TRINITY
COLLEGE LONDON

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

Charalambos Kollias
Paraskevi Kanistra

Trinity Research Report 2015/16-01

About the Authors

Charalambos Kollias is a PhD candidate at Lancaster University, investigating online (virtual) standard setting. He has been involved in language teaching since 1989 and language testing since 1990, during which time he worked for a local English language examination board as an examiner, oral examiner, clerical marker, and essay rater. Ever since, he has held several roles ranging from Oral Examiner/ Essay Rater to Oral Examiner/ Essay Rater Trainer to Chief Coordinator of the Basic Communication Certificate in English (BCCE™). His responsibilities entail managing item writers, training examiners/raters, conducting pre-test and post-test analysis, and conducting standard setting workshops.

Paraskevi (Voula) Kanistra holds an MA in Language Testing from Lancaster University. She has been involved in language teaching since 1991 and language testing since 1998, during which she worked for a local English language examination board as an examiner, oral examiner, clerical marker, and essay rater. Voula has held several roles ranging from Item Writer, Oral Examiner, Essay Rater to Oral Examiner/ Essay Rater Trainer and Team Leader for the Grammar, Reading and Listening item-writer teams for the Basic Communication Certificate in English (BCCE™) and Advanced Communication Certificate in English (ALCE™). Voula has also worked as a freelance language assessment consultant for international examination boards, providing psychometric and construct-validation support.

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 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
trinitycollege.com

Charity number England & Wales | 1014792
Charity number Scotland | SC049143
Patron | HRH The Duke of Kent KG
Chief Executive | Sarah Kemp

Published by Trinity College London
First impression, 2015
Second impression, 2016

Contents

TABLES

FIGURES

Executive Summary

This report documents the 2015 CEFR linking study for Trinity College London's revised Integrated Skills in English (ISE) exam suite, encompassing levels from ISE Foundation to ISE III. The aim of the study was to establish empirically validated cut scores that relate ISE exam results to the Common European Framework of Reference for Languages (CEFR) proficiency levels.

A multi-phase standard setting process was employed, aligned with the Council of Europe's 2009 Manual. This included panellist familiarisation, CEFR benchmarking and standard setting using three methods (Yes/No Angoff, Expected Task Score, and Performance Profile), and extensive validation. The procedures ensured that each test component - Reading, Writing, Listening, and Speaking – was robustly mapped to CEFR performance levels.

The cut scores were derived using pretest data in 2015. However, recognising the limitations of small-sample pretest analyses, decision accuracy was re-evaluated in 2016 using operational data from live administrations. This follow-up phase provided a broader and more representative evidence base for confirming classification accuracy and decision consistency.

Decision consistency and accuracy were estimated using the Livingston and Lewis (1995) method, implemented via the BB–Class software (Brennan, 2001) employing a four-parameter beta-binomial model. Classification accuracy at the key Pass threshold was high across all ISE levels and components, with correct classification rates generally above 0.80 and minimal false-positive and false-negative errors.

Internal validation was confirmed through indicators of intraparticipant and interparticipant consistency, supported by Rasch modelling and classical test theory indices. The findings substantiate the reliability, precision, and defensibility of the ISE cut scores, reinforcing their continued use.

1. Introduction

This report documents a study that links the Integrated Skills in English (ISE) exam to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR, Council of Europe, 2001), which was conducted prior to the launch of a revised version of the exam. The study 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 and benchmarking (training of panellists on the method and how to apply the method in relation to the CEFR levels)
- 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 face-to-face standard setting workshop. Step 5 was conducted after the workshop and evaluated the internal validity of the standard setting procedure.

The standard setting was conducted on pretest data. Three different methods were used. These were identified based on a detailed analysis of the ISE exam specifications, the available data, and an extensive literature review of feasible standard setting methods in the context of aligning exams to the CEFR:

- The *Yes/No Angoff method* (a test-centred method) for the objectively scored parts of the Reading and Listening components
- The *Expected Task Score approach*, a modification of the Angoff method, for polytomous scored items in the Reading and Listening components
- A modification of the *Performance Profile method* for the Writing and Speaking components, which are assessed using a rating scale.

The results of the familiarisation exercise and the cut-score study are reported here.

1.1 OVERVIEW OF THE ISE

The ISE is a multi-skill language examination suite designed for young people seeking proof of their English language proficiency for educational and employment purposes. The suite comprises four levels that target the A2 to C1 levels of the CEFR. At each level, the ISE examinations focus on key competences as outlined in the relevant CEFR descriptors.

ISE level	CEFR level
ISE Foundation	A2
ISE I	B1
ISE II	B2
ISE III	C1

Table 1.1: ISE – CEFR levelling

Each ISE level comprises two independent exam modules: Speaking & Listening and Reading & Writing. The modules can be taken together or at different times when students are ready. Once a candidate has passed both modules at the same level, they receive a certificate for the full qualification.

ISE Speaking & Listening Module Structure

The Speaking & Listening module is a one-to-one, face-to-face, oral interview between one candidate and one examiner. Table 1.2 shows the structure of the module at each ISE level.

	ISE Foundation	ISE I	ISE II	ISE III
CEFR level	A2	B1	B2	C1
Speaking assessment (including interactive listening)	13 minutes	18 minutes	20 minutes	25 minutes
	Topic task	Topic task	Topic task	Topic task
	-	-	Collaborative task	Collaborative task
	Conversation task	Conversation task	Conversation task	Conversation task
Independent listening assessment	Independent listening task	Independent listening task	Independent listening task	Independent listening task

Table 1.2: Tasks in the Speaking & Listening module

As Table 1.2 shows, the Speaking component comprises two or three tasks, each of which is designed to elicit language along different communicative dimensions:

- **Topic task:** Before the exam, the candidate prepares a topic of their own choosing. This serves as a basis for discussion during the exam. The Topic task affords the candidate the opportunity to speak about a subject of personal interest or relevance, one in which they feel confident. This task offers the candidate a degree of autonomy and control.
- **Collaborative task:** In this task, the examiner reads a prompt to the candidate that outlines a dilemma, situation, or opinion. The candidate responds to this prompt by initiating, leading, and maintaining the interaction to learn more about the examiner's background or viewpoint and engaging the examiner in a sustained discussion regarding their circumstances or views. A key element of the collaborative task is that it gives the candidate control over the interaction and encourages them to take the initiative within it.
- **Conversation task:** In this task, the examiner chooses a subject area for discussion with the candidate. A list of subject areas, organised by level, is available in the Examination Specifications. These subject areas have been carefully selected to provide a progression from 'concrete' subjects at ISE Foundation to more 'abstract' topics at ISE III.

The Listening component consists of an independent listening task, during which candidates can demonstrate the listening skills required in lessons and lectures. They listen to a pre-recorded audio track and respond to verbal questions from the examiner, who asks for further details. In a typical lesson or lecture environment, candidates may take notes while listening. The notes are optional and are not assessed.

ISE Reading & Writing Module Structure

At each level of the Reading & Writing module, candidates complete a long reading task, a multi-text reading task, a reading-to-writing task, and an extended writing task. The demands of each task are metered by level and entail the following:

- **Long reading:** The candidate reads a single text (the length varies according to the ISE level), and answers 15 questions based on what they have read. These 15 questions are presented in three groups of five, each testing a different reading skill.
    - Questions 1–5 require the candidate to select the most suitable title for each paragraph of the text. The text comprises five paragraphs, and the candidate must choose from six titles.
    - Questions 6–10 require the candidate to select the five true statements in a list of eight statements. According to the text, five statements are true, while three are false.
    - Questions 11–15 require candidates to complete sentences with a word or phrase taken from the text, using up to three words.

- **Multi-text reading:** The candidate reads several short texts (the length and number of texts vary according to the level), and answers 15 questions based on what they have read. There are three texts at ISE Foundation and four at ISE I, II, and III. One text will always include graphical information. The 15 questions are divided into three groups of five, with each group testing a different reading skill.
    - Questions 16–20 require the candidate to choose the most appropriate sentence to describe each text. There are five sentences, and each refers to one text only. The same text can be the correct answer for up to two questions.
    - Questions 21–25 require the candidate to select the five true statements from a list of eight possible answers. Five statements are true, and three are false or not given.
    - Questions 26–30 require the candidate to complete a summary of the texts with a word or phrase (up to three words) taken from the text. The completed task represents a summary in note form of all the texts in this task. At ISE Foundation, a bank of possible answers is provided for the candidate to choose from.
- **Reading into writing:** In this task, the candidate must write a short response to a prompt using the information provided in the texts from Task 2. This task assesses the candidate's ability to read cross-textually and to transform and adapt what they have read to suit a new purpose. At ISE Foundation and ISE I, the prompt includes three bullet points that guide the candidate in the information to include. In contrast, at ISE II and III, there are no bullet points, and the candidate has more independence in selecting the information to include.
- **Extended writing:** In this task, the candidate responds to a prompt where no input material is provided. The candidate must write independently about the given topic, which is related to one of the communication themes specified for each ISE level. The expected response is in the form of one of the specified genres. The task does not require creative writing skills and does not require the candidate to use their imagination outside of perhaps considering a hypothetical situation within concrete parameters. At ISE Foundation and ISE I, the prompt includes two bullet points to guide the candidate in the information to include and to assist with structuring the answer. There are no bullet points at ISE II and III, and the candidate has more independence in choosing how to respond to the prompt.

Table 1.3 shows the structure of the Reading & Writing module at each ISE level.

	ISE Foundation	ISE I	ISE II	ISE III
CEFR level	A2	B1	B2	C1
Time	2 hours	2 hours	2 hours	2 hours
Task 1	Long reading • 300 words • 15 questions	Long reading • 400 words • 15 questions	Long reading • 500 words • 15 questions	Long reading • 700 words • 15 questions
Task 2	Multi-text reading • 3 texts • 300 words • 15 questions	Multi-text reading • 4 texts • 400 words • 15 questions	Multi-text reading • 4 texts • 500 words • 15 questions	Multi-text reading • 4 texts • 700 words • 15 questions
Task 3	Reading into writing • 70-100 words	Reading into writing • 100-130 words	Reading into writing • 150-180 words	Reading into writing • 200-230 words
Task 4	Extended writing • 70-100 words	Extended writing • 100-130 words	Extended writing • 150-180 words	Extended writing • 200-230 words

Table 1.3: Tasks in the Reading & Writing module

ISE Speaking & Listening Module Assessment

The Speaking component is assessed using a rating scale, which is customised to each ISE level, carefully targeting the CEFR descriptors at those levels. There are four criteria and five levels of performance (0-4) for each criterion:

- **Communicative effectiveness:** this includes task fulfilment, appropriacy of contributions and effectiveness of communicative strategies such as turn-taking and repairing breakdowns in communication.
- **Interactive listening:** this includes the relevance of a response to a question or input, the level of understanding and the speed and accuracy of responses.
- **Language control:** this includes the range and accuracy of the language functions used and the effect on the listener.
- **Delivery:** this includes fluency, intelligibility and the effect on the listener.

The examiner dichotomously scores the Listening tasks at the ISE Foundation and ISE I level during the examination. Candidate performance in the Listening component of the test at Levels II and III is assessed using a five-point category rating scale (0-4). Table 1.4 summarises the scoring method for each ISE level.

Questions	*Task Type*	*Format of response*	*Scoring method*
ISE Foundation			
Questions 1-5	Task 1	Multiple matching (paper)	Dichotomous
Questions 6-12	Task 2	Answer questions orally	Rating scale
ISE Level I			
Questions 1-6	Task 1	Answer questions orally	Dichotomous
Questions 7- 12	Task 2	Answer questions orally	Rating scale
ISE Level I &II			
One question (opening for discussion)	Task 1	Answer the question orally	Rating scale

Table 1.4: Listening component assessment procedure

ISE Reading & Writing Module Assessment

In the Reading component of the ISE examination, some items are assessed dichotomously; a candidate receives a score of '1' for a correct answer or a score of '0' for an incorrect answer. Other items are marked using a partial credit model; a candidate receives a score based on the relative accuracy of their answer. Table 1.5 summarises the scoring method for each reading task.

Questions	*Assessment method*	*Format of response*	*Scoring method*
Task 1 – Long Reading			
Questions 1-5	Multiple matching	selected	dichotomous
Questions 6-10	True/False	selected	Partial credit
Questions 11-15	Fill in the blanks	Open-ended	dichotomous
Task 2 – Multi-text Reading			
Questions 16-20	Multiple matching	selected	dichotomous
Questions 21-25	True/False	selected	Partial credit
Questions 26-30	Fill in the blanks	Open-ended	dichotomous

Table 1.5: Reading component assessment procedure

The Writing component is assessed using rating scales; a different scale is applied to each of the writing tasks. The Reading-into-Writing (Task 3) scale has four criteria with six performance bands per criterion (0-5):

- **Reading into writing**: this includes demonstrating an understanding of source texts, use of paraphrasing and summarising, and identifying common themes across texts.
- **Task fulfilment:** this includes overall achievement of the communicative aim of the task, awareness of the reader, and adequacy of the coverage of the topic.
- **Organisation and structure:** this includes text organisation, presentation of ideas, use of format and signposting.
- **Language control:** this includes range and accuracy of grammar and lexis, and control of spelling and punctuation.

The Extended Writing (Task 4) scale has three criteria with five performance bands per criterion (0-4):

- **Task fulfilment**: this includes overall achievement of the communicative aim of the task, awareness of the reader, and adequacy of coverage of the topic.
- **Organisation and structure**: this includes text organisation, presentation of ideas, use of format and signposting.
- **Language control**: this includes range and accuracy of grammar and lexis, and control of spelling and punctuation.

The writing tasks are equally weighted; candidates receive seven scores, one for each criterion.

1.2 ORGANISATION OF THIS REPORT

There are four remaining sections in this report. The next section describes the methodology used for this linking study and offers an overview of the judgement panel. This is followed by accounts of the familiarisation and cut score setting activities. The last section explores the validity of the results.

2. Standard Setting: Processes & Participants

Standard setting is a decision-making process whereby a panel of experts classifies exam results "in several successive, but limited numbers of levels of achievement (achievement, proficiency, mastery, competency)" (Kaftandjieva, 2010, p.12). These levels of achievement may also be described as performance standards, which typically indicate a minimum level of proficiency or competence and the knowledge a candidate needs to possess in a content area to demonstrate successful mastery of the objectives described in a specific performance category (Cizek, 2012).

Advances in language testing practices, including the development of novel item types (such as polytomously scored items and partial-credit items) alongside the need to establish multiple cut scores (not merely pass/fail distinctions), as well as improvements in statistical programmes, have resulted in a proliferation of standard setting methods used for various educational and/or licensure purposes. Berk (1996) reported 50 methods/approaches, and just over a dozen years later, Kaftandjieva (2010) identified 62 standard-setting methods. The ongoing research and continued refinement of standard setting methods remain a pertinent topic, particularly due to the impact and social consequences of standard setting and cut scores on candidates and society.

In language testing and assessment, the CEFR has had a significant impact on the reporting of language test results in Europe and beyond. Most exams define and/or align their attainment levels with the six proficiency levels of the CEFR. Several alignment and standard-setting procedures are outlined in the Manual (Council of Europe, 2009), which also details the steps necessary to classify exam results into levels of achievement, as delimited by CEFR proficiency levels and their associated descriptors. The Manual recommends five steps for any alignment process:

1. Familiarisation (with the CEFR proficiency level descriptors and the CEFR categories)
2. Specification (of the test tasks, items and content in relation to the CEFR)
3. Standardisation and benchmarking (training of panellists to gain a shared understanding and use of the CEFR levels with regard to how tasks and performances can be related to the CEFR levels)
4. Standard setting (the actual relation of tests or performances to CEFR levels)
5. Validation (of the test, the panellist training, the internal standard setting results, as well as external validation)

Specification (Step 2) was completed as part of the Trinity test development procedure and independently from the standard setting process. The standard setting workshop encompassed Familiarisation, Standardisation and Benchmarking, and Standard Setting (Steps 1, 3, and 4). Validation (Step 5) involved examining the internal validity of the standard setting procedure after the workshop.

2.1 STANDARD SETTING METHODS

There is an abundance of standard setting methods suitable for different test formats, exam conditions, data, and standard setting purposes (Cizek & Bunch, 2007). Standard setting panellists (also referred to as judges) have access to candidates' work, which can range from answers to multiple-choice questions, essays and portfolios to oral performances collected in a speaking assessment context (Zieky, Perie & Livingston, 2012; Council of Europe, 2009). The methods can be grouped into two main categories: test-centred and examinee-centred methods. In the test-centred methods, panellists recommend cut scores by evaluating test items; they make a judgement about how a 'borderline candidate' would perform on a specific item and potentially on any subsequent items sharing similar characteristics with the item in question. In the examinee-centred methods, panellists recommend a cut score after evaluating either the candidates themselves or samples of the written and/or oral language candidates produced, either during their study or during an examination.

This study used two standard-setting methods: the Angoff and Performance Profile methods.

Angoff Method

The Angoff Method is a test-centred approach and is possibly one of the oldest and, thus, most researched standard setting methods, which may also account for its numerous modifications. It has been extensively employed to establish performance standards for dichotomously scored, multiple-choice tests and polytomously scored tasks (Tannenbaum, 2014). In what has become known as the unmodified Angoff Method, panellists are required to evaluate each test item; they identify the skills or subskills a

dichotomous item aims to measure to determine the cognitive processes a candidate must utilise to answer the item correctly. Subsequently, the panellists estimate the likelihood that a borderline candidate would correctly answer this item. In the modified Yes/No Angoff method, panellists define a borderline candidate and assess each dichotomous item to determine whether the borderline candidate (as defined) would be able to answer the item correctly. When the Angoff method is applied in a standard-setting workshop to establish performance standards for polytomously scored tasks such as essays, the panellists evaluate the task, identify the skills and subskills – or aspects of language in the case of speaking and writing tasks – and finally estimate how a borderline candidate would perform on average in each of the tasks. When setting cut scores for polytomously scored tasks or constructed response items, the two primary modifications of the Angoff method are the Mean Item Estimation approach and the Expected Task Score approach. In the Mean Item Estimation Approach, panellists are asked to estimate the mean performance of a borderline candidate on each task. In contrast, in the Expected Task Score approach, the panellists estimate the score a borderline candidate would receive (Plake & Cizek, 2012).

Regardless of the Angoff modification applied, the procedure for calculating the recommended cut score remains similar. The sum of each panellist's judgements is their recommended cut score. The recommended cut scores for the entire panel are then averaged using the mean, the median or the trimmed mean to calculate the group's recommended cut score (Zieky, Perie, & Livingston, 2012; Council of Europe, 2009; Plake & Cizek, 2012).

The Angoff method is the most widely used in educational and occupational testing because it is easy to use. Compared to other standard setting methods, it is easily explained to panellists, and data gathering and analysis are relatively simple. However, the Angoff method is criticised because it is difficult to operationalise the abstract notion of a borderline candidate. Studies have shown that panellists/judges struggle to provide probability estimates. Nevertheless, the Yes/No Angoff method and the Expected Task Score approach (both of which are used in this study) have been well researched and produce reliable cut scores.

