CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

CEFR alignment: Combining the best of different methods

Paraskevi (Voula) Kanistra, Trinity College London, Great Britain
Jayanti Banerjee, Worden Consulting, United States of America

https://doi.org/10.37546/JALTSIG.CEFR8-5

The alignment of language assessments to the Common European Framework of Reference for Languages (CEFR) is traditionally a complex and lengthy process. Test developers either create a test first and align it to the CEFR post-development, or they integrate CEFR standards from the outset. Both methods necessitate strict adherence to a "series of well-established and largely sequential steps" (British Council et al. 2022: 13). This article introduces a transformative shift in this traditional paradigm by adapting existing standard-setting techniques and leveraging modern tools to streamline alignment procedures. Three standard-setting methods, the Dominant Profile Judgement method, the Item Descriptor Matching method, and the Body-of-Work method, were amalgamated to structure and inform a principled approach to content creation and standard-setting preparation. The ISE Digital writing module will be used to demonstrate how this process expedited panellist alignment and contributed to panellist agreement, both within and between panels.

Keywords: CEFR, ISE Digital, multi-method standard setting, virtual standard setting, Unified Alignment and Test Development (UATD) approach, Dominant Profile Judgement method, Item Descriptor (ID) Matching method, Body-of-Work method

1 Introduction

It is an accepted requirement that exam scores must carry a defensible and interpretable meaning. The *Common European Framework of Reference for Languages: Learning, teaching, assessment* (Council of Europe [CoE] 2020) offers language examinations a shared, interpretable meaning. However, this shared meaning is, in turn, contingent on valid and reliable cut scores. Several guides have been published for developing CEFR-referenced exams and relating existing exams to the CEFR. The most recent is a handbook intended to support the alignment of all language education activities with the CEFR which states that the alignment process entails either "[c]ollecting evidence and developing an argument to show that an existing resource [...] fulfils criteria derived from the CEFR" or "[d]eveloping and documenting a new resource [...] on the basis of the CEFR criteria" (British Council et al. 2022: 13). Both processes necessitate adherence to four steps. The first two steps, familiarization and specification, and standardization, ensure that everyone involved in designing an exam, as well as in the post-hoc standard-setting process, has a thorough understanding of the performance level descriptors (PLDs), in this case, the CEFR. The third and fourth steps, standard setting and validation, establish the recommended cut scores and provide evidence to support their defensibility.

There is an extensive body of standard-setting literature; Kaftandjieva (2010) highlights that there are more than 60 methods available. With such a vast array of options, she advocates caution and warns that standard setting is subjective and context dependent. It must be approached systematically and

carefully because "[i]f the cut scores are inadequate they raise serious doubts about the validity of the interpretation of the test results" (2010: 7). Kaftandjieva makes several key recommendations (2010: 131–135) including, crucially, to use multiple (context relevant) methods in the standard-setting process and to compare and check the results from each method.

Another innovation in standard-setting methodology has been the adoption of online tools that support remote panels (Kanistra forthcoming; Kollias 2023). A key element in standard-setting workshops is the discussion period following each judgement round. If rushed, panellist alignment and the recommended cut scores can be adversely affected. However, in-person workshops are typically time-limited, and this can create pressure on several aspects of the standard-setting process. A carefully designed online workshop featuring self-paced, asynchronous activities alongside well-structured, synchronous discussion sessions can actively promote strong procedural evidence to underpin the recommended cut scores. More recently, attention has also turned to the systematic use of standard-setting methods during the preparation stage of a workshop, helping to structure and strengthen the foundations of the process (Kanistra 2025).

These innovations were central to the design of this study, which had three aims:

1. to align a new digital variant of Trinity's Integrated Skills in English exam (ISE Digital) with the CEFR throughout the test development process,
2. to set CEFR cut scores using multiple methods and triangulate those results when finalizing them, and
3. to maximize the procedural robustness of the familiarization and training steps, as well as the standard setting step, by using online tools that support multiple rounds of self-paced work, and collaboration and discussion.

2 ISE Digital

ISE Digital is a computer-delivered exam that assesses all four language skills individually and together, reflecting how language skills are used in real-life settings. There are four modules. Each module focuses primarily on one language skill and includes several task types (see the ISE Digital exam information booklet for details). The type and number of tasks that test takers receive are dependent on their ability. The reading and listening modules are fully computer-adaptive. Task selection for the speaking and writing modules is also adjusted based on the test takers' ability, as measured by a levelling test that everyone completes at the start of the exam.

