was helpful when reflecting on their own decisions. The group-recommended CEFR classifications were widely endorsed as appropriate minimum performance standards for the levels under consideration.

Across all four modules, the decision-accuracy (DA) and decision-consistency (DC) indices met accepted benchmarks for high-stakes assessments, indicating that the recommended cut scores can classify candidates accurately and consistently. As part of a responsible validation cycle, and in accordance with the expectations of the UATD framework, these indices will be recalculated once larger sets of operational data become available, particularly for modules where the current study drew on limited candidate datasets.

The final stage of the UATD framework reinforces that CEFR alignment is not a single event but an iterative process embedded in assessment design, item development, and psychometric evaluation. Robust Rasch calibrations and measurement precision, supported by the integration of CEFR-aligned DIALANG items, strengthen the validity of the listening and reading scales. For speaking and writing, MFRM analyses demonstrated that examiners applied severity in a broadly comparable manner, providing dependable support for CEFR-linked classifications.

To maintain the integrity of the scale and the defensibility of cut scores over time, Trinity will undertake several near- and medium-term monitoring actions:

- Repeat the Principled Cut Score Approach when sufficient new operational data become available, ensuring that cut scores remain empirically grounded across administrations.
- Conduct periodic Rasch and MFRM analyses to monitor item functioning, scale stability, and examiner behaviour.
- Review decision-consistency and decision-accuracy indices at regular intervals to ensure that classification decisions remain aligned with CEFR expectations.
- Continue monitoring item and task development through the UATD framework to ensure that content, assessment methods, and external-validation requirements are met as the item bank expands.

Taken together, these processes ensure that ISE Digital remains aligned with CEFR levels in a principled and evidence-based manner, and that Trinity can continue to provide reliable, interpretable, and defensible classification decisions as the test is further developed and operationalised.

.

10 References

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

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. doi:10.1207/S15324818AME1204_2

Council of Europe. (2001). *A Common European Framework for Reference for Language Learning and Teaching.* Cambridge: Cambridge University Press.

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.

Eckes, T. (October 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.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2017). Principled approaches to assessment design, development, and implementation. In A. A. Rupp, & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (1 ed., pp. 41-74). Wiley.

Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgemental task with standard setting panelist expertise: The Item-Descriptor (ID) matching method. *Journal of Applied Testing Technology, 9*(1), 1-20.

Griffiths, M. (2023). *Linking ISE Digital to teh CEFR: A Claim by Specification.* London: Trinity College London.

Jaeger, R. M. (1988). Use and Effect of Caution Indices in Detecting Aberrant Patterns of Standard-Setting Judgments. *Applied Measurement in Education, 1*(1), 17-31. doi:10.1207/s15324818ame0101_3

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

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.

Kanistra, P. (2022). Turning the page on benchmarking and standard setting studies. *Paper presented in S. Ferrara (Organizer), Evolution in Standard Setting: We Can Expect Changes in Practice, a coordinated session in the 2022 annual meeting of the National Council on Measurement in Education.* (pp. 1-12). Virtual meeting: NCME.

Kanistra, P. (2023). *The end of an era?* Retrieved from EALTA : http://www.ealta.eu.org/conference/2023/SIG%20CEFR/CEFR%20SIG%20presentations/SIG%20CEFR/CEFR-%20SIG_Kanistra.pdf

Kanistra, P. (2024 (forthcoming)). *Evaluating the Item Descriptor (ID) Matching method in a face-to-face and synchronous virtual environment.* Berlin: Peter Lang.

Kanistra, P., & Kollias, C. (2024). *Aligning the ICLE 500 written scripts to the CEFR: The technical report.* Institut Langage et Communication. Open Data @ UCLouvain.

Kollias, C. (2023). *Virtual standard setting: Setting cut scores* (Vol. 46). Berlin, Germany: Peter Lang.

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

Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice, 39*(1), 8-21. doi:doi.org/10.1111/emip.12318

Linacre, J. M. (2022). A User's Guide to FACETS. Program Manual 3.84.0. Retrieved from https://www.winsteps.com/manuals.htm

Linacre, J. M. (2022). A user's guide to WINSTEPS MINISTEP Rasch-Model computer programs. Program Manual 5.3.1.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179-197.