Performance Profile Method

The Performance Profile Method is arguably both test-centred and examinee-centred, as it focuses on test scores on one hand and candidate score profiles on the other. Being a relatively new method, it is suitable for performance tests containing a limited number of constructed response items. Ideally, there should be no more than eight items that are scored polytomously or assessed against several assessment criteria, such as through an analytic rating scale. Panellists receive an Ordered Profile Booklet that lists the candidates' profiles ordered by their total scores (from low to high). These profiles present the candidates' analytical scores on the individual items or assessment criteria. Initially, panellists familiarise themselves with the test items/tasks and the meanings of the available polytomous scores, along with the assessment criteria, scoring guides, and descriptors that define these scores and criteria. They are then asked to analyse the candidates' score profiles (they do not see their performances) to select the profile and total score that best represents the characteristics of a borderline performance between two adjacent performance levels (Zieky, Perie, & Livingston, 2012). This method assumes a hypothetical borderline candidate, represented by a total score, which can comprise various score profiles.

In brief, this method aims to establish a cut score for the overall score by reviewing individual profiles to determine whether all candidates achieving the same total score can be deemed proficient enough to be awarded a pass. As previously stated, in this method, the cut score for each judge/panellist is the total score of the profile that best represents a borderline performance between two adjacent performance levels. The group's recommended cut-score is derived by calculating the mean, median, or trimmed mean, depending on the panellists' degree of inter- and intra-consistency (Zieky, Perie, & Livingston, 2012).

The primary advantage of this method is that it enables panellists to critically evaluate the various paths (score profiles) that candidates can take to achieve the total score indicative of a borderline performance. Consequently, the panellists can form an opinion regarding the underlying abilities, knowledge, and skills. Additionally, this method has the potential to save time, as panellists do not assess the actual performances.

Preparing this type of booklet is laborious, even when sophisticated analyses such as item response theory (IRT) scaling are not used. A third issue with this method is that panellists are asked to identify the first total score in the ordered booklet, which can be classified as a 'pass'. This total score can stem

from different criterion profiles, some of which may be classified as a 'pass' and others as a 'fail'. Similarly, a higher total score with a specific criterion profile might be deemed a 'fail', while a lower total score with a different profile could be classified as a 'pass'. In such cases, panellists face the challenging task of judging which of the following types of misclassifications would be less harmful to the candidates and the purpose of the test:

- Passing candidates whose profile is not in line with a successful candidate's profile, although they might have the same total score.
- Failing candidates whose profiles would have allowed them to pass, given their total score.

Finally, the method is not yet well researched, suggesting that cut scores derived by this method would benefit from being validated using another method (Zieky, Perie, & Livingston, 2012).

In this study, panellists created the performance profile first by estimating the score that a borderline candidate would likely be awarded (as in the Expected Task Score approach) and then by explaining the suggested score by giving the combination of scores that would be acceptable as a pass. This addressed the main disadvantage of the Performance Profile method: it can be time-consuming to evaluate all possible profiles that could be arrived at from all possible combinations of scores.

2.2 STANDARD SETTING PANELLISTS

The panel members whose judgements form the basis of the calibration study are central to any standard-setting project. It is widely acknowledged that panellist selection criteria are of utmost importance, and perhaps unsurprisingly, in the plethora of recommendations available, opinions vary 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 & Pitoniak, 2006; and Cizek & Bunch, 2007).

This study drew on the extensive set of guidelines offered by Raymond & Reid (2001, p. 130). 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 examinees 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, as this would expedite the overall ISE benchmarking process.

As might be expected, not all panellists are likely to meet these requirements, particularly subject matter experts representing a diverse constituency of stakeholders, including teachers on ISE preparation programmes, parents of candidates, and educational managers in various markets. To address differences in panellist expertise, Berk (1996, p. 222) suggests identifying two panels: one comprising lay-person stakeholders and the other consisting of subject matter experts. Each panel would contribute to different aspects of the cut-score setting process. The lay-person stakeholders would engage at an initial stage, setting the expectations of various groups regarding the consequences of standard setting. Later in the standard setting process, they would provide 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 fully mitigate the logistical and practical challenges in achieving 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. In accordance with the Manual (2009, p.42) and to ensure that they represented as diverse a group of stakeholders as possible, the panel comprised judges from both within and outside the organisation, reflecting the various stages of language testing development. The group included members from Trinity's examiner panel, examiner trainers, academic consultants, and research staff. As recommended in the Manual, 12 panellists were invited (2009, p. 49). However, due to factors beyond the project's control, one of the panellists had to cancel, resulting in only 11 attending the workshop. Eight panellists were active examiners and/or freelance item writers, while the remaining three were external consultants, six males and five females.

3. Familiarisation

The familiarisation activities aimed to establish the panellists' familiarity with the CEFR levels. For the recommended cut scores to be valid, panellists must be very familiar with the CEFR levels and must rank order CEFR descriptors appropriately. The panellists' performance in the familiarisation activities was analysed using both Classical Analysis and Rasch.

3.1 PANELLIST JUDGEMENTS: CLASSICAL ANALYSIS

Exact agreement and consistency indices were calculated to investigate intra-judge consistency among the panellists' judgements and the CEFR descriptors. Following Kaftandjieva (2010), the misplacement index (MPI), developed by Kaftandjieva (2006), and Goodman and Kruskal's Gamma were computed to examine rater consistency with the correct ordering of the CEFR descriptors. The MPI index ranges from 0 to 1. A value of 0 indicates that the panellist ranks the items in a reverse manner to the correct ordering of the CEFR descriptors. In contrast, a value of 1 is obtained if the ranking of the items aligns completely with the prescribed ordering of the descriptors. Goodman and Kruskal's Gamma was also calculated, as it allows for the standardisation of the difference between each panellist's concordant and discordant pairs and the CEFR descriptors (Norusis, 2006). However, this index is slightly lower than the MPI because it disregards all pairs of cases with tied ranks.

These intra-rater rank correlation indices were used because they provide an overall metric for each rater and insights into their performance on each item. Such detailed output can facilitate an in-depth analysis of each panellist's calibration with the CEFR. Intra-class correlation coefficients (ICC) were also calculated separately for each panellist to investigate the absolute agreement of their average measures with the average measures of the CEFR task descriptors, along with Cronbach's Alpha.

A total of 124 descriptors were used in this familiarisation and calibration exercise: 30 for Speaking, 25 for Writing, 19 for Listening, 20 for Reading, and 30 for Global Descriptors. Tables 3.1 and 3.2 illustrate the intra-rater consistency analysis summary statistics during the familiarisation stage for each panellist and the group.

Index	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11
MPI	.974	.972	.954	.938	.931	.933	.969	.982	.957	.957	.971
Gamma	.948	.944	.908	.876	.863	.866	.937	.964	.915	.914	.943
ICC	.940	.934	.935	.914	.901	.911	.944	.968	.935	.935	.949
Cronbach's Alpha	.933	.940	.939	.922	.904	.911	.948	.969	.940	.944	.951
CEFR Task Mean (3.56)	3.36	3.30	3.36	3.27	3.36	3.64	3.35	3.46	3.33	3.23	3.44

Table 3.1: Intra-panellist consistency indices (individual panellists)

ICC	.985
Cronbach's Alpha	.986

Table 3.2: Intra-panellist consistency and agreement indices (whole panel)

All indices were high, indicating that the panellists ranked the descriptors according to their prescribed order. Indeed, the MPI for all panellists was significantly higher than the minimum of .70 suggested by Kaftandjieva. Furthermore, the more detailed output provided by the MPI programme for each descriptor revealed that both as a group and individually, the panellists generally did not encounter any issues in ranking the CEFR descriptors similarly to their prescribed order. However, a closer examination of the detailed tables, illustrating each panellist's MPI and Gamma index across the 124 descriptors, indicated that the raters were not always successful in accurately ranking all descriptors. Additionally, the mean (average) for all panellists, except for J6, although rather close to the CEFR descriptor mean, tended to be lower than the CEFR descriptor mean, suggesting a tendency towards some strictness in ratings. This tendency was considered during the main standard setting phase, as rater severity and leniency could influence the panellists' recommendations for a cut score for the ISE exam suite.

That said, it is worth noting that during a familiarisation exercise, the CEFR descriptors are broken down into single idea units and presented to panellists as independent items, which makes the ranking task more challenging for them. Consequently, the misplacement of certain descriptors at adjacent levels is unsurprising. It is also important to note that a cut score recommendation is not the result of an individual's efforts but rather the outcome of collective and collaborative group work. The summary statistics (high ICC and high Cronbach's Alpha) presented in Table 3.2 indicate that the panel, viewed as a group, was consistent in interpreting and ranking the CEFR descriptors. Nevertheless, further analyses were undertaken using the Multi-Faceted Rasch Model (MFRM) to investigate individual panellist behaviour and its implications for the ranking of the descriptors in the familiarisation task.

3.2 PANELLIST JUDGEMENTS: MFRM ANALYSIS

The Many-Facet Rasch Model was employed to investigate the sources of panellist and descriptor variability that could affect the outcome of the standard setting workshop. This was performed using FACETS 3.71.4 for Windows (Linacre, 1987-2014). Figure 3.1 shows the summary map produced by FACETS. Note that the map refers to the panellists as judges.

The first column displays the logit scale, an equal-interval scale, which provides a single frame of reference for all the facets of the MFRM analysis, allowing for comparisons both within and between the facets. The second column shows the descriptor facet. The descriptors are spread out over a wide range of logits. The naming convention of the descriptors is as follows:

- The first character shows language skills. For example, 'S' indicates 'Speaking'.
- The next two characters show the descriptor's running order in the descriptor list (eg 01). This is randomly assigned.
- The final two characters show the descriptor's CEFR level. The characters 'C2' indicate a descriptor that belongs to the highest proficiency level described in the CEFR.

Applying this logic, a descriptor labelled 'S01C2' is a C2 level, Speaking descriptor (01 in the running order).

Descriptors at the higher end of the second column in the FACETS map are those that the panellists considered more challenging, while descriptors at the lower end were deemed the easiest. For instance, the panellists evaluated the first speaking descriptor (S01C2) as the most difficult, whereas the writing descriptor (W25A1) was the easiest. This corresponds with the descriptors' assigned CEFR levels, C2 and A1. However, other sections of the map indicate that the progression of the descriptors, as judged by the panellists, does not always align with the progression indicated in the CEFR.

The third column presents the panellist facet (labelled 'judges'). Panellists whose judgements were consistently more stringent (severe) appear at the higher end of the column, while the more lenient ones are at the lower end. J6 is the most lenient panellist in this group, with J10 and J4 being the most severe. However, all panellists cluster within a narrow logit spread (approximately 2.0), which is about 0.08 of the logit spread observed for the CEFR descriptors. This corroborates the findings of the classical analysis, as the panellists had no issues assigning the familiarisation descriptors to their correct CEFR level, as indicated by the very high classical indices (e.g., MPI, Gamma, ICC, etc.).

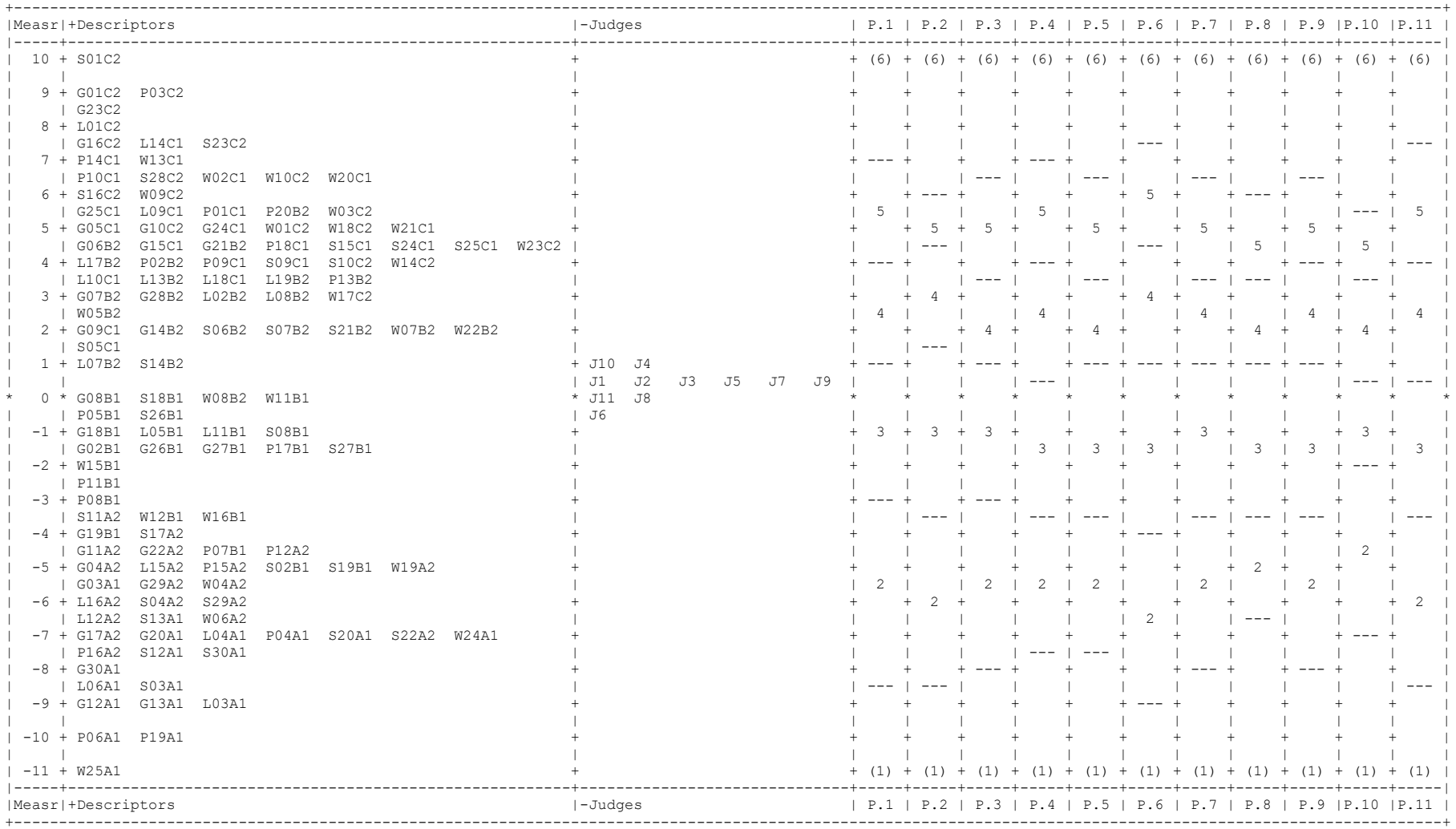

Figure 3.1: FACETS Map of the CEFR Familiarisation Task Descriptors

The final set of columns (P1 to P11) graphically illustrates how each panellist applied the six rating scale categories nested within the CEFR descriptors. Each column corresponds to a different panellist, and although the panellists apply the levels slightly differently (depending on their tendency towards harshness and leniency), they all utilise each of the CEFR levels.

The Descriptor Facet

All 124 descriptors were included in one analysis (see Annex A for all 124 descriptors). Figure 3.1 shows how all the descriptors have been ranked in relation to one another. This section will examine each skill in turn, starting with Speaking.

The Speaking Descriptors measurement report is illustrated in Table 3.3. The first column displays the descriptor ID, the second the pre-assigned CEFR level, the third the difficulty measure of the descriptor in logits based on the panellists' judgements, followed by its standard error (the precision of the measures when the data fit the Rasch model). Table 3.3 indicates that all panellists assigned item S01 a score of six (C2), resulting in a high-difficulty measure. The MPI index (see Table 3.1) also showed that all panellists had ranked this item correctly.

		Model	Infit		Outfit	
Descriptors	CEFR	Measure	S.E.	MnSq	ZStd	MnSq	ZStd
1 S01	C2	(10.52 | 1.85)	Maximum			
23 S23	C2	7.55	.60	1.28	.7	1.12	.4
28 S28	C2	6.63	.52	1.31	.8	1.31	.9
16 S16	C2	6.11	.50	.43	-1.7	.37	-2.0
15 S15	C1	4.65	.49	1.82	1.7	1.92	1.9
24 S24	C1	4.65	.49	1.32	.8	1.29	.7
25 S25	C1	4.65	.49	.38	-1.8	.39	-1.8
9 S09	C1	4.16	.50	1.42	1.0	1.43	1.0
10 S10	**C2**	**3.91**	**.50**	1.49	1.1	1.42	1.0
7 S07	B2	2.03	.53	.47	-1.5	.48	-1.4
21 S21	B2	2.03	.53	.46	-1.5	.46	-1.5
6 S06	B2	1.75	.53	1.23	.6	1.23	.6
5 S05	C1	1.47	.54	.73	-.6	.73	-.6
14 S14	B2	1.17	.55	.97	.0	1.03	.2
18 S18	B1	-.23	.65	.28	-1.7	.25	-1.6
26 S26	B1	-.67	.67	.67	-.4	.58	-.5
8 S08	B1	-1.12	.68	.14	-2.2	.10	-2.4
27 S27	B1	-1.58	.66	.50	-.9	.52	-.8
11 S11	A2	-3.50	.61	.74	-.6	.69	-.7
17 S17	A2	-3.88	.62	.61	-.9	.54	-1.1
2 S02	**B1**	**-5.12**	**.66**	.11	-2.7	.10	-2.6
19 S19	**B1**	**-5.12**	**.66**	.63	-.6	.53	-.8
4 S04	A2	-5.97	.64	1.76	1.5	1.66	1.2
29 S29	A2	-5.97	.64	.53	-1.1	.48	-1.1
13 S13	A1	-6.37	.63	.98	.0	1.03	.2
20 S20	A1	-7.15	.63	1.23	.8	1.25	.7
22 S22	**A2**	**-7.15**	**.63**	.91	-.2	.88	-.2
12 S12	A1	-7.55	.64	.79	-.6	.73	-.6
30 S30	A1	-7.55	.64	1.16	.5	1.24	.7
3 S03	A1	-8.48	.74	.80	-.3	.71	-.4
Mean (Count: 30)		-.54	.63	.87	-.3	.84	-.4
S.D.		5.37	.24	.46	1.2	.47	1.2

Sample: RMSE .68 Adj (True) S.D. 5.33 Separation 7.87 Strata 10.83 Reliability .98
Fixed (all same) chi-square: 2098.0 d.f.: 29 significance (probability): .00

Table 3.3: Speaking Descriptors Measurement Report

If panellists had a perfect agreement, the lower-level items would appear to be the easiest. Disagreement amongst panellists and the CEFR results either in panellists' mis-scaling some of the items, if quite a few of the panellists disagreed substantially with the CEFR, or with high infit or outfit mean square statistics. The final four columns in the table (the Infit and Outfit figures) show the fit statistics, each of which shows the size of variation of the measurement system. The expected values of the Infit and Outfit Mean Square statistics are 1.0. Values less than 1.0 indicate minimal variation, i.e. observations were too predictable. Values greater than 1.0 indicate excess variation, i.e. random behaviour. In the case of both fit statistics, tolerance is allowed. Ideally, the fit statistics for these descriptors should not exceed ±2.75 for Infit [Infit range = .99 ± (.88 × 2) = ± 2.75] and ± 2.65 for Outfit [Outfit range = .97 + (.84 × 2) = ± 2.65]. The Infit and Outfit Zstd statistic shows the significance of the Infit and Outfit Mean Square statistics; the expected value is 0.0. Less than 0.00 indicates too predictable behaviour, while values greater than 0.0 indicate a lack of predictability and, therefore, behaviour that is significant enough to be surprising. Values greater than or equal to 2 (Zstd ≥ ± 2) indicate statistical significance. Consequently, items are not displaying an appropriate fit when mean-square values are outside the acceptable range, and Zstd values are greater than or equal to 2. However, if mean-squares are acceptable (i.e., within infit and outfit ranges), Zstd values can be ignored. Misfit occurs when infit values exceed mean + 2 × S.D. (i.e., misfit = ≥ 2.75 + Zstd ≥ 2.0).