The test development process followed the Principled Approaches to Assessment Design, Development, and Implementation model (PADDI, Ferrara et al. 2017) and the Unified Alignment and Test Development (UATD) approach developed by Kanistra (forthcoming). The CEFR was a core resource for the draft specifications, which also drew on theoretical and empirical research in communicative language models and the relevant language skills, as well as research into the language demands of the target language contexts. The draft specifications were used to prepare task blueprints and draft tasks. The pilot testing phase informed revisions to the specifications and additional task revision cycles. Prior to final task decisions, Trinity commissioned a claim-by-specification study (Griffiths 2023) which critically reviewed the exam's alignment with the CEFR and offered some recommendations for improvements. The finalized test design reflects close attention to theory, the target language contexts, and the CEFR, positioning it well for the empirical linking phase.

3 Methodology

This article will focus on the empirical linking phase for the performance-based skills and will be exemplified with data from the writing module linking process. This phase comprised two stages: a

preparation stage and the standard-setting workshop. As recommended by Kaftandjieva (2010), this study incorporated several standard-setting methods, each of which was selected for its appropriateness for the context. The chosen methods were:

- Item Descriptor (ID) Matching method
- Dominant Profile Judgement Method
- Body-of-Work method

The ID Matching method (Ferrara and Lewis 2012; Harsch and Kanistra 2020) entails a two-step process. Panellists first identify the knowledge, skills, and abilities (KSAs) required to answer a task correctly. Then, they map these KSAs to performance level descriptors (PLDs), answering the following question: "Which PLD most closely matches the knowledge and skills required to respond successfully to this item?" Cut scores are established by locating threshold regions—the set of items where judgments alternate between two adjacent PLDs (e.g., below basic basic). The method is applied over two or three judgement rounds. In the first round, panellists review the tasks, analyse the KSAs, identify threshold regions and propose cut scores. The second round involves feedback and discussion centred on the threshold regions, after which panellists revise their judgements. The third round is optional. Here, panellists consider normative and impact data, which may lead them to adjust their recommended cut scores. During this round, the emphasis is on the entire threshold region and the recommended cut scores rather than individual tasks. The final cut scores are set by analysing the panellist judgements using item response theory (IRT) modelling to account for panellist variations. This approach is accessible because it requires expert panellists to perform a familiar activity of aligning task demands with PLDs. It also avoids cognitively demanding concepts such as "minimally competent candidate" and probabilistic judgements (as used in the Angoff method). The final cut scores thus reflect a content-driven, transparent alignment between exam tasks and PLDs.

The Dominant Profile Judgement method (Plake et al. 1997) has been designed for complex performance assessments where test takers' responses have multiple scoring dimensions, including various tasks and assessment criteria. In this method, panellists review performance profiles across the relevant dimensions and identify the dominant profile—the performance pattern most representative of a minimally competent examinee at a given level—which is then used to determine the appropriate cut score. This approach is less cognitively demanding for panellists than estimating probabilities and encourages them to focus on authentic performance patterns. It is also more defensible in high-stakes assessments where the cut scores should offer a transparent link between the observed performance patterns and the performance-level descriptors.

The Body-of-Work method (Kingston and Tiemann 2012) focuses on full performances by test takers (i.e., their responses to all tasks in the module) and comprises two rounds. The first round is known as "range finding" and its purpose is to make an initial estimate of the dividing point between performance levels. The panellists receive an ordered set of full performances and must sort them into performance levels. Once this round is complete and the general location of each cut score has been established, there is a *pinpointing* round to determine a more precise location of the cut scores. In this round, panellists receive several performance examples close to the estimated cut scores. This round requires fine-grained decision making to arrive at a more precise cut-score recommendation.

All three methods were ideal for high-stakes performance assessments, such as ISE Digital, where stakeholders require an evidential link between the test performance and the score interpretation. These methods also offer a concrete judgement process that replicates the panellists' professional experience and expertise. However, they all approach the standard setting task slightly differently, with a focus on descriptor matching, score profiles, or test takers' overall performance on all tasks. As such, they offer an opportunity to triangulate cut score judgements.