MacCann, R. G., & Gordon, S. (2004). Estimating the standard error of the judging in a modified-Angoff standards setting procedure. *Practical Assessment, Research and Evaluation, 9*, Article 5. doi:10.7275/n78q-6g60

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4 ed., pp. 257-305). Westport: American Council on Education/Praeger.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing, & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Myford, C. M., & Wolfe, E. W. (2004a). Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I. In E. V. Jr., & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (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 II. In *Introduction to Rasch Measurement: Theory, models, and applications* (pp. 518-575). 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

North, b., & Jones, N. (2009, January). Further material on maintaining standards across languages, contexts and administrations by exploring teacher judgment and IRT scaling. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR).* Strasbourg, France. Retrieved November 2024, from Council of Europe: https://www.coe.int/en/web/common-european-framework-reference-languages/additional-material#Further

O'Sullivan, B. (2013). *Linking the Aptis reporting scales to the CEFR.* London: British Council.

Pallant, J. (2016). *SPSS survival manual. A step by step guide to data analysis using IBM SPSS* (6 ed.). Berkshire: McGraw Hill Education.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard-setting method for performance assessments: the dominant profile judgement method and some field-test results. *Educational and Psychological Measurement, 57*(3), 400-411.

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.

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.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (sixth ed.). Essex: Pearson.

11 Appendices

APPENDIX A: PANELLIST MEASUREMENT REPORT (SPEAKING)

Round 1

Weightd Score	Weightd Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Kappa*	Nu	Judges
785	118	6.65	6.39	-1.02	.12	1.16	1.2	.98	.0	.90	.96	.97	37.5	32.5	0.07	61	DJ01
852.5	120	7.10	7.03	-1.93	.13	.74	-1.8	.88	-.6	.36	.97	.96	42.8	41.4	0.03	69	DJ09
854.5	120	7.12	7.11	-2.04	.13	.89	-.6	1.24	1.2	.31	.96	.96	41.9	42.1	0.02	64	DJ04
848.5	115.5	7.35	7.20	-2.15	.13	1.06	.4	1.37	1.8	.03	.97	.96	43.9	42.0	-0.01	72	DJ12
791.5	110	7.20	7.31	-2.30	.14	1.17	1.0	1.01	.1	-.18	.96	.96	45.5	42.7	0.05	67	DJ07
892	120	7.43	7.43	-2.46	.13	1.08	.5	.93	-.3	-.23	.96	.96	47.7	43.9	0.04	70	DJ10
881.5	120	7.35	7.45	-2.49	.13	.83	-1.1	.94	-.2	-.12	.96	.96	46.4	44.0	0.03	68	DJ08
902	120	7.52	7.59	-2.69	.13	1.16	1.0	1.17	.8	-.60	.97	.96	49.3	45.7	0.01	71	DJ11
944	120	7.87	8.06	-3.38	.13	1.02	.2	.82	-.6	-2.39	.95	.95	45.2	43.0	0.00	62	DJ02
908.5	116	7.83	8.17	-3.54	.13	1.13	.9	1.09	.4	-3.08	.95	.95	40.5	45.5	-0.01	75	DJ15
946.5	119	7.95	8.22	-3.61	.13	1.16	1.1	1.36	1.2	-2.72	.94	.95	46.4	45.4	0.04	74	DJ14
973.5	120	8.11	8.30	-3.73	.13	.77	-1.7	.75	-.9	-3.69	.95	.95	42.2	41.1	0.02	65	DJ05
990.5	120	8.25	8.41	-3.88	.13	1.21	1.5	1.01	.1	-4.50	.94	.94	42.5	44.0	0.03	73	DJ13
992.5	120	8.27	8.59	-4.12	.13	.68	-2.7	.61	-1.2	-5.60	.95	.94	40.6	38.1	0.02	63	DJ03
991.5	114	8.70	8.89	-4.50	.14	1.28	1.8	1.05	.2	-7.89	.92	.93	35.5	35.5	0.00	66	DJ06
903.6	118.2	7.65	7.74	-2.92	.13	1.02	.1	1.01	.1		.95						Mean (Count: 15)
66.8	2.9	.53	.67	.94	.00	.18	1.4	.20	.9		.01						S.D. (Population)