None of the descriptors demonstrated serious misfit, indicating that the variation in panellists' responses was not of concern. Four descriptors (indicated in bold) were not scaled in accordance with their pre-defined CEFR levels, suggesting that panellists could not rank-order S02, S10, S19, and S22 descriptors according to the CEFR. Nevertheless, upon inspecting the raw data, the misclassified items were found to be ranked at adjacent levels.

Table 3.4 shows that the panellists were more consistent in their ranking of the Listening descriptors; only two were misplaced. The infit mean square statistics for all listening descriptors were within the acceptable range, indicating that the variation in the rankings was in line with the model's expectations.

Descriptors	CEFR	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
56 L01	C2	7.94	.66	1.05	.2	1.05	.2
69 L14	C1	7.55	.60	.90	-.1	.89	-.1
64 L09	C1	5.61	.51	.54	-1.1	.60	-.9
72 L17	**B2**	**3.91**	**.50**	.64	-.8	.67	-.7
65 L10	C1	3.65	.51	1.03	.2	1.02	.1
68 L13	B2	3.65	.51	.98	.0	.98	.0
73 L18	**C1**	**3.39**	**.51**	1.44	1.0	1.53	1.1
74 L19	B2	3.39	.51	.74	-.4	.72	-.5
57 L02	B2	2.86	.52	1.18	.5	1.14	.4
63 L08	B2	2.86	.52	.64	-.7	.69	-.6
62 L07	B2	1.17	.55	.88	-.2	.85	-.2
60 L05	B1	-1.12	.68	1.67	1.1	1.64	1.0
66 L11	B1	-1.12	.68	.14	-2.2	.10	-2.4
70 L15	A2	-5.12	.66	1.54	1.0	1.48	.9
71 L16	A2	-5.97	.64	.95	.0	.98	.1
67 L12	A2	-6.37	.63	.65	-.9	.59	-1.0
59 L04	A1	-6.76	.62	1.25	.8	1.31	.8
61 L06	A1	-8.48	.74	.85	-.2	.85	-.1
58 L03	A1	-9.10	.85	.62	-.6	.43	-.6
Mean (Count: 19)		.10	.60	.93	-.1	.92	-.1
S.D.		5.47	.10	.38	.9	.39	.9

RMSE .61 Adj (True) S.D. 5.44 Separation 8.95 Strata 12.26 Reliability .99
Fixed (all same) chi-square: 1287.2 d.f.: 18 significance (probability): .00

Table 3.4: Listening Descriptors Measurement Report

Table 3.5 shows the ranking of the Reading descriptors. Although three descriptors were misplaced (R20, R02, and R16), the infit mean squares for all descriptors were within the acceptable range, indicating that the variation in the rankings of these descriptors was aligned with the model's expectations. An inspection of the raw data showed that the panellists had assigned adjacent ratings for R16. In the case of R20 and R02, a few panellists had assigned either one or two levels above the pre-defined CEFR level for both descriptors.

Descriptors	CEFR	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
77 R03	C2	9.25	1.05	.89	.1	.59	-.1
88 R14	C1	7.21	.56	.68	-.8	.67	-.8
84 R10	C1	6.36	.51	.84	-.3	.83	-.3
75 R01	C1	5.62	.49	.46	-1.5	.55	-1.1
94 **R20**	**B2**	**5.62**	**.49**	1.52	1.2	1.52	1.2
92 R18	C1	4.41	.50	1.44	1.0	1.42	1.0
76 R02	**B2**	**3.91**	**.50**	.96	.0	.96	.0
83 R09	C1	3.91	.50	.41	-1.6	.41	-1.6
87 R13	B2	3.65	.51	1.38	.9	1.42	.9
79 R05	B1	-.67	.67	1.15	.4	1.19	.4
91 R17	B1	-1.58	.66	.38	-1.4	.33	-1.5
85 R11	B1	-2.40	.62	.75	-.6	.70	-.6
82 R08	B1	-3.14	.60	.78	-.6	.72	-.6
81 R07	B1	-4.27	.64	.69	-.5	.62	-.7
86 R12	A2	-4.68	.65	1.70	1.2	1.69	1.2
89 R15	A2	-5.12	.66	.88	.0	.94	.0
78 R04	A1	-6.76	.62	.86	-.3	.85	-.3
90 R16	**A2**	**-7.55**	**.64**	.77	-.7	.70	-.7
80 R06	A1	-10.03	1.12	1.21	.5	.78	.2
93 R19	A1	-10.03	1.12	1.21	.5	.78	.2
Mean (Count: 20)		-.31	.66	.95	-.1	.88	-.2
S.D.		6.03	.20	.37	.9	.38	.9

RMSE .68 Adj (True) S.D. 5.99 Separation 8.75 Strata 12.00 Reliability .99
Model, Fixed (all same) chi-square: 1468.2 d.f.: 19 significance (probability): .00

Table 3.5: Reading Descriptors Measurement Report

Table 3.6 illustrates the panellists' ranking of the Writing CEFR descriptors. It shows that the panellists misplaced descriptors W13, W02, W20, W23, W14 and W17, but an inspection of the raw data revealed that most panellists had ranked the misclassified items at adjacent levels. Interestingly, the misclassification occurs only with C1 and C2 level descriptors, and this could be partly explained by the fact that some of the panellists had not realised that descriptors at both C1 and C2 levels were included in the familiarisation activity and consequently very often ranked some descriptors as C1. Another notable pattern is the infit mean square statistics for descriptors W09, W18, and W19, which are greater than 2.75. This indicates that, for these descriptors, some panellists' rankings were at non-adjacent levels.

Descriptors	CEFR	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
43 W13	**C1**	**6.91**	**.54**	.76	-.5	.79	-.5
32 W02	**C1**	**6.63**	**.52**	.67	-.8	.69	-.8
40 W10	C2	6.63	.52	.55	-1.3	.55	-1.3
50 W20	**C1**	**6.36**	**.51**	.54	-1.3	.53	-1.3
39 W09	C2	5.86	.50	**2.80**	**3.2**	**2.90**	**3.2**
33 W03	C2	5.62	.49	1.95	1.9	2.13	2.2
31 W01	C2	4.90	.49	1.25	.6	1.32	.8
48 W18	C2	4.90	.49	**3.92**	**4.3**	**4.11**	**4.5**
51 W21	C1	4.90	.49	.71	-.6	.69	-.7
53 **W23**	**C2**	**4.41**	**.50**	1.66	1.4	1.67	1.5
44 **W14**	**C2**	**4.16**	**.50**	1.00	.1	1.00	.1
47 **W17**	**C2**	**2.86**	**.52**	.98	.0	1.04	.2
35 W05	B2	2.58	.52	.87	-.1	.87	-.1
52 W22	B2	2.03	.53	.42	-1.7	.43	-1.6
37 W07	B2	1.75	.53	.55	-1.2	.56	-1.2
38 W08	B2	-.23	.65	.98	.1	.93	.0
41 W11	B1	-.23	.65	1.47	.9	1.50	.9
45 W15	B1	-2.00	.64	1.00	.1	.94	.0
42 W12	B1	-3.34	.64	.98	.0	1.01	.1
46 W16	B1	-3.50	.61	1.98	2.3	1.79	1.7
49 W19	A2	-5.12	.66	**8.69**	**6.1**	**7.59**	**5.2**
34 W04	A2	-5.55	.65	1.45	.9	1.54	1.0
36 W06	A2	-6.37	.63	.66	-.8	.59	-1.0
54 W24	A1	-7.15	.63	.85	-.4	.79	-.5
55 W25	A1	(-11.42	1.90)	Minimum			
Mean (Count: 25)		1.02	.61	1.53	.6	1.50	.5
S.D.		5.22	.28	1.72	1.9	1.55	1.8

RMSE .67 Adj (True) S.D. 5.17 Separation 7.74 Strata 10.65 Reliability .98
Fixed (all same) chi-square: 1477.0 d.f.: 24 significance (probability): .00

Table 3.6: Writing Descriptors Measurement Report

In addition to the skill-specific descriptors, the familiarisation step included several global descriptors. Table 3.7 shows that four of these were also misplaced (G10, G06, G09 and G03), but all the infit mean square statistics were well within the acceptable fit criterion. This indicates that, overall, the variation in the ranking of the global descriptors was within the model's expectations.

It should also be noted that, similar to skill-specific descriptors, the global descriptors were presented to the panellists as independent items. Consequently, dependencies and connections between claims remain obscured. Therefore, the misplacement of certain items should not imply that the panellists had difficulty ranking all the global descriptors according to the CEFR. Indeed, the separation figures (7.78, 7.74, 8.95, 8.75, and 8.57 for the Speaking, Writing, Listening, Reading, and Global descriptors, respectively) confirm that the panellists could reliably (reliability ≥ .98) distinguish among the various levels of difficulty of the CEFR descriptors. The significant chi-square for all descriptors (0.00) further corroborates this.

Descriptors		CEFR	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
95	G01	C2	9.25	1.05	.89	.1	.59	-.1
117	G23	C2	8.45	.78	.92	.0	.98	.1
110	G16	C2	7.55	.60	.88	-.1	.85	-.2
119	G25	C1	5.62	.49	.97	.0	1.00	.1
99	G05	C1	5.14	.49	.81	-.3	.82	-.3
104	**G10**	**C2**	**4.90**	**.49**	.35	-1.9	.37	-1.9
118	G24	C1	4.90	.49	.52	-1.3	.49	-1.4
100	**G06**	**B2**	**4.41**	**.50**	.47	-1.4	.48	-1.4
109	G15	C1	4.41	.50	.48	-1.4	.48	-1.4
115	G21	B2	4.41	.50	.57	-1.1	.56	-1.1
122	G28	B2	3.13	.52	.19	-2.6	.19	-2.6
101	G07	B2	2.86	.52	.57	-.9	.65	-.7
108	G14	B2	2.03	.53	.90	-.1	.82	-.3
103	**G09**	**C1**	**1.75**	**.53**	1.16	.5	1.18	.5
102	G08	B1	.17	.62	.52	-1.0	.51	-.9
112	G18	B1	-1.12	.68	.14	-2.2	.10	-2.4
96	G02	B1	-1.58	.66	.85	-.1	.77	-.2
120	G26	B1	-1.58	.66	1.51	1.0	1.47	.9
121	G27	B1	-1.58	.66	.50	-.9	.52	-.8
113	G19	B1	-3.88	.62	1.32	.8	1.31	.7
116	G22	A2	-4.27	.64	.45	-1.3	.36	-1.5
105	G11	A2	-4.68	.65	.17	-2.4	.15	-2.4
98	G04	A2	-5.12	.66	.59	-.7	.56	-.7
97	**G03**	**A1**	**-5.55**	**.65**	.75	-.3	.65	-.5
123	G29	A2	-5.55	.65	.39	-1.4	.29	-1.6
111	G17	A2	-6.76	.62	.83	-.4	.81	-.4
114	G20	A1	-7.15	.63	.75	-.7	.70	-.8
124	G30	A1	-7.99	.68	.77	-.6	.68	-.7
106	G12	A1	-9.10	.85	.82	-.1	.83	.0
107	G13	A1	-9.10	.85	1.73	1.2	3.38	2.1
Mean (Count: 30)			-.20	.63	.73	-.7	.75	-.7
S.D. (Sample)			5.51	.13	.37	1.0	.59	1.1

RMSE .64 Adj (True) S.D. 5.47 Separation 8.57 Strata 11.76 Reliability .99
Fixed (all same) chi-square: 2059.2 d.f.: 29 significance (probability): .00

Table 3.7: Global Descriptors Measurement Report

The Rater Facet

The analysis of the Descriptor facet has revealed that some panellists found it challenging to rank certain descriptors according to the CEFR. Consequently, it is also prudent to investigate the rater facet for inconsistencies in panellist behaviour. Table 3.8 displays the panellists listed in order of the severity they applied while ranking the descriptors. J10 was the strictest panellist; they assigned some higher-level descriptors to lower CEFR levels, whereas J6 was the most lenient, assigning some lower-level descriptors to higher CEFR levels.

The acceptable infit range for the panellists was .78 to 1.77. J5 exceeded the maximum criterion for serious misfit (Infit Mean Square = 2.00), indicating that some of his ratings were very inconsistent. An inspection of the raw data also indicated that, on two occasions, J5 ranked the descriptors in exactly the reverse order from the CEFR. This has implications for the cut-score setting process and will be discussed in Section 3.3. Nevertheless, the mean infit for the group is close to the desirable 1.00, indicating that, as a group, the panellists exhibited the appropriate amount of variation (Myford & Wolfe, 2004, p.495; Linacre, 2010).

Column 11 of Table 3.8 presents the Corr. PtBis statistic, which reflects the correlation between a single rater and the rest of the raters (SR/ROR). According to Myford & Wolfe (2004, p. 498), "SR/ROR correlations less than .30 are considered to be rather low, while correlations greater than .70 are considered to be high for a rating scale composed of several categories. However, as the number of rating scale categories decreases, these rule-of-thumb values should be relaxed". On average, the panellists demonstrated a high level of agreement (average inter-rater correlation of .66), indicating that, overall, the panellists ranked descriptors in a similar manner.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Judge
400	124	3.23	2.90	1.11	.17	.98	-.1	1.10	.6	.66	48.4	50.5	10	J10
405	124	3.27	3.14	.89	.18	1.25	1.8	1.26	1.6	.64	49.9	54.0	4	J4
409	124	3.30	3.04	.66	.17	.65	-2.8	.61	-2.8	.66	55.3	53.5	2	J2
415	124	3.35	3.09	.61	.17	.87	-.9	.78	-1.6	.66	56.5	54.6	7	J7
413	124	3.33	3.12	.60	.17	.84	-1.2	.80	-1.4	.67	56.1	54.4	9	J9
417	124	3.36	3.08	.59	.17	.82	-1.4	.82	-1.2	.67	55.2	54.4	3	J3
417	123	3.39	3.18	.47	.17	2.00	5.9	1.92	4.8	.61	54.0	54.3	5	J5
417	123	3.39	3.15	.38	.18	.64	-3.0	.55	-3.5	.67	59.4	54.8	1	J1
429	124	3.46	3.20	.24	.17	.69	-2.5	.63	-2.3	.68	56.2	52.5	8	J8
426	124	3.44	3.31	.22	.18	1.13	1.0	1.21	1.4	.66	50.2	54.3	11	J11
451	124	3.64	3.36	-.66	.18	1.04	.3	.97	-.1	.65	49.2	50.5	6	J6
418.1	123.8	3.38	3.14	.46	.17	.99	-.3	.97	-.4	.66			Mean(Count: 11)	
13.7	.4	.11	.12	.46	.00	.39	2.6	.40	2.5	.02			S.D. (Sample)	

RMSE .17 Adj (True) S.D. .42 Separation 2.42 Strata 3.55 Reliability (not inter-rater) .85
Fixed (all same) chi-square: 66.1 d.f.: 10 significance (probability): .00
Inter-Rater agreement opportunities: 6800 Exact agreements: 3650 = 53.7% Expected: 3632.1 = 53.4%

Table 3.8: Rater Facet Report

Other important statistics to inspect include the following:

- **Strata:** As a group, the panellists exercised almost three severity levels (Strata 3.55).
- **Reliability:** Ideally, low-reliability indices are preferable for the rater facet (Linacre, 2009; Myford & Wolfe, 2004). The panellists demonstrated that they ranked the CEFR descriptors reliably (.85).
- **Chi-square:** This was 66.1 (df = 10), p = .00, which rejected the null hypothesis that there was no significant difference among the panellists in the severity/leniency levels they exercised.
- **Inter-rater agreement:** The dataset contained 6800 opportunities for inter-rater agreement. With the spread of rater severities exhibited by the panellists, the Model expected panellists to achieve exact agreement on 53.4% (expected agreement) of the observations. The exact agreement observed was very close to the expected one (53.7%), indicating that the panellists as a group were consistent in their ranking of the descriptors.

Finally, the Rasch Kappa for this group of panellists was .006. The Rasch version of the kappa index and Cohen's (1960) kappa are conceptually similar. According to Eckes (2011, p.71), "under Rasch model conditions, the Rasch Kappa index should be close to 0.0. Values much larger than 0 indicate overly high interrater agreement and, consequently, a high degree of local rater dependence; large negative values indicate much less interrater agreement than expected based on the Rasch model, which may be due to unmodeled sources of variation in the ratings (e.g., hidden facets)". The Rasch Kappa for the rater facet is very close to the desired 0, supporting the conclusion that the panellists in the current session exhibited the desirable amount of interrater agreement.

3.3 FAMILIARISATION OUTCOME

Reviewing the outcomes of the Classical and Rasch analyses for the descriptor facet and rater facet analyses, the following conclusions are defensible:

- Descriptor Facet Analyses
  - The overall ranking of the majority of the descriptors was in accordance with the CEFR, indicating that panellists overall understood how language ability advances from one CEFR level to the next.
  - The panellists could not place some of the descriptors at the appropriate level, but such behaviour is not surprising given the abstract nature of the task. The familiarization exercise was identical to the one Papageorgiou (2007) used to link a previous version of the ISE to the CEFR. The panellists for that project also misplaced some of the descriptors. Interestingly, eight of the 19 descriptors that were misplaced in this study were also misplaced by the panellists in the 2007 linking study.
  - Panellists struggled particularly with higher-level descriptors; they had problems distinguishing between B2 and C1 and between C1 and C2. It is possible that the underlying methodology of the familiarisation stage, which deconstructs the descriptor paragraphs at each CEFR level into single statements, coupled with the inconsistencies characterising CEFR descriptors at higher levels, may partly explain the panellists' difficulties with the task (Papageorgiou 2009, pp. 106-115).
- Rater Facet Analyses: Though the results from the rater facet analyses were satisfactory and indicated sufficient rater agreement to proceed with subsequent linking stages, there were specific areas where panellists could not understand and distinguish between adjacent CEFR levels. While placing a C1 descriptor at the C2 level may not be surprising, systematic misplacements might jeopardise the validity of the linking outcomes. To avoid this, the following measures were taken:
  - The panellists reviewed the descriptive statistics to better understand their disagreements regarding the CEFR level of the descriptors.
  - "Problematic" descriptor statements were discussed.
  - Panellists received a copy of all the CEFR scales to use during the Standard Setting workshop.

  The panellists indicated that the discussion was beneficial, as it helped them clarify the differences between adjacent levels and examine the descriptors more carefully. Although there is no empirical evidence to confirm the positive impact of the discussion (the descriptor sorting task was not repeated), one would expect that the steps taken to explain the familiarisation results and provide feedback to the panellists guided them in the right direction.

4. Setting Cut Scores

This section presents the cut-score judgements for each ISE level and language skill. The cut score judgements were established in two rounds. In each round, using the standard setting methods described in Section 2, the panellists arrived at a cut score for each performance level. After the first judgement round, the panellists received feedback on their judgements and were offered opportunities to discuss areas of incongruence. Judgement round 2 was conducted immediately following the post-round 1 discussion. The round 2 judgements were the panellists' recommended cut scores.

4.1 SPEAKING CUT SCORES

The cut scores for the Speaking component were set using the Performance Profile method (described in Section 2.1). This section presents a table of cut-score judgements for each ISE level – four tables in total. For each table, the first column identifies the panellist's ID number. Since each ISE level has three results bands - Pass, Merit, and Distinction, where the Pass cut-score delimits the minimally competent candidate at that ISE level, the subsequent columns show each panellist's suggested cut score for each judgement round by results band. The bottom section of each table shows the group's descriptive summary statistics. For this study, the final recommended cut score is the mean (average) of the panellist judgements.

Table 4.1 shows that the recommended cut scores (after judgement round 2) for ISE Foundation were: Pass = 7.76, Merit = 12.45, and Distinction = 15.09. In the Pass and Distinction results bands, the cut scores decreased slightly between judgement rounds for the Pass and Distinction groups but remained the same for the Merit group.

	Borderline		**Merit**		**Distinction**	
Judge ID	**Round 1**	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
J01	7.50	7.50	11.75	11.75	14.50	14.50
J02	7.60	7.60	12.00	12.00	14.70	14.70
J03	7.80	7.80	13.10	13.10	15.60	15.60
J04	7.80	7.80	11.40	11.40	15.40	15.40
J05	8.00	8.00	13.65	13.65	16.00	16.00
J06	6.75	6.75	11.70	11.70	15.50	15.50
J07	8.20	7.90	13.20	13.20	15.00	15.00
J08	8.00	8.00	11.00	11.00	14.00	14.00
J09	10.30	8.00	13.80	13.80	15.80	15.80
J10	8.00	8.00	13.80	13.80	16.00	15.50
J11	8.00	8.00	11.60	11.60	14.00	14.00
Min	6.75	6.75	11.00	11.00	14.00	14.00
Max	10.30	8.00	13.80	13.80	16.00	16.00
Mean	8.00	7.76	12.45	12.45	15.14	15.09
SD	.82	.36	1.01	1.01	.71	.67

Table 4.1: Cut score judgements for the Speaking component: ISE Foundation

Table 4.2 shows the recommended cut scores for ISE I (after judgement round 2). They were Pass = 8.75, Merit = 12.67, and Distinction = 15.27. The recommended cut score for all three results bands increased slightly between the judgement rounds.

	Borderline		Merit		Distinction	
Judge ID	**Round 1**	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
J01	7.80	7.80	11.80	11.80	15.00	15.00
J02	7.50	8.00	11.50	11.50	15.50	15.50
J03	8.00	8.00	12.40	12.80	15.20	15.60
J04	7.80	7.80	11.40	12.20	15.40	15.40
J05	8.00	8.40	13.65	13.65	16.00	16.00
J06	7.80	7.80	11.30	11.30	14.90	14.90
J07	8.20	9.00	12.30	12.60	14.40	15.00
J08	11.00	11.00	14.00	14.00	16.00	16.00
J09	10.40	10.50	13.50	13.50	15.80	15.80
J10	7.60	7.90	11.20	12.80	14.60	14.60
J11	9.80	10.00	13.20	13.20	14.40	14.20
Min	7.50	7.80	11.20	11.30	14.40	14.20
Max	11.00	11.00	14.00	14.00	16.00	16.00
Mean	8.54	8.75	12.39	12.67	15.20	15.27
SD	1.18	1.15	.99	.85	.57	.56

Table 4.2: Cut score judgements for the Speaking component: ISE I

Table 4.3 shows the recommended cut scores for ISE II (after judgement round 2). They were Pass = 8.81, Merit = 12.28, and Distinction = 15.01. For the Pass and Merit results bands, the recommended cut score increased slightly between the judgement rounds, but for the Distinction results band, the recommended cut score decreased by .06 of a raw score point.

	Borderline		Merit		Distinction	
Judge ID	**Round 1**	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
J01	7.80	8.00	11.60	12.10	15.50	15.25
J02	7.80	7.80	11.40	11.40	15.40	15.40
J03	9.20	9.20	13.20	13.20	15.60	15.60
J04	7.80	7.80	11.60	11.60	15.60	15.40
J05	8.60	8.60	12.80	12.80	15.20	15.20
J06	7.30	7.30	9.90	10.90	14.10	14.10
J07	8.60	9.20	13.50	13.10	14.90	14.70
J08	12.00	12.00	14.00	14.00	16.00	16.00
J09	9.90	9.40	12.00	12.00	14.20	14.20
J10	8.60	8.60	12.80	12.80	15.40	15.40
J11	8.20	9.00	11.20	11.20	13.90	13.90
Min	7.30	7.30	9.90	10.90	13.90	13.90
Max	12.00	12.00	14.00	14.00	16.00	16.00
Mean	8.71	8.81	12.18	12.28	15.07	15.01
SD	1.25	1.20	1.14	.93	.67	.65

Table 4.3: Cut score judgements for the Speaking component: ISE II

Table 4.4 shows the recommended cut scores for ISE III (after judgement round 2). They were Pass = 8.69, Merit = 12.17, and Distinction = 14.91. For the Pass and Merit results bands, the recommended cut score increased slightly between the judgement rounds, but for the Distinction results band, the recommended cut score decreased by .03 of a raw score point.

	Borderline		Merit		Distinction	
Judge ID	**Round 1**	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
J01	8.30	8.30	12.20	12.20	15.60	15.60
J02	8.00	8.00	11.80	11.80	15.20	15.20
J03	9.50	9.50	13.40	13.40	15.60	15.60
J04	8.00	8.00	11.60	11.80	14.80	14.80
J05	9.20	9.20	12.70	12.70	14.70	14.70
J06	6.90	6.90	10.90	10.90	14.70	14.70
J07	9.50	9.70	13.10	13.50	15.10	15.10
J08	11.00	11.00	15.00	15.00	16.00	16.00
J09	8.00	8.00	10.20	10.20	14.00	14.00
J10	7.70	8.20	10.80	10.80	14.80	14.50
J11	8.80	8.80	11.60	11.60	13.80	13.80
Min	6.90	6.90	10.20	10.20	13.80	13.80
Max	11.00	11.00	15.00	15.00	16.00	16.00
Mean	8.63	8.69	12.12	12.17	14.94	14.91
SD	1.07	1.06	1.31	1.33	.63	.65

Table 4.4: Cut score judgements for the Speaking component: ISE III

4.2 LISTENING CUT SCORES

The cut scores for the Listening component were set using the Yes/No Angoff method (described in Section 2.1). This section presents a table of cut-score judgements for each ISE level – four tables in total. For each table, the first column identifies the panellist's ID number. Since each ISE level has three results bands - Pass, Merit, and Distinction, where the Pass cut-score delimits the minimally competent candidate at that ISE level, the subsequent columns show each panellist's suggested cut score for each judgement round by results band. The bottom section of each table shows the group's descriptive summary statistics. For this study, the final recommended cut score is the mean (average) of the panellist judgements.

Table 4.5 shows that the recommended cut scores (after judgement round 2) for ISE Foundation were: Pass = 4.00, Merit = 7.46, and Distinction = 8.74. The recommended cut score for all three results bands increased slightly between the judgement rounds.

Table 4.6 shows the recommended cut scores for ISE I (after judgement round 2). They were Pass = 4.71, Merit = 8.17, and Distinction = 9.35. The recommended cut score for all three results bands decreased slightly between the judgement rounds.

	Borderline		Merit		Distinction	
Judge ID	**Round 1**	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
J01	5.00	4.00	8.00	8.00	9.00	9.00
J02	2.00	3.00	5.20	8.20	8.80	8.80
J03	2.50	2.50	6.50	6.50	8.40	8.40
J04	4.00	3.80	6.80	7.80	8.80	8.80
J05	4.00	5.00	8.50	8.50	9.00	9.00
J06	2.70	2.70	5.70	5.70	8.50	8.50
J07	2.80	3.80	6.80	6.80	8.50	8.50
J08	6.00	5.00	8.00	8.00	9.00	9.00
J09	4.00	5.20	7.80	7.80	8.70	8.70
J10	2.00	3.80	7.80	7.80	8.80	8.80
J11	5.30	5.10	8.00	7.00	8.50	8.60
Min	2.00	2.50	5.20	5.70	8.40	8.40
Max	6.00	5.20	8.50	8.50	9.00	9.00
Mean	3.70	4.00	7.19	7.46	8.70	8.74
SD	1.30	.90	1.02	.81	.20	.21

Table 4.5: Cut score judgements for the Listening component: ISE Foundation

	Borderline		Merit		Distinction	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	5.00	6.00	8.90	8.90	9.90	9.70
J02	5.00	4.00	8.80	8.80	9.80	9.80
J03	5.80	3.80	9.00	9.00	9.80	9.80
J04	5.80	4.20	7.60	7.40	9.70	8.50
J05	5.00	4.00	9.50	9.00	10.00	10.00
J06	6.70	5.70	8.50	8.70	9.70	9.50
J07	4.80	4.80	7.90	7.90	9.50	9.50
J08	6.00	4.00	8.00	8.00	9.00	9.00
J09	5.00	6.20	6.80	5.60	8.50	8.50
J10	6.00	4.00	9.50	9.50	10.00	10.00
J11	5.30	5.10	8.10	7.10	9.60	8.60
Min	4.80	3.80	6.80	5.60	8.50	8.50
Max	6.70	6.20	9.50	9.50	10.00	10.00
Mean	5.49	4.71	8.42	8.17	9.59	9.35
SD	.58	.86	.79	1.08	.44	.57

Table 4.6: Cut score judgements for the Listening component: ISE I

Table 4.7 shows the recommended cut scores for ISE II (after judgement round 2). They were Pass = 2.22, Merit = 3.00, and Distinction = 3.69. For the Pass and Merit results bands, the recommended cut score increased slightly between the judgement rounds, but for the Distinction results band, the recommended cut score decreased by .01 of a raw score point.

	Borderline		**Merit**		**Distinction**	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	2.00	2.00	3.00	3.00	3.90	3.90
J02	2.00	2.20	2.90	3.00	3.50	3.50
J03	2.20	2.20	3.20	3.20	3.80	3.80
J04	2.00	2.00	2.70	2.80	3.60	3.60
J05	2.20	2.20	3.00	3.00	3.80	3.80
J06	2.00	2.00	2.80	2.80	3.50	3.50
J07	2.20	2.20	3.20	3.20	3.80	3.70
J08	3.00	3.00	3.00	3.00	4.00	4.00
J09	2.50	2.50	3.00	3.00	3.50	3.50
J10	2.00	2.00	3.20	3.20	3.80	3.80
J11	2.10	2.10	2.80	2.80	3.50	3.50
Min	2.00	2.00	2.70	2.80	3.50	3.50
Max	3.00	3.00	3.20	3.20	4.00	4.00
Mean	2.20	2.22	2.98	3.00	3.70	3.69
SD	.29	.29	.16	.15	.18	.17

Table 4.7: Cut score judgements for the Listening component: ISE II

Table 4.8 shows the recommended cut scores for ISE III (after judgement round 2). They were Pass = 2.34, Merit = 3.01, and Distinction = 3.69. The recommended cut score for the Pass and Merit results bands increased slightly between the judgement rounds, but for the Distinction results band remained the same.

	Borderline		**Merit**		**Distinction**	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	2.00	2.10	2.93	2.95	3.70	3.70
J02	2.20	2.20	3.00	3.00	3.60	3.60
J03	2.20	2.20	3.30	3.30	3.80	3.80
J04	1.50	1.90	2.70	2.70	3.50	3.50
J05	2.50	2.80	3.30	3.30	3.80	3.80
J06	1.40	1.70	2.60	2.60	3.60	3.60
J07	2.20	2.50	3.20	3.30	3.80	3.80
J08	3.00	3.00	3.00	3.00	4.00	4.00
J09	2.50	2.80	3.00	3.20	3.50	3.50
J10	1.80	2.10	2.80	2.90	3.80	3.80
J11	2.20	2.40	2.90	2.90	3.50	3.50
Min	1.40	1.70	2.60	2.60	3.50	3.50
Max	3.00	3.00	3.30	3.30	4.00	4.00
Mean	2.14	2.34	2.98	3.01	3.69	3.69
SD	.44	.39	.22	.23	.16	.16

Table 4.8: Cut score judgements for the Listening component: ISE III

4.3 READING CUT SCORES

The cut scores for the Reading component were set using the Yes/No Angoff method (described in Section 2.1). This section presents a table of cut-score judgements for each ISE level – four tables in total. For each table, the first column identifies the panellist's ID number. Since each ISE level has three results bands - Pass, Merit, and Distinction, where the Pass cut-score delimits the minimally competent candidate at that ISE level, the subsequent columns show each panellist's suggested cut score for each judgement round by results band. The bottom section of each table shows the group's descriptive summary statistics. For this study, the final recommended cut score is the mean (average) of the panellist judgements.

Table 4.9 shows that the recommended cut scores (after judgement round 2) for ISE Foundation were: Pass = 14.10, Merit = 22.60, and Distinction = 28.80. For the Pass and Distinction results bands, the recommended cut score increased between the judgement rounds, but for the Merit results band, the recommended cut score decreased by .20 of a raw score point.

	Borderline		**Merit**		**Distinction**	
Judge ID	**Round 1**	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
J01	10	11	22	22	30	30
J02	16	17	22	23	29	30
J03[1]	-	-	-	-	-	-
J04	16	14	18	22	24	27
J05	17	17	25	25	29	29
J06	12	13	26	26	30	30
J07	08	10	18	19	24	27
J08	13	13	18	22	30	29
J09	15	20	23	22	30	27
J10	15	08	18	20	30	30
J11	18	18	26	23	30	29
Min	08	08	18	19	24	27
Max	18	20	26	26	30	30
Mean	14.00	14.10	21.60	22.40	28.60	28.80
SD	3.03	3.65	3.23	1.96	2.33	1.25

Table 4.9: Cut score judgements for the Reading component: ISE Foundation

[1] Judge J03 could not attend this session

Table 4.10 shows the recommended cut scores for ISE I (after judgement round 2). They were Pass = 14.55, Merit = 23.18, and Distinction = 29.00. The recommended cut score for the Pass results band increased between the judgement rounds, but for the Merit and Distinction results bands, the recommended cut score decreased.

	Borderline		**Merit**		**Distinction**	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	13	12	23	24	30	30
J02	12	12	22	22	30	30
J03	14	14	23	23	30	28
J04	14	15	26	23	30	27
J05	12	12	24	24	30	30
J06	17	17	25	25	30	30
J07	14	15	23	23	30	30
J08	17	17	27	24	29	29
J09	14	17	20	21	28	28
J10	15	10	28	24	30	30
J11	17	19	24	22	29	27
Min	12	10	20	21	28	27
Max	17	19	28	25	30	30
Mean	14.45	14.55	24.09	23.18	29.64	29.00
SD	1.78	2.68	2.19	1.11	.64	1.21

Table 4.10: Cut score judgements for the Reading component: ISE I

Table 4.11 shows the recommended cut scores for ISE II (after judgement round 2). They were Pass = 14.00, Merit = 23.82, and Distinction = 28.91. For all three results bands, the recommended cut score increased slightly between the judgement rounds.

	Borderline		**Merit**		**Distinction**	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	10	15	28	29	30	30
J02	09	09	22	22	29	29
J03	11	13	22	22	30	30
J04	12	12	19	22	25	28
J05	16	13	25	25	30	30
J06	17	15	27	27	30	30
J07	15	15	21	21	30	30
J08	13	13	25	25	30	30
J09	17	19	20	23	25	25
J10	12	11	20	22	27	27
J11	17	19	23	24	27	29
Min	09	09	19	21	25	25
Max	17	19	28	29	30	30
Mean	13.55	14.00	22.91	23.82	28.45	28.91
SD	2.84	2.92	2.84	2.37	1.97	1.56

Table 4.11: Cut score judgements for the Reading component: ISE II

Table 4.12 shows the recommended cut scores for ISE III (after judgement round 2). They were Pass = 15.82, Merit = 23.82, and Distinction = 28.82. The recommended cut score for all three results bands increased slightly between the judgement rounds.

	Borderline		**Merit**		**Distinction**	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	13	14	21	25	30	30
J02	16	16	24	26	29	29
J03	09	15	22	23	30	30
J04	16	15	21	21	28	27
J05	13	13	24	24	30	30
J06	17	17	25	25	30	30
J07	14	13	21	21	28	28
J08	15	16	23	24	27	27
J09	17	17	23	23	27	28
J10	20	17	27	24	30	30
J11	20	21	23	26	26	28
Min	09	13	21	21	26	27
Max	20	21	27	26	30	30
Mean	15.45	15.82	23.09	23.82	28.64	28.82
SD	3.06	2.17	1.78	1.64	1.43	1.19

Table 4.12: Cut score judgements for the Reading component: ISE III

4.4 WRITING CUT SCORES

The cut scores for the Writing component were determined using the Performance Profile method, as described in Section 2.1. This section presents a table of cut-score judgements for each ISE level – four tables in total. For each table, the first column identifies the panellist's ID number. Since each ISE level has three results bands - Pass, Merit, and Distinction, where the Pass cut-score delimits the minimally competent candidate at that ISE level, the subsequent columns show each panellist's suggested cut score for each judgement round by results band. The bottom section of each table shows the group's descriptive summary statistics. For this study, the final recommended cut score is the mean (average) of the panellist judgements.

Table 4.13 shows that the recommended cut scores (after judgement round 2) for ISE Foundation were: Pass = 13.07, Merit = 19.34, and Distinction = 24.29. The recommended cut score decreased between the judgement rounds for all three result bands.

Table 4.14 shows that the recommended cut scores (after judgement round 2) for ISE I were: Pass = 13.53, Merit = 19.68, and Distinction = 24.96. The recommended cut score decreased between the judgement rounds for all three results bands.

	Borderline		Merit		Distinction	
Judge ID	**Round 1**	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
J01	13.00	13.00	21.00	21.00	26.50	26.50
J02	13.00	11.60	19.50	17.80	25.50	24.40
J03[2]	-	-	-	-	-	-
J04	12.30	12.30	18.80	18.80	25.60	25.60
J05	13.70	11.60	21.00	17.80	27.25	24.40
J06	16.00	15.00	22.00	22.00	25.00	25.25
J07	13.00	13.00	19.00	19.00	25.00	24.75
J08	17.00	13.00	23.00	18.00	25.00	20.00
J09	14.50	14.20	18.90	18.90	21.80	21.40
J10	13.75	12.95	19.75	19.75	27.25	27.20
J11	14.00	14.00	20.20	20.30	23.50	23.40
Min	12.30	11.60	18.80	17.80	21.80	20.00
Max	17.00	15.00	23.00	22.00	27.25	27.20
Mean	14.03	13.07	20.32	19.34	25.24	24.29
SD	1.39	1.04	1.34	1.34	1.58	2.09

Table 4.13: Cut score judgements for the Writing component: ISE Foundation

	Borderline		Merit		Distinction	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	13.40	14.00	20.40	21.10	25.00	26.00
J02	13.20	13.20	20.35	20.35	25.75	25.75
J03	13.20	13.20	22.10	22.10	26.00	26.00
J04	12.30	12.30	18.80	18.80	25.50	25.50
J05	13.50	13.40	21.30	20.40	27.40	25.00
J06	12.50	12.50	19.30	19.30	25.10	25.10
J07	13.70	13.70	21.20	21.20	24.80	24.80
J08	15.00	15.00	15.00	15.00	22.00	22.00
J09	14.00	14.00	19.40	18.10	24.00	23.80
J10	13.50	13.50	20.20	20.20	26.30	26.30
J11	14.00	14.00	19.90	19.90	24.10	24.30
Min	12.30	12.30	15.00	15.00	22.00	22.00
Max	15.00	15.00	22.10	22.10	27.40	26.30
Mean	13.48	13.53	19.81	19.68	25.09	24.96
SD	.70	.72	1.78	1.83	1.35	1.19

Table 4.14: Cut score judgements for the Writing component: ISE I

[2] Judge J03 could not attend this session

Table 4.15 shows that the recommended cut scores (after judgement round 2) for ISE II were: Pass = 15.51, Merit = 20.92, and Distinction = 25.82. For the Pass and Distinction results bands, the recommended cut score increased between the judgement rounds, but for the Merit results band, the recommended cut score decreased by .06 of a raw score point.