Model, Populn: RMSE .13 Adj (True) S.D. .93 Separation 7.11 Strata 9.82 Reliability (not inter-rater) .98
Model, Fixed (all same) chi-squared: 792.0 d.f.: 14 significance (probability): .00
Inter-Rater agreement opportunities: 12806 Exact agreements: 5494 = 42.9% Expected: 5274.3 = 41.2%

*Facets do not calculate Rasch Kappa

APPENDIX B: PANELLIST MEASUREMENT REPORT (SPEAKING)

Round 2

Weightd Score	Weightd Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Kappa*	Nu Judges
817	119.5	6.84	6.60	-1.35	.12	1.03	.2	.86	-.9	1.13	.97	.97	42.9	37.4	0.09	61 DJ01
854.5	120	7.12	7.02	-1.92	.13	.80	-1.4	.99	.0	.68	.97	.97	45.0	43.0	0.04	69 DJ09
859	120	7.16	7.13	-2.07	.13	.85	-1.0	1.10	.6	.69	.97	.97	44.9	44.0	0.02	64 DJ04
812	112.5	7.22	7.23	-2.19	.14	1.31	1.8	1.23	1.2	.20	.96	.96	45.6	44.1	0.03	67 DJ07
860.5	115.5	7.45	7.31	-2.31	.14	.79	-1.3	.94	-.3	.50	.97	.96	46.8	43.9	0.05	72 DJ12
881	120	7.34	7.42	-2.45	.13	.72	-1.8	.79	-1.1	.40	.97	.96	50.4	45.7	0.09	68 DJ08
897.5	120	7.48	7.47	-2.53	.13	.90	-.6	.77	-1.2	.22	.97	.96	51.1	45.9	0.10	70 DJ10
929.5	120	7.75	7.88	-3.11	.13	.84	-1.0	.87	-.6	-.77	.97	.96	54.5	47.7	0.13	71 DJ11
937.5	120	7.81	7.96	-3.23	.13	.97	-.1	.89	-.4	-1.55	.96	.96	47.9	44.9	0.05	62 DJ02
893.5	114.5	7.80	8.16	-3.52	.13	1.18	1.2	1.01	.1	-2.55	.96	.95	42.0	46.4	-0.08	75 DJ15
956.5	120	7.97	8.18	-3.56	.13	1.14	1.0	1.82	2.9	-2.43	.94	.95	50.6	47.0	0.07	74 DJ14
966	120	8.05	8.22	-3.61	.13	.86	-1.0	.85	-.6	-2.80	.96	.95	43.4	42.7	0.01	65 DJ05
986	120	8.22	8.33	-3.77	.13	1.15	1.1	.99	.0	-3.48	.94	.95	44.9	45.9	-0.02	73 DJ13
970.5	120	8.09	8.35	-3.80	.13	.75	-2.0	.69	-1.3	-3.61	.95	.95	44.3	41.2	0.05	63 DJ03
976	116.5	8.38	8.65	-4.19	.13	1.16	1.1	.96	.0	-5.87	.93	.94	37.8	37.7	0.00	66 DJ06
906.5	118.6	7.64	7.73	-2.91	.13	.96	-.2	.98	-.1		.96					Mean (Count: 15)
56.4	2.4	.44	.58	.81	.00	.18	1.2	.26	1.1		.01					S.D. (Population)
58.4	2.5	.45	.60	.84	.00	.19	1.3	.27	1.1		.01					S.D. (Sample)

Model, Populn: RMSE .13 Adj (True) S.D. .80 Separation 6.12 Strata 8.49 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-squared: 595.3 d.f.: 14 significance (probability): .00
Inter-Rater agreement opportunities: 12918 Exact agreements: 5915 = 45.8% Expected: 5593.5 = 43.3%

***FACETS DO NOT CALCULATE RASCH KAPPA**

APPENDIX C: PANELLIST MEASUREMENT REPORT (WRITING_WOC TASK)