	Borderline		**Merit**		**Distinction**	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	14.00	14.00	21.10	21.10	26.00	26.00
J02	13.00	14.00	19.50	20.20	25.10	25.40
J03	15.00	15.00	23.00	23.00	26.70	26.70
J04	12.90	13.30	20.50	20.50	26.00	26.00
J05	14.00	13.70	21.90	21.00	27.40	27.25
J06	12.30	12.30	18.90	18.90	24.80	24.80
J07	14.80	15.20	21.60	22.00	24.90	25.00
J08	18.00	18.00	22.00	22.00	27.00	27.00
J09	15.00	15.00	19.50	20.10	24.20	24.20
J10	15.10	15.10	23.70	21.30	27.00	27.00
J11	14.00	14.00	19.10	20.00	24.70	24.70
Min	12.30	12.30	18.90	18.90	24.20	24.20
Max	18.00	18.00	23.70	23.00	27.40	27.25
Mean	14.37	14.51	20.98	20.92	25.80	25.82
SD	1.46	1.39	1.54	1.09	1.06	1.02

Table 4.15: Cut score judgements for Writing component: ISE II

Table 4.16 shows that the recommended cut scores (after judgement round 2) for ISE II were: Pass = 15.62, Merit = 21.27, and Distinction = 25.81. For the Pass and Merit results bands, the recommended cut score increased between the judgement rounds, but for the Distinction results band, the recommended cut score decreased by .09 of a raw score point.

	Borderline		**Merit**		**Distinction**	
Judge ID	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
J01	14.00	15.40	21.10	21.20	25.70	25.00
J02	14.60	15.00	20.80	20.80	25.20	25.20
J03	16.10	16.10	23.30	23.30	27.00	27.00
J04	13.70	13.70	20.30	20.30	26.50	26.50
J05	18.80	18.60	22.70	22.70	26.65	26.65
J06	11.40	11.40	17.30	17.30	24.80	24.80
J07	15.20	16.00	21.60	21.60	25.40	25.40
J08	19.00	18.00	24.00	24.00	28.00	28.00
J09	16.10	15.90	19.70	19.70	24.90	24.90
J10	15.40	15.70	21.00	21.00	26.10	25.90
J11	16.00	16.00	22.10	22.10	24.60	24.60
Min	11.40	11.40	17.30	17.30	24.60	24.60
Max	19.00	18.60	24.00	24.00	28.00	28.00
Mean	15.48	15.62	21.26	21.27	25.90	25.81
SD	2.07	1.84	1.76	1.76	1.01	1.04

Table 4.16: Cut score judgements for Writing component: ISE III

4.5 SUMMARY OF RECOMMENDED CUT SCORES

Tables 4.17 to 4.20 summarise the recommended cut scores for each ISE level. They are based on the panellist mean ratings from judgement round 2.

	Borderline	Merit	Distinction
Reading	14.10	22.40	28.80
Writing	13.07	19.34	24.29
Listening	4.00	7.46	8.74
Speaking	7.76	12.45	15.09

Table 4.17: Round 2 Cut score recommendations: ISE Foundation

	Borderline	Merit	Distinction
Reading	14.55	23.18	29.00
Writing	13.53	19.68	24.96
Listening	4.71	8.17	9.35
Speaking	8.75	12.67	15.27

Table 4.18: Round 2 Cut score recommendations: ISE I

	Borderline	Merit	Distinction
Reading	14.00	23.82	28.91
Writing	14.51	20.92	25.82
Listening	2.22	3.00	3.69
Speaking	8.81	12.28	15.01

Table 4.19: Round 2 Cut score recommendations: ISE II

	Borderline	Merit	Distinction
Reading	15.82	23.82	28.82
Writing	15.62	21.27	25.81
Listening	2.34	3.01	3.69
Speaking	8.69	12.17	14.91

Table 4.20: Round 2 Cut score recommendations: ISE III

Section 5 presents the analyses conducted to confirm the validity of these cut scores.

5. Cut Score Validation

Standard setting studies are evaluated in terms of three types of validity evidence: Procedural, Internal, and External, as illustrated in Table 5.1 (source: Hambleton & Pitoniak, 2006). Elements of Procedural validity were described in the methodology section (Section 2). This section presents evidence of internal validation.

Procedural	Internal	External
◗ Explicitness ◗ Practicability ◗ Implementation of procedures ◗ Panellist feedback ◗ Documentation	◗ Intraparticipant consistency ◗ Interparticipant consistency ◗ Consistency within method ◗ Decision consistency	◗ Comparisons to other methods ◗ Comparisons to other sources of information ◗ Reasonableness of performance levels

Table 5.1: Standard setting evaluation elements

5.1 CUT-SCORE VALIDATION ANALYSIS

This section presents the cut score validation analyses conducted to examine intraparticipant consistency, interparticipant consistency, consistency within the method, and decision accuracy. To ensure methodological rigour and coherence across phases of the standard setting process, a two-stage approach was adopted.The cut scores were established in 2015 using pretest data; however, limited sample sizes at that stage restricted the robustness of analyses related to classification accuracy. In line with the original methodological framework, the cut scores were subsequently re-evaluated in 2016 using operational (live) test data from a full year of test administrations. This more representative dataset enabled a reliable evaluation of decision accuracy and consistency, while intraparticipant and interparticipant consistency were examined using judgment data from the 2015 standard setting workshop.

It is important to note that test forms are constructed to be comparable in difficulty across years. For Listening and Reading, statistical equating procedures are applied, while for Speaking and Writing, comparability is maintained through careful task selection, rigorous rater and examiner training, and supporting evidence from small-scale pilots. These practices uphold the validity of applying the original cut scores to the 2016 dataset, ensuring that the resulting decision metrics are both meaningful and generalisable.

5.2 INTRAPARTICIPANT CONSISTENCY

Hambleton and Pitoniak (2006, p. 458) define intraparticipant consistency as "the degree to which a panellist can provide ratings that are consistent with the empirical difficulties and the degree to which ratings change across rounds". As pretesting data was used, and the sample size was small, intraparticipant consistency was investigated solely by examining the degree to which ratings of each panellist changed across rounds, keeping in mind the warning by Hambleton, Pitoniak, and Copella (2012) that when panellists do not change their ratings, they may not be considering the feedback provided between rounds. We would expect to see some changes in judgements between rounds. However, in cases in which panellists do not make any changes, it may imply that panellists are happy with their original Round 1 ratings.

In what follows, intraparticipant consistency is inspected by component and by ISE level, beginning with the speaking component.

Speaking Component

Tables 5.2 – 5.5 show the intraparticipant consistency for the ISE Foundation – ISE III speaking components. Rating changes between round 1 and round 2 are highlighted in grey. Apart from ISE I, where there were several changes between rounds 1 and 2, panellists tended not to make changes to their judgements between round 1 and round 2 for any of the results bands. Possible reasons for this could be that:

- For this component, only total scores were presented, and minor changes in ratings may not be observable when total scores remain unchanged or when panellists are satisfied with their ratings.
- A four-band scale was used for each of the four criteria.
- Panellists were overall satisfied with their Round 1 ratings.

	Borderline	Merit	Distinction
Judge ID	**Round 1 – Round 2**	**Round 1 – Round 2**	**Round 1 – Round 2**
J01	.00	.00	.00
J02	.00	.00	.00
J03	.00	.00	.00
J04	.00	.00	.00
J05	.00	.00	.00
J06	.00	.00	.00
J07	.30	.00	.00
J08	.00	.00	.00
J09	2.30	.00	.00
J10	.00	.00	.50
J11	.00	.00	.00

Table 5.2: Intraparticipant consistency: Changes in ratings across rounds ISE Foundation Speaking component

	Borderline	Merit	Distinction
Judge ID	**Round 1 – Round 2**	**Round 1 – Round 2**	**Round 1 – Round 2**
J01	.00	.00	.00
J02	-.50	.00	.00
J03	.00	-.40	-.40
J04	.00	-.80	.00
J05	-.40	.00	.00
J06	.00	.00	.00
J07	-.80	-.30	-.60
J08	.00	.00	.00
J09	-.10	.00	.00
J10	-.30	-1.60	.00
J11	-.20	.00	.20

Table 5.3: Intraparticipant consistency: Changes in ratings across rounds ISE I Speaking component

	Borderline	Merit	Distinction
Judge ID	**Round 1 – Round 2**	**Round 1 – Round 2**	**Round 1 – Round 2**
J01	-.20	-.50	.25
J02	.00	.00	.00
J03	.00	.00	.00
J04	.00	.00	.20
J05	.00	.00	.00
J06	.00	-1.00	.00
J07	-.60	.40	.20
J08	.00	.00	.00
J09	.50	.00	.00
J10	.00	.00	.00
J11	-.80	.00	.00

Table 5.4: Intraparticipant consistency: Changes in ratings across rounds ISE II Speaking component

	Borderline	Merit	Distinction
Judge ID	**Round 1 – Round 2**	**Round 1 – Round 2**	**Round 1 – Round 2**
J01	.00	.00	.00
J02	.00	.00	.00
J03	.00	.00	.00
J04	.00	-.20	.00
J05	.00	.00	.00
J06	.00	.00	.00
J07	-.20	-.40	.00
J08	.00	.00	.00
J09	.00	.00	.00
J10	-.50	.00	.30
J11	.00	.00	.00

Table 5.5: Intraparticipant consistency: Changes in ratings across rounds ISE III Speaking component

Overall, the panellists made few changes across rounds, and when they did change their judgements between rounds, the changes tended to be small. The minimal nature of the changes (when they occurred) is attributable to the nature of the test and the rating scale, which entails assigning a score out of four for each criterion. The relative stability of the judgements between rounds, therefore, signals intraparticipant consistency for the speaking component.

Listening Component

Tables 5.6-5.9 display the intraparticipant consistency for the ISE Foundation – ISE III listening components. Rating changes between round 1 and round 2 are highlighted in grey. The tables show that, for ISE Foundation and ISE I, the panellists made several changes in their judgements between rounds 1 and 2. However, there were relatively fewer changes in judgements between round 1 and round 2 for ISE II and III. This can likely be explained by the differing scale lengths for ISE II and ISE III, where candidates receive only one score from a rating scale.

It is also important to note that some panellists tended not to change their judgements (see J03). Though it is possible to argue that these panellists did not absorb the feedback between judgement rounds, it is more likely that they were generally more satisfied with their initial cut scores and chose to retain them.

	Borderline	Merit	Distinction
Judge ID	**Round 1 – Round 2**	**Round 1 – Round 2**	**Round 1 – Round 2**
J01	1.00	.00	.00
J02	-1.00	-3.00	.00
J03	.00	.00	.00
J04	.20	-1.00	.00
J05	-1.00	.00	.00
J06	.00	.00	.00
J07	-1.00	.00	.00
J08	1.00	.00	.00
J09	-1.20	.00	.00
J10	-1.80	.00	.00
J11	.20	1.00	-.10

Table 5.6: Intraparticipant consistency: Changes in ratings across rounds, ISE Foundation Listening component

	Borderline	Merit	Distinction
Judge ID	**Round 1 – Round 2**	**Round 1 – Round 2**	**Round 1 – Round 2**
J01	-1.00	.00	.20
J02	1.00	.00	.00
J03	2.00	.00	.00
J04	1.60	.20	1.20
J05	1.00	.50	.00
J06	1.00	-.20	.20
J07	.00	.00	.00
J08	2.00	.00	.00
J09	-1.20	1.20	.00
J10	2.00	.00	.00
J11	.20	1.00	1.00

Table 5.7: Intraparticipant consistency: Changes in ratings across rounds, ISE I Listening component

	Borderline	Merit	Distinction
Judge ID	**Round 1 – Round 2**	**Round 1 – Round 2**	**Round 1 – Round 2**
J01	.00	.00	.00
J02	-.20	-.10	.00
J03	.00	.00	.00
J04	.00	-.10	.00
J05	.00	.00	.00
J06	.00	.00	.00
J07	.00	.00	.10
J08	.00	.00	.00
J09	.00	.00	.00
J10	.00	.00	.00
J11	.00	.00	.00

Table 5.8: Intraparticipant consistency: Changes in ratings across rounds, ISE II Listening component

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	-.10	-.02	.00
J02	.00	.00	.00
J03	.00	.00	.00
J04	-.40	.00	.00
J05	-.30	.00	.00
J06	-.30	.00	.00
J07	-.30	-.10	.00
J08	.00	.00	.00
J09	-.30	-.20	.00
J10	-.30	-.10	.00
J11	-.20	.00	.00

Table 5.9: Intraparticipant consistency: Changes in ratings across rounds, ISE III Listening component

When interpreting the data for the listening component, it is important to remember that the length of the scale varies between components. For ISE Foundation and ISE I, the scale is 9 and 10 raw score points, respectively. For ISE II and III, the scale is 4 raw score points. Bearing this in mind, though there were several judgement changes between rounds, the changes tended to be small adjustments rather than large recalibrations. It is also notable that most of the panellists' cut score changes for every ISE level were in their estimation of a passing score. This suggests that the discussion between judgement rounds was constructive in clarifying the panellists' understanding of the minimal competence required to be at the level. Overall, the data signals intraparticipant consistency for the listening component.

Reading Component

Tables 5.10 – 5.13 show the intraparticipant consistency for the ISE Foundation – ISE III reading components. Rating changes between round 1 and round 2 are highlighted in grey. The tables show that, for ISE Foundation the panellists made several changes in their judgements between rounds 1 and 2 and that the changes in judgements between round 1 and round 2 for the other ISE levels tended to be for the Pass results band, indicating that the discussions helped the panellists to clarify their view of the minimal level of competence required to pass each ISE level.

It is also important to note that some panellists tended not to change their judgements (see J05). As with other cases where panellists did not change their judgements, we would argue that these panellists absorbed the feedback between judgement rounds but were generally satisfied with their initial cut scores and chose to retain them.

When interpreting the data for the reading component, it is essential to remember that the scale length is 30 raw score points. Bearing this in mind, though there were several judgement changes between rounds across all ISE levels and results bands, the changes tended to be minor adjustments relative to the length of the scale, rather than large recalibrations. As with the speaking and listening components, for the reading component, most of the panellists' cut score changes at every ISE level were in their estimation of a passing score. This suggests that the discussion between judgement rounds was constructive in clarifying the panellists' understanding of the minimal competence required to be at the level. Overall, the relative stability of the judgements between rounds signals intraparticipant consistency for the reading component.

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	-1.00	.00	.00
J02	-1.00	-1.00	-1.00
J03[3]	-	-	-
J04	2.00	-4.00	-3.00
J05	.00	.00	.00
J06	-1.00	.00	.00
J07	-2.00	-1.00	-3.00
J08	.00	-4.00	1.00
J09	-5.00	1.00	3.00
J10	7.00	-2.00	.00
J11	.00	3.00	1.00

Table 5.10: Intraparticipant consistency: Changes in ratings across rounds, ISE Foundation Reading component

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	1.00	-1.00	.00
J02	.00	.00	.00
J03	.00	.00	2.00
J04	-1.00	3.00	3.00
J05	.00	.00	.00
J06	.00	.00	.00
J07	-1.00	.00	.00
J08	.00	3.00	.00
J09	-3.00	-1.00	.00
J10	5.00	4.00	.00
J11	-2.00	2.00	2.00

Table 5.11: Intraparticipant consistency: Changes in ratings across rounds, ISE I Reading component

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	-5.00	-1.00	.00
J02	.00	.00	.00
J03	-2.00	.00	.00
J04	.00	-3.00	-3.00
J05	3.00	.00	.00
J06	2.00	.00	.00
J07	.00	.00	.00
J08	.00	.00	.00
J09	-2.00	-3.00	.00
J10	1.00	-2.00	.00
J11	-2.00	-1.00	-2.00

Table 5.12: Intraparticipant consistency: Changes in ratings across rounds, ISE II Reading component

[3] Judge J03 could not attend this session

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	-1.00	-4.00	.00
J02	.00	-2.00	.00
J03	-6.00	-1.00	.00
J04	1.00	.00	1.00
J05	.00	.00	.00
J06	.00	.00	.00
J07	1.00	.00	.00
J08	-1.00	-1.00	.00
J09	.00	.00	-1.00
J10	3.00	3.00	.00
J11	-1.00	-3.00	-2.00

Table 5.13: Intraparticipant consistency: Changes in ratings across rounds, ISE III Reading component

Writing Component

Tables 5.14 – 5.17 show the intraparticipant consistency for the ISE Foundation – ISE III writing components. Rating changes between round 1 and round 2 are highlighted in grey. The tables show that, for ISE Foundation, the panellists made several changes in their judgements between rounds 1 and 2 for all the results bands. The changes in judgements between round 1 and round 2 for the other ISE levels tended to be for the Pass results band, indicating that the discussions were particularly helpful in clarifying the panellists' view of the minimal level of competence required to pass each ISE level.

It is also important to note that some panellists tended not to change their judgements (see J04). We acknowledge that an argument could be made that these panellists did not absorb the feedback between judgement rounds but would maintain that it is more likely that they were generally more satisfied with their initial cut scores and chose to retain them.

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	.00	.00	.00
J02	1.40	1.70	1.10
J03[4]	-	-	-
J04	.00	.00	.00
J05	2.10	3.20	2.85
J06	1.00	.00	-.25
J07	.00	.00	.25
J08	4.00	5.00	5.00
J09	.30	.00	.40
J10	.80	.00	.05
J11	.00	-0.10	.10

Table 5.14: Intraparticipant consistency: Changes in ratings across rounds, ISE Foundation Writing component

[4] Judge J03 could not attend this session

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	-.60	-.70	-1.00
J02	.00	.00	.00
J03	.00	.00	.00
J04	.00	.00	.00
J05	.10	.90	2.40
J06	.00	.00	.00
J07	.00	.00	.00
J08	.00	.00	.00
J09	.00	1.30	.20
J10	.00	.00	.00
J11	.00	.00	-.20

Table 5.15: Intraparticipant consistency: Changes in ratings across rounds, ISE I Writing component

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	.00	.00	.00
J02	-1.00	-.70	-.30
J03	.00	.00	.00
J04	-.40	.00	.00
J05	.30	.90	.15
J06	.00	.00	.00
J07	-.40	-.40	-.10
J08	.00	.00	.00
J09	.00	-.60	.00
J10	.00	2.40	.00
J11	.00	-.90	.00

Table 5.16: Intraparticipant consistency: Changes in ratings across rounds, ISE II Writing component

Judge ID	Borderline Round 1 – Round 2	Merit Round 1 – Round 2	Distinction Round 1 – Round 2
J01	-1.40	-.10	.70
J02	-.40	.00	.00
J03	.00	.00	.00
J04	.00	.00	.00
J05	.20	.00	.00
J06	.00	.00	.00
J07	-.80	.00	.00
J08	1.00	.00	.00
J09	.20	.00	.00
J10	-.30	.00	.20
J11	.00	.00	.00

Table 5.17: Intraparticipant consistency: Changes in ratings across rounds ISE I Writing component

When interpreting the data for the writing component, it is essential to remember that the scale length for this component is 28 raw score points, a maximum of four score points per each of the seven criteria.

Bearing this in mind, though there were several judgement changes between rounds across all ISE levels and results bands (particularly for ISE Foundation), the changes tended to be minor adjustments relative to the scale length rather than large recalibrations. As with all the other components, the panellists' cut score changes at every ISE level tended to be in their estimation of a passing score. This suggests that the discussion between judgement rounds was constructive in clarifying the panellists' understanding of the minimal competence required to be at the level. Overall, the relative stability of the judgements between rounds signals intraparticipant consistency for the writing component.

Taking the results for all four components together, the intraparticipant consistency analysis shows that panellists were willing to and did make changes to their judgements between rounds. These changes were typically small relative to the raw scale for each component and constituted appropriate refinements of their initial judgements. Therefore, evidence supports the claim that most panellists considered the feedback presented between rounds, adding further evidence of intraparticipant consistency.

5.3 INTERPARTICIPANT CONSISTENCY: CLASSICAL TEST THEORY (CTT) ANALYSIS

Hambleton and Pitoniak (2006, p. 458) define interparticipant consistency as "the consistency of item ratings and performance standards across panellists". As "there is no single perfect statistical index for the estimation of inter-rater reliability" (Kaftandjieva & Takala, 2002, p. 111), this analysis presents both Cronbach's alpha and the intraclass correlation (ICC, McGraw & Wong, 1996; Shrout & Fleiss, 1979) of the panellists' estimates. Cronbach's alpha and ICC are consensus interrater agreement indices. The model used to calculate the ICC was the two-way mixed model, average measures for exact agreement. A high alpha estimate indicates that panellists' ratings measure a common dimension. A high intraclass correlation (close to 1) suggests that panellists have achieved excellent interrater reliability (Stemler & Tsai, 2008). For both measures, reliability estimates should be at least .80 to "reflect good dependability of scores" (Hyot, 2010, p. 152). The tables that follow present the interparticipant consistency indices by ISE level. ICC confidence intervals are provided in brackets.

	Round 1		Round 2	
Section	**Alpha**	**ICC**	**Alpha**	**ICC**
Reading	.84	.84 CI [.78 – .88]	.88	.87 CI [.79 – .93]
Writing	.96	.96 CI [.92 – .98]	.98	.97 CI [.95 – .99]
Listening	.99	.99 CI [.98 – 1.00]	.99	.99 CI [.98 – 1.00]
Speaking	.99	.99 CI [.97 – 1.00]	.99	.99 CI [.98 – 1.00]

Table 5.18: Interparticipant consistency: ISE Foundation

	Round 1		Round 2	
Section	**Alpha**	**ICC**	**Alpha**	**ICC**
Reading	.91	.91 CI [.87 – .93]	.91	.91 CI [.89 – .94]
Writing	.99	.99 CI [.98 – .99]	.99	.99 CI [.98 – .99]
Listening	.99	.99 CI [.98 – .99]	.99	.99 CI [.98 – .99]
Speaking	.99	.99 CI [.96 – 1.00]	.99	.99 CI [.97 – 1.00]

Table 5.19: Interparticipant consistency: ISE I

	Round 1		Round 2	
Section	**Alpha**	**ICC**	**Alpha**	**ICC**
Reading	.83	.83 CI [.77 – .88]	.88	.88 CI [.83 – .91]
Writing	.99	.98 CI [.97 – .99]	.99	.98 CI [.97 – .99]
Listening	.99	.99 CI [.96 – 1.00]	.99	.99 CI [.97 – 1.00]
Speaking	.99	.98 CI [.96 – .99]	.99	.98 CI [.98 – 1.00]

Table 5.20: Interparticipant consistency: ISE II

	Round 1		Round 2	
Section	**Alpha**	**ICC**	**Alpha**	**ICC**
Reading	.81	.81 CI [.74 - .86]	.85	.85 CI [.80 - .89]
Writing	.99	.98 CI [.96 - .99]	.99	.98 CI [.96 - .99]
Listening	.99	.99 CI [.94 – 1.00]	.99	.98 CI [.93 – 1.00]
Speaking	.99	.98 CI [.95 - .99]	.99	.98 CI [.95 - .99]

Table 5.21: Interparticipant consistency: ISE III

The tables clearly show that, for every ISE level, inter-rater consistency in the round 1 judgements met the minimum recommended by Hyot (2010). The consistency estimates unilaterally rose between rounds 1 and 2, ranging from .85 to .99 and generally reaching close to or over .90. The ICC confidence intervals were also very high. Taken together, it is clear that the panellists were "very homogeneous in terms of exact agreement as well as in terms of association" (Kaftandjieva & Takala, 2002, p. 113).

5.4 RASCH ANALYSIS OF INTRA- & INTERPARTICIPANT CONSISTENCY

This section also analyses interparticipant consistency using Multifaceted Rasch measurement (MFRM) analysis. Measures are presented in logits and were retrieved through FACETS (Linacre, 2014). MFRM was also used to investigate intraparticipant consistency. The tables that follow present the relevant statistics for each consistency analysis:

- **Separation (G)** and **Strata (H):** Both are separation indices, indicators of how widely different panellists are in terms of their severity/leniency. Ideally, G should be close to 0, and H should be close to 1, indicating that panellists are not substantially different in their severity of judgements.
- **Reliability (R):** This captures differences in judgements between panellists. Ideally, this number should be low, indicating that differences in panellist judgements are attributable to chance.
- **Observed Agreement (%)** and **Expected Agreement (%):** These figures show the actual number of times panellists gave the same score/judgement (observed, expressed as a percentage), contrasted with the expectations of the model (expected agreement). Ideally, these numbers should be close, as this indicates good correspondence between the actual data and the model's expectations. The same index also shows whether the panellists acted as independent experts. For this reason, ideally, indices lower than .90% should be observed.
- **Rasch–Kappa:** Rasch Kappa is an inter-rater agreement index that should be close to 0.00. In a standard setting context, this index allows practitioners to evaluate the degree of rater dependence in a given dataset. Values much larger than 0.00 indicate overly high interrater agreement and, consequently, a high degree of local rater dependence; large negative values indicate much less interrater agreement than expected based on the Rasch model, which may be due to unmodeled sources of variation in the ratings (e.g., hidden facets).
- **Mean Infit** and **Mean Infit Standard Deviation (SD)**: The acceptable infit range is calculated as follows: infit mean± twice the infit standard deviation (Pollitt & Hutchinson, 1987). Ideally, all panellists should fall within the acceptable infit range. Panellists identified as misfitting should be eliminated from the analysis. To retain maximum data at each decision point, this can be done case-by-case, at the ISE level, and by component.

A more detailed discussion of the tables (by component) follows. Misfitting panellists were removed in each case before the consistency statistics were finalised. After eliminating misfitting panellists, all the tables list close to zero separation indices (strata) H and G, showing that the panellists did not exercise different severity levels when assigning the recommended cut scores across the different skills and all ISE levels. Additionally, the consistently low-reliability index suggests that the panellists were not reliably different, and any minute differences could have been attributed to chance. Such findings are highly desirable, as judge severity did not directly impact the recommended cut scores. The Rasch Kappa figure also corroborates the finding of high agreement in the panellists' ratings.

Speaking Component

The initial analysis for ISE Foundation revealed one misfitting panellist (J08, Infit measure of 5.72 with a Zstd of 2.1). After this panellist was eliminated and the data re-run, another panellist was misfitting (J04) and had to be removed. None of the remaining nine panellists was misfitting, and, as a consequence of this exercise, the overall mean for the Pass result band for Round 2 dropped slightly from 7.8 to 7.7. The standard deviation of the judges remained at 0.4. The group's Alpha increased from 0.99 to 1.00, and the ICC remained at 0.99. The separation indices G, H, and R remained the same. The mean infit was 174 with a standard deviation of 0.94.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.00	0.00	0.34	0.34	0.00	0.00
Strata (H)	0.33	0.33	0.78	0.78	0.33	0.33
Reliability (R)	0.00	0.00	0.10	0.10	0.00	0.00
Obs. Agree (%)	33.6	43.2	13.6	13.6	23.6	21.4
Exp. agree (%)	34.9	43.8	13.9	13.9	26.2	24.4
Rasch – Kappa	-0.02	-0.01	0.00	0.00	-0.04	-0.04
Min. *Infit (ZStd)*	0.02 (1.5)	0.03 (1.3)	0.06 (0.9)	0.06 (.9)	0.02 (1.6)	0.06 (0.9)
Max. *Infit (ZStd)*	8.24 (2.4)	5.72 (2.1)	2.42 (1.7)	2.42 (1.7)	2.76 (1.2)	2.33 (1.1)
Mean *Infit*	1.12	1.03	1.02	1.02	0.87	0.94
Mean *Infit SD*	2.28	1.60	0.79	0.79	0.84	0.76

Table 5.22: Rasch Interparticipant and interparticipant consistency: ISE Foundation Speaking section

No panellist exhibited misfit for ISE I, and all were retained for the analysis. The separation ratio (G) was 0.00, the strata (H) index was 0.00, and the reliability index was 0.00 in all rounds and result bands. The Rasch Kappa index ranged from -0.01 to -0.04.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.00	0.00	0.00	0.00	0.00	0.00
Strata (H)	0.33	0.33	0.33	0.33	0.33	0.33
Reliability (R)	0.00	0.00	0.00	0.00	0.00	0.00
Obs. Agree (%)	24.5	23.2	11.4	11.8	20.0	23.6
Exp. agree (%)	24.9	24.0	14.3	14.2	23.3	26.3
Rasch – Kappa	-0.01	-0.01	-0.03	-0.03	-0.04	-0.04
Min. *Infit (ZStd)*	0.08 (.2)	0.13 (1.6)	0.15 (.5)	0.15 (0.5)	0.05 (1.4)	0.09 (1.6)
Max. *Infit (ZStd)*	2.98 (1.4)	3.75 (1.5)	2.09 (1.4)	4.28 (1.8)	1.39 (1.1)	1.58 (.9)
Mean *Infit*	0.95	0.81	0.90	1.02	0.99	0.87
Mean *Infit SD*	1.00	1.02	0.71	1.14	1.11	0.71

Table 5.23: Rasch Interparticipant and interparticipant consistency: ISE I Speaking section

As with ISE I, no panellist exhibited misfit for ISE II. The separation ratio (G) ranged from 0.00 to 1.54. The strata (H) index ranged from 0.33 to 2.39 and the reliability index ranged from 0.00 to 0.70. The Rasch Kappa index ranged from -0.04 to 0.00.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	1.54	0.39	0.96	0.00	0.00	0.00
Strata (H)	2.39	0.86	1.62	0.33	0.33	0.33
Reliability (R)	0.70	0.13	0.48	0.00	0.00	0.00
Obs. Agree (%)	16.4	17.7	11.4	12.3	15.9	14.1
Exp. agree (%)	16.4	18.3	15.2	15.9	18.8	17.6
Rasch – Kappa	0.00	-0.01	-0.04	-0.04	-0.04	-0.04
Min. *Infit (ZStd)*	0.35 (1.3)	0.01 (4.6)	0.29 (1.0)	0.17 (.5)	0.07 (0.7)	0.08 (0.6)
Max. *Infit (ZStd)*	2.47 (1.3)	1.79 (1.1)	2.28 (1.2)	2.87 (1.4)	2.12 (1.3)	1.99 (1.2)
Mean *Infit*	0.82	0.78	0.75	0.93	0.94	0.94
Mean *Infit SD*	0.61	0.49	0.75	0.78	0.57	0.56

Table 5.24: Rasch Interparticipant and interparticipant consistency: ISE II Speaking section

Once again, no panellist exhibited misfit for ISE III, and all 11 panellists were retained for the analysis. The separation ratio (G) ranged from 0.00 to .94. The strata (H) index ranged from 0.33 to 1.59, and the reliability index ranged from 0.00 to 0.47. The Rasch Kappa index ranged from -0.01 to -0.04.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.00	0.00	0.82	0.94	0.00	0.00
Strata (H)	0.33	0.33	1.43	1.59	0.33	0.33
Reliability (R)	0.00	0.00	0.40	0.47	0.00	0.00
Obs. Agree (%)	14.1	13.2	11.8	9.1	15.0	14.5
Exp. agree (%)	17.6	16.9	12.5	11.2	17.7	17.0
Rasch – Kappa	-0.04	-0.04	-0.01	-0.02	-0.03	-0.03
Min. *Infit (ZStd)*	0.03 (1.7)	3.72 (2.0)	.13 (.6)	.07 (.6)	.01 (1.1)	.03 (1.0)
Max. *Infit (ZStd)*	2.82 (1.4)	.02 (1.8)	2.65 (1.5)	3.35 (1.6)	2.08 (1.2)	2.47 (1.2)
Mean *Infit*	1.00	.89	1.03	1.00	.91	.90
Mean *Infit SD*	1.35	1.18	0.83	1.00	.69	.78

Table 5.25: Rasch Interparticipant and interparticipant consistency: ISE III Speaking section

Listening Component

As has already been stated, the length of the listening scale is slightly different depending on the ISE level. The short (4-point scale) for ISE II and ISE III rendered the data as extreme, resulting in a recurring cycle of misfitting panellists. Therefore, no analysis was possible for these levels.

The initial analysis for ISE Foundation revealed one misfitting panellist (J08, Infit measure of 3.76 with a Zstd of 2.7). After this panellist was eliminated, none of the remaining panellists were misfitting, and the Infit values for the remaining panellists ranged from 0.30 (-1.1) to 1.83 (1.1). As a consequence of this exercise, the overall mean for the Pass result band for Round 2 dropped from 4.0 to 3.9. The standard deviation of the panellists remained the same. The SEj slightly increased from .27 to .28, and the SEj/SEM slightly increased from .18 to .19. The group's Alpha and ICC remained at .99. The separation G index was .00, strata was .33, and reliability was .00.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	1.31	0.32	0.00	0.00	-	-
Strata (H)	2.08	0.76	0.33	0.33	-	-
Reliability (R)	0.63	0.09	0.00	0.00	-	-
Obs. Agree (%)	60.1	61.2	73.7	81.8	-	-
Exp. agree (%)	62.1	64.0	75.2	83.0	-	-
Rasch – Kappa	-0.05	-0.08	-0.06	-0.07	-	-
Min. *Infit (ZStd)*	0.20 (-1.0)	0.29 (-1.0)	0.06 (0.6)	0.19 (0.7)	-	-
Max. *Infit (ZStd)*	6.18 (2.9)	3.76 (2.7)	1.78 (1.4)	1.61 (0.9)	-	-
Mean *Infit*	1.75	1.23	0.73	0.71	-	-
Mean *Infit SD*	1.69	0.93	0.59	0.53	-	-

Table 5.26: Rasch Interparticipant and interparticipant consistency: ISE Foundation Listening section

For ISE I no panellist exhibited misfit, and all 11 panellists were retained for the analysis. The separation ratio (G) was .00, the strata (H) index was .00, and the reliability index ranged from .00 in all rounds and levels. The Rasch Kappa index ranged from -.09 to -.01.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.00	0.00	0.00	0.00	0.00	0.00
Strata (H)	0.33	0.33	0.33	0.33	0.33	0.33
Reliability (R)	0.00	0.00	0.00	0.00	0.00	0.00
Obs. Agree (%)	70.1	71.1	85.2	80.2	96.9	92.5
Exp. agree (%)	72.5	73.4	86.2	81.3	97.0	92.6
Rasch – Kappa	-0.09	-0.09	-0.07	-0.06	-0.03	-0.01
Min. *Infit (ZStd)*	0.30 (-1.0)	0.44 (-0.2)	0.14 (0.6)	0.09 (0.6)	1.00 (0.9)	0.29 (0.6)
Max. *Infit (ZStd)*	2.33 (1.7)	2.86 (1.5)	3.02 (1.6)	4.05 (2.7)	1.00 (.9)	1.28 (0.8)
Mean *Infit*	1.17	1.33	1.51	1.44	1.00	0.62
Mean *Infit SD*	0.92	0.77	1.03	1.35	0.00	0.47

Table 5.27: Rasch Interparticipant and interparticipant consistency: ISE I Listening section

Reading Component

The initial analysis for ISE Foundation revealed no misfitting panellists, but the round 2 ratings for the Pass results band exhibited inconsistency within the method. As all panellists were within the acceptable Infit range, the group's judgments were trimmed by eliminating the lowest and highest grades. Panellists J08 and J10 were eliminated from the analysis, and all analyses were rerun. After this exercise, the round 2 mean score remained at 14.1. However, the standard deviation decreased from 3.6 to 2.8. The SEj decreased from 1.15 to .89. The criterion for consistency within the method (SEj/SEM) was met as SEj/SEM decreased from .89 to .43. The group's Alpha decreased slightly from .88 to .87. The ICC decreased from .87 to .85. The separation ratio (G) ranged from .00 to 1.65. The strata (H) index ranged from .33 to 2.53, and the reliability index ranged from .00 to .73. The Rasch Kappa index ranged from -.09 to -.06.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.95	1.65	1.16	0.49	0.39	0.00
Strata (H)	1.61	2.53	1.88	0.98	0.86	0.33
Reliability (R)	0.48	0.73	0.57	0.19	0.13	0.00
Obs. Agree (%)	61.6	69.4	65.6	75.1	91.1	92.7
Exp. agree (%)	65.2	72.0	68.7	77.4	91.6	93.3
Rasch – Kappa	-0.10	-0.09	-0.10	-0.10	-0.06	-0.09
Min. *Infit (ZStd)*	0.60 (-1.7)	1.36 (1.2)	0.69 (-1.5)	0.58 (-1.9	0.33 (-0.7)	0.57 (-1.6)
Max. *Infit (ZStd)*	1.63 (2.7)	1.39 (1.1)	1.40 (2.0)	1.49 (1.8)	1.35 (3.7)	1.34 (1.1)
Mean *Infit*	0.99	1.00	0.98	1.01	0.75	1.00
Mean *Infit SD*	0.33	0.25	0.25	0.28	0.44	0.29

Table 5.28: Rasch Interparticipant and interparticipant consistency: ISE Foundation Reading section

For ISE I, no panellist exhibited misfit, and all 11 panellists were retained for the analysis. The separation ratio (G) ranged from 0.00 to 0.94. The strata (H) index ranged from 0.33 to 1.54, and the reliability index ranged from 0.00 to 0.47. The Rasch Kappa index ranged from -0.09 to -0.06.

	Pass		**Merit**		**Distinction**	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.00	0.94	0.77	0.00	0.00	0.00
Strata (H)	0.33	1.54	1.36	0.33	0.33	0.33
Reliability (R)	0.00	0.47	0.37	0.00	0.00	0.00
Obs. Agree (%)	68.6	70.3	74.7	76.6	97.6	94.4
Exp. agree (%)	71.4	72.8	76.8	78.7	97.7	94.7
Rasch – Kappa	-0.10	-0.09	-0.09	-0.10	-0.04	-0.06
Min. *Infit (ZStd)*	.58 (-1.7)	.47 (-2.0)	.59 (-1.4)	.67 (-1.1)	1.00 (.1)	.34 (-1.2)
Max. *Infit (ZStd)*	1.47 (1.5)	1.56 (1.6)	1.45 (2.1)	.1.21 (0.9)	1.00 (0.0)	2.70 (2.1)
Mean *Infit*	1.00	1.01	0.97	0.98	1.00	1.00
Mean *Infit SD*	0.25	0.35	0.25	0.19	0.00	0.87

Table 5.29: Rasch Interparticipant and interparticipant consistency: ISE I Reading section

As with ISE I, no panellist exhibited misfit for ISE II. The separation ratio (G) ranged from .00 to 1.23. The strata (H) index ranged from 0.33 to 1.79, and the reliability index ranged from 0.00 to 0.60. The Rasch Kappa index ranged from -0.09 to -0.06.