Round 1

Weightd Score	Weightd Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Kappa*	Nu Judges
135.5	29	4.67	5.21	1.9198	.2315	.83	-.6	.94	-.1	.99	.96	.95	19.9	20.2	0.00	58 CJ13
155.5	30	5.18	5.66	1.3129	.2221	.47	-2.5	.43	-2.5	1.34	.97	.95	24.2	26.2	-0.03	50 CJ05
162	30	5.40	5.85	.9942	.2210	.61	-1.6	.60	-1.6	1.15	.97	.94	28.0	28.9	-0.02	46 CJ01
163.5	29	5.64	5.99	.7255	.2241	.84	-.5	.82	-.6	1.13	.96	.94	32.6	30.8	0.01	52 CJ07
176	30	5.87	6.20	.3109	.2212	1.02	.1	1.03	.1	.91	.95	.94	29.0	32.9	-0.07	59 CJ14
175	29.5	5.93	6.20	.3027	.2226	1.24	.9	1.30	1.1	.75	.95	.94	28.4	32.8	-0.08	51 CJ06
179.5	30	5.98	6.28	.1394	.2216	.78	-.8	.70	-1.1	1.53	.96	.94	39.8	33.4	0.08	49 CJ04
180	30	6.00	6.29	.1148	.2217	.66	-1.4	.70	-1.2	1.19	.97	.94	31.6	33.5	-0.04	60 CJ15
184	30	6.13	6.39	-.0827	.2229	.54	-2.1	.56	-1.9	1.33	.96	.94	36.7	33.7	0.03	53 CJ08
169.5	28	6.05	6.55	-.3686	.2298	1.03	.2	1.20	.7	.74	.88	.92	29.7	32.8	-0.06	47 CJ02
198.5	30	6.62	6.74	-.7000	.2325	.55	-1.9	.57	-1.7	1.18	.96	.94	31.0	32.9	-0.05	54 CJ09
207	30	6.90	6.99	-1.094	.2425	1.22	.8	1.11	.4	.93	.95	.94	28.3	31.3	-0.05	48 CJ03
210	30	7.00	7.17	-1.397	.2499	.85	-.4	.82	-.5	1.17	.95	.94	28.1	29.5	-0.06	55 CJ10
214	30	7.13	7.23	-1.492	.2518	.64	-1.2	.68	-1.0	1.33	.96	.94	29.6	28.9	-0.03	56 CJ11
215.5	29.5	7.31	7.52	-2.028	.2591	.91	-.1	.89	-.2	1.09	.93	.94	26.8	24.9	0.02	57 CJ12
181.7	29.7	6.12	6.42	-.0895	.2316	.81	-.8	.82	-.7		.95			4		Mean (Count: 15)
22.6	.6	.73	.61	1.0693	.0125	.23	1.0	.25	1.0		.02			4		S.D. (Population)

Model, Populn: RMSE .2320 Adj (True) S.D. 1.043 Separation 4.50 Strata 6.33 Reliability (not inter-rater) .95
Model, Fixed (all same) chi-squared: 298.7 d.f.: 14 significance (probability): .00

* Facets do not calculate Rasch Kappa

APPENDIX D: PANELLIST MEASUREMENT REPORT (WRITING_WOC TASK)

Round 2

Weightd Score	Weightd Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Kappa*	Nu Judges
155	30	5.17	5.64	1.3508	.2240	.69	-1.2	.77	-.7	1.09	.96	.95	26.6	29.9	-0.06	50 CJ05
146.5	26	5.63	5.83	1.0352	.2340	.93	-.1	.97	.0	.87	.93	.94	32.2	32.1	-0.01	58 CJ13
163	30	5.43	5.88	.9479	.2251	.75	-.9	.73	-.9	1.06	.96	.95	35.0	33.5	0.01	46 CJ01
163	30	5.43	5.88	.9479	.2251	.96	.0	.87	-.3	1.05	.95	.95	31.0	33.5	-0.06	51 CJ06
164.5	29	5.67	6.02	.6698	.2290	.85	-.4	.82	-.5	1.10	.96	.95	39.3	35.1	0.05	52 CJ07
173	29.5	5.86	6.19	.3320	.2281	.35	-3.1	.37	-2.9	1.58	.97	.94	47.5	36.5	0.15	54 CJ09
178	30	5.93	6.26	.1842	.2253	1.18	.7	1.29	1.0	.85	.95	.94	36.1	36.5	-0.03	59 CJ14
179	30	5.97	6.28	.1334	.2252	.75	-.9	.66	-1.3	1.50	.96	.94	45.0	36.5	0.11	49 CJ04
180.5	30	6.02	6.32	.0574	.2251	.96	.0	.89	-.3	1.07	.93	.94	41.7	36.5	0.06	47 CJ02
183	30	6.10	6.39	-.0692	.2250	.34	-3.4	.38	-3.0	1.40	.97	.94	42.9	36.3	0.08	53 CJ08
184.5	30	6.15	6.43	-.1451	.2251	.62	-1.6	.63	-1.5	1.19	.97	.94	38.1	36.1	0.01	60 CJ15
199	30	6.63	6.80	-.7936	.2327	.63	-1.6	.71	-1.1	1.10	.96	.94	32.4	32.9	-0.02	55 CJ10
202.5	30	6.75	6.85	-.8755	.2346	.91	-.2	.89	-.3	1.03	.96	.94	27.6	32.3	-0.08	48 CJ03
173.5	26.5	6.55	6.94	-1.017	.2507	.70	-1.0	.69	-1.0	1.25	.96	.93	34.1	31.4	0.03	57 CJ12
214	30	7.13	7.25	-1.525	.2514	.67	-1.1	.71	-.9	1.31	.96	.94	29.6	27.0	0.03	56 CJ11
177.3	29.4	6.03	6.33	.0822	.2307	.75	-1.0	.76	-1.0		.96			4	Mean (Count: 15)	
17.5	1.3	.53	.44	.8160	.0087	.22	1.1	.22	1.0		.01			4	S.D. (Population)	

Model, Populn: RMSE .2309 Adj (True) S.D. .7827 Separation 3.39 Strata 4.85 Reliability (not inter-rater) .92
Model, Fixed (all same) chi-squared: 178.2 d.f.: 14 significance (probability): .00

* Facets do not calculate Rasch Kappa

APPENDIX E: PANELLIST MEASUREMENT REPORT (WRITING_WFS TASK)

Round 1

Weightd Score	Weightd Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Kappa*	Nu Judges
198	30	6.60	6.46	-.2046	.2397	1.48	1.7	1.41	1.3	.42	.91	.93	29.6	31.4	-0.04	67 DJ07
179	26.5	6.75	6.61	-.4827	.2517	1.26	.9	1.24	.8	.93	.92	.94	32.4	31.2	0.01	65 DJ05
217	29.5	7.36	6.87	-.9070	.2585	.82	-.5	.86	-.3	1.08	.92	.92	42.2	40.0	0.02	61 DJ01
233.5	32	7.30	7.01	-1.140	.2463	.96	.0	1.05	.2	.91	.92	.92	38.0	40.8	-0.07	73 DJ13
231.5	32	7.23	7.09	-1.269	.2468	.48	-2.2	.46	-2.1	1.46	.94	.92	46.5	41.4	0.06	68 DJ08
233	31	7.52	7.23	-1.490	.2571	.68	-1.1	.62	-1.3	1.29	.95	.91	44.9	42.7	0.01	74 DJ14
226.5	31	7.31	7.44	-1.882	.2495	.96	.0	.87	-.3	1.20	.91	.92	47.0	40.7	0.08	64 DJ04
243	31.5	7.71	7.45	-1.903	.2566	1.10	.4	1.01	.1	1.01	.89	.91	45.7	43.2	0.01	62 DJ02
248	32.5	7.63	7.52	-2.038	.2494	.96	.0	.92	-.1	.96	.89	.91	44.1	42.8	-0.01	63 DJ03
227	30	7.57	7.73	-2.529	.2480	1.36	1.2	1.29	1.0	.86	.91	.92	38.3	37.8	-0.02	72 DJ12
255.5	32.5	7.86	7.75	-2.561	.2465	.73	-1.0	.66	-1.3	1.30	.93	.90	45.6	40.9	0.06	66 DJ06
258	31.5	8.19	7.84	-2.810	.2513	.65	-1.4	.67	-1.3	1.38	.90	.89	43.1	39.7	0.04	69 DJ09
263	32.5	8.09	7.89	-2.952	.2443	1.05	.2	.85	-.5	1.19	.92	.89	44.9	38.2	0.09	70 DJ10
251	31.5	7.97	7.89	-2.952	.2443	.76	-.9	.85	-.5	1.01	.93	.90	40.9	36.7	0.05	75 DJ15
286.5	32	8.95	8.35	-4.199	.2686	.77	-.8	.63	-1.5	1.31	.87	.87	20.2	26.5	-0.08	71 DJ11
236.7	31.1	7.60	7.41	-1.955	.2506	.93	-.2	.89	-.4		.91			5		Mean (Count: 15)
25.6	1.5	.56	.51	1.0377	.0070	.27	1.0	.26	1.0		.02			5		S.D. (Population)
26.5	1.6	.58	.53	1.0741	.0072	.28	1.1	.27	1.0		.02			5		S.D. (Sample)