	Pass		**Merit**		**Distinction**	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.65	0.85	1.09	1.23	0.00	0.00
Strata (H)	1.20	1.46	1.79	1.98	0.33	0.33
Reliability (R)	0.30	0.42	0.54	0.60	0.00	0.00
Obs. Agree (%)	57.3	61.9	68.4	74.7	89.9	93.0
Exp. agree (%)	61.0	65.2	71.0	76.7	90.7	93.4
Rasch – Kappa	-0.09	-0.09	-0.09	-0.09	-0.09	-0.06
Min. *Infit (ZStd)*	0.76 (-1.14)	0.74 (1.4)	0.63 (-0.5)	0.35 (-0.7)	0.86 (-0.6)	0.87 (-0.9)
Max. *Infit (ZStd*	1.17 (1.1)	1.29 (1.03)	1.46 (2.4)	1.62 (1.2)	1.17 (.8)	1.12 (.4)
Mean *Infit*	1.01	1.01	1.00	0.98	1.00	1.01
Mean *Infit SD*	0.15	0.20	0.23	0.30	0.11	0.10

Table 5.30: Rasch Interparticipant and interparticipant consistency: ISE II Reading section

Once again, no panellist exhibited misfit for ISE III, and all 11 panellists were retained for the analysis. The separation ratio (G) ranged from 0.00 to 0.81. The strata (H) index ranged from 0.33 to 1.41, and the reliability index ranged from 0.00 to 0.40. The Rasch Kappa index ranged from -0.09 to -0.07.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.81	0.00	0.00	0.00	0.00	0.00
Strata (H)	1.41	0.33	0.33	0.33	0.33	0.33
Reliability (R)	0.40	0.00	0.00	0.00	0.00	0.00
Obs. Agree (%)	56.7	60.7	68.8	73.0	92.0	93.6
Exp. agree (%)	60.4	64.2	71.6	75.3	92.5	94.0
Rasch – Kappa	-0.09	-0.10	-0.10	-0.09	-0.07	-0.07
Min. Infit (ZStd)	.76 (-1.2)	.82 (-1.0)	.81 (-.7)	.80 (-.5)	.44 (-1.8)	.39 (-1.8)
Max. Infit (ZStd)	1.23 (1.12)	1.44 (2.1)	1.26 (1.12)	1.41 (1.7)	1.50 (1.5)	1.60 (1.3)
Mean Infit	1.00	1.01	0.99	1.00	1.00	1.01
Mean Infit SD	0.15	0.21	0.12	0.20	0.32	0.40

Table 5.31: Rasch Interparticipant and interparticipant consistency: ISE IIII Reading section

Writing Component

The initial analysis for ISE Foundation revealed no misfitting panellists. The separation ratio (G) ranged from 0.00 to 0.22. The stratum (H) index ranged from 0.33 to 0.62, and the reliability index ranged from 0.00 to 0.05. The Rasch Kappa index ranged from -0.03 to -0.02.

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.00	0.00	0.00	0.00	0.00	0.22
Strata (H)	0.33	0.33	0.33	0.33	0.33	0.62
Reliability (R)	0.00	0.00	0.00	0.00	0.00	0.05
Obs. Agree (%)	44.1	43.5	26.3	27.0	21.6	11.1
Exp. agree (%)	45.8	45.2	28.2	28.5	23.8	13.8
Rasch – Kappa	-0.03	-0.03	-0.03	-0.02	-0.03	-0.03
Min. *Infit (ZStd)*	0.03 (0.4)	0.00 (0.48)	.28 (1.0)	0.45 (0.3)	.37 (1.3)	.30 (1.6)
Max. *Infit (ZStd)*	6.99 (1.9)	2.65 (1.4)	4.44 (1.6)	1.73 (0.9)	2.29 (1.1)	1.47 (0.8)
Mean *Infit*	1.09	0.98	1.01	0.93	0.87	0.90
Mean *Infit SD*	1.99	0.70	1.18	0.39	0.70	0.43

Table 5.32: Rasch Interparticipant and interparticipant consistency: ISE Foundation Writing section

As with ISE Foundation, no panellist exhibited misfit for ISE I and all 11 panellists were retained for the analysis. The separation ratio (G) was 0.00, the strata (H) index was 0.00, and the reliability index ranged from 0.00 in all rounds and levels. The Rasch Kappa index ranged from -.05 to .04.

	Pass		**Merit**		**Distinction**	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.00	0.00	0.00	0.00	0.00	0.00
Strata (H)	0.33	0.33	0.33	0.33	0.33	0.33
Reliability (R)	0.00	0.00	0.00	0.00	0.00	0.00
Obs. Agree (%)	41.0	41.8	16.9	12.5	14.5	15.6
Exp. agree (%)	38.3	40.5	18.9	16.4	18.4	18.9
Rasch – Kappa	0.04	0.02	-0.02	-0.05	-0.05	-.04
Min. *Infit (ZStd)*	.11 (1.6)	0.08 (1.3)	.24 (.7)	.35 (.7)	.22 (.4)	0.08 (0.4)
Max. *Infit (ZStd)*	3.86 (1.5)	4.10 (1.6)	2.74 (1.4)	3.74 (1.5)	3.49 (1.4)	3.51 (1.4)
Mean *Infit*	0.99	0.93	1.02	1.10	1.03	1.04
Mean *Infit SD*	1.08	1.16	0.86	0.92	0.86	0.92

Table 5.33: Rasch Interparticipant and interparticipant consistency: ISE I Writing section

As with the lower ISE levels, no panellist exhibited misfit, and all 11 panellists were retained for the analysis. The separation ratio (G) was 0.00, the strata (H) index was 0.00, and the reliability index ranged from 0.00 in all rounds and levels. The Rasch Kappa index ranged from -.04 to -.01.

	Pass		**Merit**		**Distinction**	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.00	0.00	0.00	0.00	0.00	0.00
Strata (H)	0.33	0.33	0.33	0.33	0.33	0.33
Reliability (R)	0.00	0.00	0.00	0.00	0.00	0.00
Obs. Agree (%)	34.5	31.2	13.8	12.5	18.7	17.1
Exp. agree (%)	35.4	33.8	16.5	15.6	19.9	18.6
Rasch – Kappa	-0.01	-0.04	-0.03	-0.04	-0.01	-0.02
Min. *Infit (ZStd)*	0.00 (5.0)	0.04 (1.3)	0.33 (0.0)	0.08 (0.2)	0.22 (0.8)	0.22 (1.1)
Max. *Infit (ZStd)*	1.74 (.9)	3.30 (1.4)	4.42 (1.6)	3.91 (1.5)	4.51 (1.6)	4.17 (1.5)
Mean *Infit*	0.74	0.83	1.03	1.01	0.94	0.91
Mean *Infit SD*	0.61	0.96	1.11	1.02	1.18	1.07

Table 5.34: Rasch Interparticipant and interparticipant consistency: ISE II Writing section

As with all previous ISE levels for this component, no panellist exhibited misfit, and all 11 panellists were retained for the analysis. The separation ratio (G) ranged from 0.00 to 0.53. The strata (H) index ranged from 0.33 to 1.04, and the reliability index ranged from 0.00 to 0.22. The Rasch Kappa index ranged from -0.03 to 0.00.

	Pass		**Merit**		**Distinction**	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Separation (G)	0.28	0.53	0.00	0.00	0.00	0.00
Strata (H)	0.71	1.04	0.33	0.33	0.33	0.33
Reliability (R)	0.07	0.22	0.00	0.00	0.00	0.00
Obs. Agree (%)	13.2	15.3	10.1	11.4	14.0	16.4
Exp. agree (%)	14.2	16.6	11.1	11.6	16.2	17.2
Rasch – Kappa	-0.01	-0.02	-0.01	0.00	-0.03	-0.01
Min. *Infit (ZStd)*	0.11 (0.2)	0.17 (0.7)	.03 (-.2)	0.04 (-0.1)	0.19 (0.7)	0.19 (0.7)
Max. *Infit (ZStd)*	3.41 (1.4)	2.31 (1.1)	2.78 (1.3)	2.19 (1.1)	3.03 (1.7)	2.65 (1.2)
Mean *Infit*	0.97	0.81	0.96	0.95	0.94	0.88
Mean *Infit SD*	1.15	0.69	0.81	0.68	0.82	0.76

Table 5.35: Rasch Interparticipant and interparticipant consistency: ISE III Writing section

5.5 CONSISTENCY WITHIN THE METHOD

Hambleton and Pitoniak (2006, p. 458) define consistency within the method as "the extent to which same performance standards would be obtained if the method were replicated". In this study, method consistency was examined by estimating the standard error of the cut score (SEj). The equation used to calculate the standard error of the cut score is the following:

$$SEj = \frac{\text{SDs}}{\sqrt{n}}$$

The standard error of judgment (SE_j) is equal to the standard deviation of the individual cut scores (SDs) divided by the square root of the number of panellists (Cizek & Bunch, 2007). The SEj is "one of the classical indices indicative of replicability of the obtained results" (Kaftandjieva, 2010, p. 103) and is compared to the standard error of measurement (SEM) of the test. Several criteria have been suggested when comparing the SEj to the SEM. Jaeger (1991) suggests that the SE should be no greater than one-quarter of the SEM, while Cohen, Kane, and Crooks state that the SE should not be greater than half the SEM to "have relatively little impact on the misclassification rates" (1999, p. 364). On the other hand, Kaftandjieva (2010) recommends a compromise between the previous two criteria and suggests that the SEj should be no greater than a third of the SEM.

In this study, the criterion used for evaluating the recommended cut scores was the criterion proposed by Cohen, Kane, and Crooks, whereby SEj/SEM ≤ 0.50 for Round 2 measures. The following tables present the standard error of the cut score and the standard error of measurement for each component (Speaking, Listening, Reading, Writing) by ISE level.

Speaking component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.26	0.11	0.30	0.30	0.19	0.18
SEM	0.86	0.86	0.86	0.86	0.86	0.86
SEj/SEM	0.33	0.14	0.37	0.37	0.25	0.23

Table 5.36: Standard error of cut scores and measurement: ISE Foundation Speaking component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.35	0.37	0.30	0.26	0.18	0.17
SEM	0.96	0.96	0.96	0.96	0.96	0.96
SEj/SEM	0.36	0.38	0.31	0.27	0.17	0.17

Table 5.37: Standard error of cut scores and measurement: ISE I Speaking component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.38	0.36	0.34	0.28	0.20	0.20
SEM	1.00	1.00	1.00	1.00	1.00	1.00
SEj/SEM	0.38	0.36	0.34	0.28	0.20	0.20

Table 5.38: Standard error of cut scores and measurement: ISE II Speaking component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.32	0.32	0.39	0.40	0.19	0.19
SEM	1.08	1.08	1.08	1.08	1.08	1.08
SEj/SEM	0.30	0.30	0.36	0.37	0.18	0.18

Table 5.39: Standard error of cut scores and measurement: ISE III Speaking component

In line with the criterion proposed by Cohen, Kane, and Crooks, whereby SEj/SEM ≤ 0.50 is considered acceptable for Round 2 estimates, the Speaking component cut scores across all ISE levels met the threshold across all performance bands (Pass, Merit, Distinction). As shown in Tables 5.36 to 5.39, the SEj/SEM ratios remained well below the 0.50 benchmark in both Rounds 1 and 2, indicating stable and precise cut score estimates and supporting the reliability of the standard setting outcomes.

Listening component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.39	0.28	0.31	0.24	0.06	0.06
SEM	0.96	0.96	0.96	0.96	0.96	0.96
SEj/SEM	0.41	0.30	0.32	0.25	0.07	0.06

Table 5.40: Standard error of cut scores and measurement: ISE Foundation Listening component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.21	0.31	0.25	0.34	0.15	.17
SEM	1.14	1.14	1.14	1.14	1.14	1.14
SEj/SEM	0.18	0.27	0.22	0.30	0.13	0.15

Table 5.41: Standard error of cut scores and measurement: ISE I Listening component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.09	0.09	0.05	0.04	0.05	0.05
SEM	0.71	0.71	0.71	0.71	0.71	0.71
SEj/SEM	0.13	0.13	0.07	0.06	0.07	0.07

Table 5.42: Standard error of cut scores and measurement: ISE II Listening component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.13	0.12	0.07	0.07	0.05	0.05
SEM	0.71	0.71	0.71	0.71	0.71	0.71
SEj/SEM	0.19	0.16	0.09	0.10	0.07	0.07

Table 5.43: Standard error of cut scores and measurement: ISE III Listening component

All bands for all ISE levels also satisfied the SEj/SEM ≤ 0.50 criterion in both rounds. The ratios were consistently low, across all levels and bands, suggesting high confidence in cut score placement for this component across all levels and bands.

Reading component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	1.06	0.87	1.02	0.62	0.74	0.39
SEM	2.10	2.10	2.10	2.10	2.10	2.10
SEj/SEM	0.50	0.42	0.49	0.30	0.35	0.19

Table 5.44: Standard error of cut scores and measurement: ISE Foundation Reading component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.54	0.72	0.72	0.34	0.19	0.36
SEM	1.61	1.61	1.61	1.61	1.61	1.61
SEj/SEM	0.34	0.45	0.45	0.21	0.12	0.22

Table 5.45: Standard error of cut scores and measurement: ISE I Reading component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.86	0.88	0.86	0.71	0.59	0.47
SEM	2.10	2.10	2.10	2.10	2.10	2.10
SEj/SEM	0.41	0.42	0.41	0.34	0.28	0.22

Table 5.46: Standard error of cut scores and measurement: ISE II Reading component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.92	0.65	0.54	0.49	.43	.36
SEM	2.07	2.07	2.07	2.07	2.07	2.07
SEj/SEM	0.45	0.32	0.26	0.24	.20	.17

Table 5.47: Standard error of cut scores and measurement: ISE III Reading component

The criterion was met for every band and level (ISE Foundation to ISE III) in both rounds, reinforcing the reliability of the locations of the cut scores at all proficiency levels.

Writing component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.44	0.33	0.42	0.42	0.50	.66
SEM	1.39	1.39	1.39	1.39	1.39	1.39
SEj/SEM	0.32	0.24	0.31	0.31	0.36	0.48

Table 5.48: Standard error of cut scores and measurement: ISE Foundation Writing component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.21	0.22	0.51	0.53	0.41	0.36
SEM	1.36	1.36	1.36	1.36	1.36	1.36
SEj/SEM	0.16	0.16	0.38	0.39	0.30	0.26

Table 5.49: Standard error of cut scores and measurement: ISE I Writing component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.44	0.42	0.47	0.33	.32	0.31
SEM	1.45	1.45	1.45	1.45	1.45	1.45
SEj/SEM	0.30	0.29	0.32	0.23	0.22	0.21

Table 5.50: Standard error of cut scores and measurement: ISE II Writing component

	Pass		Merit		Distinction	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
SEj	0.63	0.55	0.53	.53	0.31	0.31
SEM	1.33	1.33	1.33	1.33	1.33	1.33
SEj/SEM	0.47	0.42	0.40	0.40	0.24	0.24

Table 5.51: Standard error of cut scores and measurement: ISE III Writing component

All ISE levels and result bands achieved SEj/SEM ≤ 0.50 in both rounds. Some moderate variation was observed in SEj values, but in no case did the SEj/SEM ratio exceed the threshold, confirming acceptable precision of the cut scores thus confirming the reliability of the cut score placement.

Overall Conclusion

Across all components and ISE levels, the recommended cut scores demonstrated sufficient measurement precision, as all SEj/SEM ratios remained within the acceptable limit of 0.50 in both standard setting rounds. This supports the robustness and defensibility of the cut score recommendations, providing confidence in their consistent application in operational settings.

5.6 DECISION CONSISTENCY & ACCURACY

Decision consistency refers to the agreement between the classifications of the same candidates on two different examinations with the same test" (Kaftandjieva, 2004, p. 26). To compute such coefficients, candidates would have to take the same examination twice, which is not typically feasible. Some methods (Livingston & Lewis, 1995; Subkoviak, 1988) estimate decision consistency and accuracy based on a single administration. These methods provide the "likelihood that an examinee classified as passing (or failing) on one administration of an examination will be classified similarly on a second administration" (Cizek & Bunch, 2007, p. 309).

This study employed the Livingston and Lewis method, which is "a generally applicable method for using data from one form of a test to estimate the accuracy and consistency of classifications based on the scores" (Livingston & Lewis, 1995, p. 179). The Livingston and Lewis decision consistency and accuracy estimates were obtained using the BB–Class software (Brennan, 2001). The four-parameter beta-binomial model was selected for analysis. The method produces candidate classification estimates that "tend to be within one percentage point of their actual values" (Livingston & Lewis, 1995, p. 196).

The tables that follow illustrate the operational cut scores for each ISE level by component and results band with their corresponding decision accuracy. The first row of each table shows the probability of correct classification for each raw cut score, and the next two rows show the probability of false-positive and false-negative errors. False-positive errors occur when candidates are estimated to be above the cut score when, in fact, they are not. Similarly, false-negative errors occur when candidates are estimated to be below the cut score when, in fact, they are not (Hambleton & Novick, 1973).

ISE Foundation

Table 5.52 presents the classification accuracy statistics for the ISE Foundation test across the four components: Speaking, Listening, Reading, and Writing. For each component, the probability of correct classification is consistently high across the Pass, Merit, and Distinction thresholds. Specifically, correct classification probabilities range from 0.86 to 0.89 for Speaking, 0.84 to 0.88 for Listening, 0.70 to 0.99 for Reading, and 0.87 to 0.90 for Writing. Corresponding false-positive rates range from less than 0.005 to 0.13, while false-negative rates range from less than 0.005 to 0.19.

At the critical Pass threshold, classification accuracy is particularly strong, with probabilities of correct classification of 0.86 for Speaking, 0.88 for Listening, 0.99 for Reading, and 0.87 for Writing. False-positive rates at the Pass level range from <0.005 in Reading to 0.13 in Speaking and Writing, while false-negative rates at Pass are minimal—<0.005 in Reading and Writing, 0.01 in Speaking, and 0.07 in Listening. These results are in line with Subkoviak's (1980, 1988) guidelines, which indicate a low likelihood of misclassification at the Pass level and provide strong support for the validity and reliability of the cut scores when applied to live operational data.