Model, Populn: RMSE .2507 Adj (True) S.D. 1.007 Separation 4.02 Strata 5.69 Reliability (not inter-rater) .94
Model, Fixed (all same) chi-squared: 252.3 d.f.: 14 significance (probability): .00

* Facets do not calculate Rasch Kappa

APPENDIX F: PANELLIST MEASUREMENT REPORT (WRITING_WFS TASK)

Round 2

Weightd Score	Weightd Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Kappa*	Nu Judges
Group 5																
184.5	27.5	6.71	6.61	-.4768	.2508	1.26	.9	1.18	.6	.95	.92	.94	34.6	34.0	0.01	65 DJ05
222.5	32	6.95	6.70	-.6250	.2468	.89	-.3	.92	-.1	.90	.93	.93	43.0	39.6	0.06	61 DJ01
213.5	30.5	7.00	6.86	-.8871	.2554	1.20	.7	1.18	.6	.83	.91	.93	42.6	41.7	0.02	67 DJ07
233	32	7.28	7.05	-1.193	.2566	1.04	.2	1.06	.3	.95	.93	.92	44.3	44.0	0.01	73 DJ13
233	31	7.52	7.25	-1.535	.2646	.96	.0	.81	-.5	1.15	.95	.92	48.8	45.3	0.06	74 DJ14
240.5	32.5	7.40	7.34	-1.697	.2544	.33	-3.0	.26	-3.4	1.67	.94	.92	58.2	45.2	0.24	68 DJ08
238.5	31.5	7.57	7.36	-1.719	.2614	.46	-2.1	.42	-2.2	1.42	.93	.92	54.3	45.4	0.16	69 DJ09
241.5	31.5	7.67	7.43	-1.855	.2606	.66	-1.2	.63	-1.2	1.20	.92	.91	50.2	45.3	0.09	62 DJ02
226.5	31	7.31	7.48	-1.954	.2523	.94	-.1	.88	-.3	1.20	.92	.93	50.2	42.7	0.13	64 DJ04
247.5	32.5	7.62	7.54	-2.081	.2510	.63	-1.4	.62	-1.4	1.24	.91	.91	50.7	44.5	0.11	63 DJ03
244.5	31.5	7.76	7.58	-2.157	.2576	.44	-2.4	.41	-2.5	1.31	.92	.91	49.9	44.5	0.10	66 DJ06
223	30	7.43	7.66	-2.358	.2494	1.24	.9	1.23	.8	.91	.93	.93	43.3	40.1	0.05	72 DJ12
258	32.5	7.94	7.81	-2.725	.2450	.48	-2.4	.47	-2.5	1.54	.94	.90	48.8	40.6	0.14	70 DJ10
254	31.5	8.06	7.94	-3.113	.2438	.83	-.6	.88	-.4	.92	.93	.90	41.8	35.4	0.10	75 DJ15
279.5	31.5	8.87	8.24	-3.952	.2608	.70	-1.2	.61	-1.8	1.41	.88	.88	20.1	28.1	-0.11	71 DJ11
236.0	31.3	7.54	7.39	-1.888	.2540	.80	-.8	.77	-.9		.92			5		Mean (Count: 15)
20.9	1.2	.50	.44	.8967	.0061	.30	1.3	.30	1.3		.02			5		S.D. (Population)
21.7	1.3	.52	.45	.9282	.0063	.31	1.3	.31	1.3		.02			5		S.D. (Sample)

Model, Populn: RMSE .2541 Adj (True) S.D. .8599 Separation 3.38 Strata 4.85 Reliability (not inter-rater) .92
Model, Sample: RMSE .2541 Adj (True) S.D. .8927 Separation 3.51 Strata 5.02 Reliability (not inter-rater) .93
Model, Random (normal) chi-squared: 13.1 d.f.: 13 significance (probability): .44

*Facets do not calculate Rasch Kappa