SPEAKING	**Pass**	**Merit**	**Distinction**
	8	**12**	**15**
Probability of correct classification	0.86	0.86	0.89
False-positive	0.13	0.13	0.09
False-negative	0.01	0.01	0.01

LISTENING	**Pass**	**Merit**	**Distinction**
	3	**5**	**7**
Probability of correct classification	0.88	0.84	0.85
False-positive	0.05	0.09	0.10
False-negative	0.07	0.07	0.05

READING	**Pass**	**Merit**	**Distinction**
	15	**23**	**28**
Probability of correct classification	0.85	0.81	0.94
False-positive	0.14	0.17	0.06
False-negative	0.01	0.02	0.00

WRITING	**Pass**	**Merit**	**Distinction**
	14	**20**	**25**
Probability of correct classification	0.87	0.87	0.90
False-positive	0.13	0.12	0.09
False-negative	<0.005	0.01	0.01

Table 5.52: Accuracy relative to observed scores: ISE Foundation

ISE I

Table 5.53 presents the classification accuracy statistics for the ISE I test across the four components: Speaking, Listening, Reading, and Writing. The probability of correct classification across components ranges from 0.71 to 0.99, indicating a generally strong level of decision accuracy (Subkoviak, 1980, 1988). Speaking and Writing show particularly high classification accuracy, with probabilities ranging from 0.92 to 0.99 across all grade bands. Reading also performs strongly, with correct classification ranging from 0.81 to 0.99. Listening shows slightly lower accuracy at the Pass threshold (0.71), but this remains within acceptable bounds, given the relatively short length of the Listening component.

False-positive rates range from less than 0.005 to 0.27, while false-negative rates remain low across all components, ranging from less than 0.005 to 0.06. At the Pass level, correct classification is consistently high for Speaking (0.92), Reading (0.99), and Writing (0.90). Although Listening shows more variability at the Pass level, the accuracy remains fit for purpose within the broader assessment framework.

These findings support the reliability and validity of the ISE I cut scores when applied to live operational data, particularly for the Speaking, Reading, and Writing components. For Listening, the results are acceptable, given the test design.

SPEAKING	**Pass**	**Merit**	**Distinction**
	8	**12**	**15**
Probability of correct classification	0.92	0.93	0.95
False-positive	0.05	0.04	0.04
False-negative	0.03	0.03	0.01

LISTENING	**Pass**	**Merit**	**Distinction**
	3	**5**	**7**
Probability of correct classification	0.71	0.92	0.82
False-positive	0.27	0.06	0.18
False-negative	0.03	0.02	0.00

READING	**Pass**	**Merit**	**Distinction**
	15	**23**	**28**
Probability of correct classification	0.99	0.93	0.81
False-positive	0.01	0.05	0.13
False-negative	<0.005	0.03	0.06

WRITING	**Pass**	**Merit**	**Distinction**
	14	**20**	**25**
Probability of correct classification	0.90	0.94	0.99
False-positive	0.03	0.02	< 0.005
False-negative	0.06	0.04	< 0.005

Table 5.53: Accuracy relative to observed scores: ISE I

ISE II

Table 5.54 presents the classification accuracy statistics for the ISE II test across the four components: Speaking, Listening, Reading, and Writing. The probability of correct classification ranges from 0.81 to 0.99, indicating a consistently high level of classification accuracy across all domains (Subkoviak, 1980, 1988).

False-positive rates range from 0.01 to 0.17, and false-negative rates remain low, between less than 0.005 and 0.04. At the Pass threshold, correct classification rates are 0.91 for Speaking, 0.90 for Writing, 0.85 for Reading, and 0.81 for Listening. These results reflect stable classification performance at the key decision points, with a low risk of misclassification across all components.

Together, the findings provide strong evidence for the validity and consistency of the ISE II cut scores, supporting their continued use in operational testing contexts.

SPEAKING	**Pass**	**Merit**	**Distinction**
	8	**12**	**15**
Probability of correct classification	0.91	0.93	0.96
False-positive	0.06	0.04	0.03
False-negative	0.03	0.03	0.01

LISTENING	**Pass**	**Merit**	**Distinction**
	3	**5**	**7**
Probability of correct classification	0.81	0.90	0.90
False-positive	0.16	0.08	0.10
False-negative	0.03	0.02	0.00

READING	**Pass**	**Merit**	**Distinction**
	15	**23**	**28**
Probability of correct classification	0.85	0.81	0.94
False-positive	0.14	0.17	0.06
False-negative	0.01	0.02	0.00

WRITING	**Pass**	**Merit**	**Distinction**
	14	**20**	**25**
Probability of correct classification	0.90	0.96	0.99
False-positive	0.06	0.02	0.01
False-negative	0.04	0.02	<0.005

Table 5.54: Accuracy relative to observed scores: ISE II

ISE III

Table 5.51 presents the classification accuracy statistics for the ISE III test across the four components: Speaking, Listening, Reading, and Writing. The probability of correct classification ranges from 0.76 to 0.99, with consistently strong performance across components, particularly at the Pass threshold, where accuracy is highest and most consequential for test-takers. According to Subkoviak's interpretive guidelines (1980, 1988), probabilities above 0.80 represent strong decision consistency, and all components meet or exceed this threshold at the Pass level. At the Pass level, classification accuracy is notably high: 0.91 for Speaking, 0.84 for Writing, 0.82 for Reading, and 0.80 for Listening.

False-positive rates range from 0.01 to 0.23, and false-negative rates remain low throughout, ranging from less than 0.005 to 0.03. These results indicate reliable decision-making at the key cut score, with minimal risk of misclassification.

Overall, these findings provide strong evidence for the reliability and validity of the ISE III cut scores. Consistent classification performance at the Pass threshold, aligned with recognised benchmarks for criterion-referenced test decisions, supports their continued use in operational testing contexts.

SPEAKING	**Pass**	**Merit**	**Distinction**
	8	**12**	**15**
Probability of correct classification	0.91	0.93	0.95
False-positive	0.06	0.04	0.04
False-negative	0.03	0.03	0.01

LISTENING	**Pass**	**Merit**	**Distinction**
	3	**5**	**7**
Probability of correct classification	0.80	0.91	0.93
False-positive	0.17	0.07	0.07
False-negative	0.03	0.02	0.00

READING	**Pass**	**Merit**	**Distinction**
	16	**23**	**28**
Probability of correct classification	0.82	0.76	0.88
False-positive	0.18	0.23	0.12
False-negative	< 0.005	0.01	0.01

WRITING	**Pass**	**Merit**	**Distinction**
	15	**20**	**25**
Probability of correct classification	0.84	0.95	0.99
False-positive	0.16	0.05	0.01
False-negative	<0.005	<0.005	<0.005

Table 5.55: Accuracy relative to observed scores: ISE III

6. Conclusion

This study undertook a comprehensive validation of the cut scores for the revised ISE test suite (ISE Foundation to ISE III), following the three-part framework for standard setting validation proposed by Hambleton and Pitoniak (2006): procedural, internal, and external validity.

Procedural validity was positively demonstrated through the structured implementation of the standard setting process. Clear documentation, panel training, adherence to best practices, and structured feedback loops ensured that the methods were transparent, replicable, and appropriate for the high-stakes decisions involved. Panel composition was balanced and representative, and multiple judgement rounds were used to promote informed reflection and convergence of scores.

Internal validity was supported through several lines of evidence. First, consistency within the method was confirmed by low SEj/SEM ratios, indicating strong replicability of cut score judgments. Second, intra-participant consistency was evidenced by small and appropriate judgement shifts between rounds, suggesting meaningful engagement with empirical feedback. Third, inter-participant consistency was excellent across all ISE levels, as demonstrated by high Cronbach's alpha and intraclass correlation coefficients, supported further by Rasch-based agreement statistics. Lastly, decision consistency and accuracy, estimated using the Livingston and Lewis method and interpreted using Subkoviak's (1980, 1988) criteria, showed that classification reliability was high across components and especially strong at the critical Pass threshold. False-positive and false-negative rates remained consistently low. External validity is outside the scope of this report; it will be addressed in future work to strengthen further the interpretive claims associated with the cut scores.

In sum, this study presents strong and converging validity evidence for the cut scores established in 2015 for all ISE levels. The results confirm that the scores are defensible, consistent, and well aligned with the intended performance standards. These findings support the operational use of the cut scores in high-stakes assessment contexts and provide confidence in their continued application for CEFR-referenced reporting.

References

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (Version 1.0)

Cizek, G. J. (2012). An Introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, innovations* (pp. 3-14). New York, NY: Routledge.

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

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

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg, France: Council of Europe, Modern Languages Division & Cambridge University Press.

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: A Manual*. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d.

Impara, J. C., & Blake, B. (1997). Standard Setting: An alternative approach. *Journal of Educational Measurement*, 34(4), 353-366.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1), 69-81.

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, The Netherlands: Cito.

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.

Longford, N. (1996). Reconciling experts' differences in setting cut scores for pass-fail decisions. *Journal of Educational and Behavioral Statistics*, 21(3), 203-213. doi:10.3102/10769986021003203

Papageorgiou, S. (2007). Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference: [online report] https://www.trinitycollege.com/about-us/recognition/english-language/cefr-alignment

Papageorgiou, S. (2009). *Setting performance standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*. Frankfurt, Germany: Peter Lang GmbH.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, 27(2), 261-282. doi:10.1177/0265532209349472

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The Modified Angoff, Extended Angoff and Yes/No standard setting methods. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods and Innovations* (pp.118-201). New York & London, NY & United Kingdom: Routledge.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing, *Language Testing*, 72 (4), 72-92. doi:10.1177/026553228700400107

Putnam, S. E., Pence, P., & Jaeger, R. M. (1995). A Multi-Stage Dominant Profile Method for setting standards on complex performance assessments. *Applied Measurement in Education*, 8(1), 57-83.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based tests: In pursuit of improved construct representation. In S. G. Sireci, A. L. Zenisky, S. M. Downing, & T. M. Haldyna (Eds.), *Handbook of Test Development* (pp. 329-348). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Subkoviak, M. J. (1980). Consistency of decisions based on test scores: A review of the reliability of criterion-referenced tests. *Journal of Educational Measurement, 17*(3), 221–230.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47–55. Tannenbaum, I. R. (2014). Comparison of web-based and face-to-face standard setting using the Angoff method. *Journal of Applied Testing Technology*, 15(1), 1-17.

Zieky, M. J., Perie, M., & Livingston, S. A. (2012). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Lexington, KY: ETS.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: ETS. Retrieved from https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

Annex

ANNEX A CEFR DESCRIPTORS: ITEM MEASUREMENT (ALL DESCRIPTORS)

Descriptors	CEFR	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
1 S01	C2	(10.52	1.85)	Maximum			
77 P03	C2	9.25	1.05	.89	.1	.59	-.1
95 G01	C2	9.25	1.05	.89	.1	.59	-.1
117 G23	C2	8.45	.78	.92	.0	.98	.1
56 L01	C2	7.94	.66	1.05	.2	1.05	.2
23 S23	C2	7.55	.60	1.28	.7	1.12	.4
69 L14	C1	7.55	.60	.90	-.1	.89	-.1
110 G16	C2	7.55	.60	.88	-.1	.85	-.2
88 P14	C1	7.21	.56	.68	-.8	.67	-.8
43 W13	C1	6.91	.54	.76	-.5	.79	-.5
28 S28	C2	6.63	.52	1.31	.8	1.31	.9
32 W02	C1	6.63	.52	.67	-.8	.69	-.8
40 W10	C2	6.63	.52	.55	-1.3	.55	-1.3
50 W20	C1	6.36	.51	.54	-1.3	.53	-1.3
84 P10	C1	6.36	.51	.84	-.3	.83	-.3
16 S16	C2	6.11	.50	.43	-1.7	.37	-2.0
39 W09	C2	5.86	.50	2.80	3.2	2.90	3.2
33 W03	C2	5.62	.49	1.95	1.9	2.13	2.2
75 P01	C1	5.62	.49	.46	-1.5	.55	-1.1
94 P20	B2	5.62	.49	1.52	1.2	1.52	1.2
119 G25	C1	5.62	.49	.97	.0	1.00	.1
64 L09	C1	5.61	.51	.54	-1.1	.60	-.9
99 G05	C1	5.14	.49	.81	-.3	.82	-.3
31 W01	C2	4.90	.49	1.25	.6	1.32	.8
48 W18	C2	4.90	.49	3.92	4.3	4.11	4.5
51 W21	C1	4.90	.49	.71	-.6	.69	-.7
104 G10	C2	4.90	.49	.35	-1.9	.37	-1.9
118 G24	C1	4.90	.49	.52	-1.3	.49	-1.4
15 S15	C1	4.65	.49	1.82	1.7	1.92	1.9
24 S24	C1	4.65	.49	1.32	.8	1.29	.7
25 S25	C1	4.65	.49	.38	-1.8	.39	-1.8
53 W23	C2	4.41	.50	1.66	1.4	1.67	1.5
92 P18	C1	4.41	.50	1.44	1.0	1.42	1.0
100 G06	B2	4.41	.50	.47	-1.4	.48	-1.4
109 G15	C1	4.41	.50	.48	-1.4	.48	-1.4
115 G21	B2	4.41	.50	.57	-1.1	.56	-1.1
9 S09	C1	4.16	.50	1.42	1.0	1.43	1.0
44 W14	C2	4.16	.50	1.00	.1	1.00	.1
10 S10	C2	3.91	.50	1.49	1.1	1.42	1.0
72 L17	B2	3.91	.50	.64	-.8	.67	-.7
76 P02	B2	3.91	.50	.96	.0	.96	.0
83 P09	C1	3.91	.50	.41	-1.6	.41	-1.6
65 L10	C1	3.65	.51	1.03	.2	1.02	.1
68 L13	B2	3.65	.51	.98	.0	.98	.0
87 P13	B2	3.65	.51	1.38	.9	1.42	.9
73 L18	C1	3.39	.51	1.44	1.0	1.53	1.1
74 L19	B2	3.39	.51	.74	-.4	.72	-.5
122 G28	B2	3.13	.52	.19	-2.6	.19	-2.6
47 W17	C2	2.86	.52	.98	.0	1.04	.2
57 L02	B2	2.86	.52	1.18	.5	1.14	.4
63 L08	B2	2.86	.52	.64	-.7	.69	-.6
101 G07	B2	2.86	.52	.57	-.9	.65	-.7
35 W05	B2	2.58	.52	.87	-.1	.87	-.1

Descriptors	CEFR	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
7 S07	B2	2.03	.53	.47	-1.5	.48	-1.4
21 S21	B2	2.03	.53	.46	-1.5	.46	-1.5
52 W22	B2	2.03	.53	.42	-1.7	.43	-1.6
108 G14	B2	2.03	.53	.90	-.1	.82	-.3
6 S06	B2	1.75	.53	1.23	.6	1.23	.6
37 W07	B2	1.75	.53	.55	-1.2	.56	-1.2
103 G09	C1	1.75	.53	1.16	.5	1.18	.5
5 S05	C1	1.47	.54	.73	-.6	.73	-.6
14 S14	B2	1.17	.55	.97	.0	1.03	.2
62 L07	B2	1.17	.55	.88	-.2	.85	-.2
102 G08	B1	.17	.62	.52	-1.0	.51	-.9
18 S18	B1	-.23	.65	.28	-1.7	.25	-1.6
38 W08	B2	-.23	.65	.98	.1	.93	.0
41 W11	B1	-.23	.65	1.47	.9	1.50	.9
26 S26	B1	-.67	.67	.67	-.4	.58	-.5
79 P05	B1	-.67	.67	1.15	.4	1.19	.4
8 S08	B1	-1.12	.68	.14	-2.2	.10	-2.4
60 L05	B1	-1.12	.68	1.67	1.1	1.64	1.0
66 L11	B1	-1.12	.68	.14	-2.2	.10	-2.4
112 G18	B1	-1.12	.68	.14	-2.2	.10	-2.4
27 S27	B1	-1.58	.66	.50	-.9	.52	-.8
91 P17	B1	-1.58	.66	.38	-1.4	.33	-1.5
96 G02	B1	-1.58	.66	.85	-.1	.77	-.2
120 G26	B1	-1.58	.66	1.51	1.0	1.47	.9
121 G27	B1	-1.58	.66	.50	-.9	.52	-.8
45 W15	B1	-2.00	.64	1.00	.1	.94	.0
85 P11	B1	-2.40	.62	.75	-.6	.70	-.6
82 P08	B1	-3.14	.60	.78	-.6	.72	-.6
42 W12	B1	-3.34	.64	.98	.0	1.01	.1
11 S11	A2	-3.50	.61	.74	-.6	.69	-.7
46 W16	B1	-3.50	.61	1.98	2.3	1.79	1.7
17 S17	A2	-3.88	.62	.61	-.9	.54	-1.1
113 G19	B1	-3.88	.62	1.32	.8	1.31	.7
81 P07	B1	-4.27	.64	.69	-.5	.62	-.7
116 G22	A2	-4.27	.64	.45	-1.3	.36	-1.5
86 P12	A2	-4.68	.65	1.70	1.2	1.69	1.2
105 G11	A2	-4.68	.65	.17	-2.4	.15	-2.4
2 S02	B1	-5.12	.66	.11	-2.7	.10	-2.6
19 S19	B1	-5.12	.66	.63	-.6	.53	-.8
49 W19	A2	-5.12	.66	8.69	6.1	7.59	5.2
70 L15	A2	-5.12	.66	1.54	1.0	1.48	.9
89 P15	A2	-5.12	.66	.88	.0	.94	.0
98 G04	A2	-5.12	.66	.59	-.7	.56	-.7
34 W04	A2	-5.55	.65	1.45	.9	1.54	1.0
97 G03	A1	-5.55	.65	.75	-.3	.65	-.5
123 G29	A2	-5.55	.65	.39	-1.4	.29	-1.6
4 S04	A2	-5.97	.64	1.76	1.5	1.66	1.2
29 S29	A2	-5.97	.64	.53	-1.1	.48	-1.1
71 L16	A2	-5.97	.64	.95	.0	.98	.1
13 S13	A1	-6.37	.63	.98	.0	1.03	.2
36 W06	A2	-6.37	.63	.66	-.8	.59	-1.0
67 L12	A2	-6.37	.63	.65	-.9	.59	-1.0
59 L04	A1	-6.76	.62	1.25	.8	1.31	.8
78 P04	A1	-6.76	.62	.86	-.3	.85	-.3
111 G17	A2	-6.76	.62	.83	-.4	.81	-.4
20 S20	A1	-7.15	.63	1.23	.8	1.25	.7
22 S22	A2	-7.15	.63	.91	-.2	.88	-.2

		Model	Infit		Outfit	
Descriptors	CEFR	Measure	S.E.	MnSq	ZStd	MnSq	ZStd
54 W24	A1	-7.15	.63	.85	-.4	.79	-.5
114 G20	A1	-7.15	.63	.75	-.7	.70	-.8
12 S12	A1	-7.55	.64	.79	-.6	.73	-.6
30 S30	A1	-7.55	.64	1.16	.5	1.24	.7
90 P16	A2	-7.55	.64	.77	-.7	.70	-.7
124 G30	A1	-7.99	.68	.77	-.6	.68	-.7
3 S03	A1	-8.48	.74	.80	-.3	.71	-.4
61 L06	A1	-8.48	.74	.85	-.2	.85	-.1
58 L03	A1	-9.10	.85	.62	-.6	.43	-.6
106 G12	A1	-9.10	.85	.82	-.1	.83	.0
107 G13	A1	-9.10	.85	1.73	1.2	3.38	2.1
80 P06	A1	-10.03	1.12	1.21	.5	.78	.2
93 P19	A1	-10.03	1.12	1.21	.5	.78	.2
55 W25	A1	(-11.42 | 1.90)	Minimum			
Mean (Count: 124)		-.01	.63	.99	-.2	.97	-.2
S.D.		5.44	.20	.88	1.3	.84	1.3

RMSE .66 Adj (True) S.D. 5.40 Separation 8.22 Strata 11.29 Reliability .99
Fixed (all same) chi-square: 8615.5 d.f.: 123 significance (probability): .00