The preparation stage was completed by the ISE Digital development team. The team refreshed their understanding of the CEFR, the writing construct, tasks, and assessment criteria. After this re-familiarisation process, the team reviewed both writing task types, *written online communication* (WOC) and *writing from sources* (WS). They mapped these task types to the relevant KSAs using the ID Matching method, ensuring alignment between task demands, construct coverage and CEFR alignment. They then used a modification of the Dominant Profile method to map the writing assessment criteria to the CEFR. This activity enabled the team to select the CEFR scales and descriptors that best aligned with the writing construct, establish score profiles aligned with the target CEFR levels (A1-C2), and predict the expected cut scores. Subsequently, the facilitator applied the range-finding techniques described in the Body-of-Work method, both to define the score range for the targeted CEFR levels (i.e., A1-C2) and to select appropriate responses for the external benchmarking study, thereby helping to confirm the tentative cut scores derived from the Dominant Profile Method. Performances that were clearly outside this range (e.g., far below the expected cut score level) were excluded. The aim was to reduce the external panellists' fatigue and cognitive load, ensuring they focused on relevant scripts and improving the quality and consistency of their judgments.

The order of presentation can influence judgements as panellists have a natural tendency to compare performances or items (Kanistra forthcoming; Wyse and Babcock 2020). Therefore, the facilitator arranged the selected responses in ascending order, from lower to higher scores. The sequence included several tied responses (those receiving the same score). These ties acted as pinpointing tasks, enabling panellists to validate their decisions and mitigate any biases introduced by the response order, thereby ensuring greater consistency in panellist judgments.

The external standard-setting workshop was completed by 15 panellists, all of whom met Raymond and Reid's (2001: 130) criteria for panel selection, especially representativeness and expertise. The panellists were grouped as two sub-panels, one with only external panellists (n = 10) and one with only internal panellists (n = 5). The creation of sub-panels, which were kept separate during the standard-setting workshops, supported a post-hoc check of the recommended cut scores.

The workshop was conducted online over several days using Adobe Connect. There were synchronous and asynchronous sessions, which supported focused individual work and collaborative discussion. The workshop facilitator was available online even when panellists were working asynchronously. The sessions covered four key standard setting steps: orientation, familiarization, training in the method, and standard setting and benchmarking. The panellists were required to complete the writing module under test-taking conditions. This gave them first-hand experience of the cognitive and linguistic demands of the tasks, enabling insights into the targeted knowledge, skills, and abilities. An additional aim of this activity was to help panellists assess task difficulty more accurately, thereby reducing the potential for bias in their cut-score decisions. The panellists also individually completed a CEFR familiarization task. The synchronous activities comprised a briefing on the writing module construct, after which they received a written summary of the construct for reference throughout the workshop, and a review of the outcomes of their CEFR and test familiarization activities. The third workshop step entailed training and practice in using the ID Matching method, conducted synchronously to support an easy exchange of information and clarification questions. Finally, in step four, the panellists completed three judgement rounds with a group discussion of the judgement outcomes between rounds one and two and then again between rounds two and three. The judgements were performed asynchronously, but the group discussions were synchronous.

All panellists completed an evaluation questionnaire after step three (the training phase) and again after step four (the judgement phase). These gathered feedback from the panellists on the procedural adequacy of the standard-setting procedures (Cizek 2012), especially their confidence in the process and the resulting cut scores. Feedback from the step three questionnaire was reviewed before step four, so that any remaining concerns and/or queries about the ID Matching method could be addressed before the judgement step.

4 Results

For the recommended cut scores to be valid, panellists must be very familiar with the CEFR levels and demonstrate their ability to rank order CEFR descriptors accurately. Therefore, a minimum score of 80% was set as the pass criterion for the CEFR familiarization activity following Cicchetti and Sparrow's (1981, cited in Cicchetti 1994) suggestions for rater agreement. Five of the external judges did not meet the minimum familiarity criterion for one scale (this differed by judge), but all demonstrated an average familiarity of 88% or higher. Importantly, panellists received scoring feedback and repeated a task until they achieved 80% accuracy. This ensured that all panellists achieved an acceptable percentage of correct answers on every scale before proceeding to the standard-setting tasks. One of the benefits of working online is that familiarization activities can be phased to ensure that every panellist reaches the required level of expertise before the judgement phase proceeds.

In accordance with Harsch and Kanistra (2020), panellists evaluated the students' written scripts—15 WOC performances (on six different tasks) and 13 WS performances (on three different tasks)—using the Written Assessment Grid (CoE 2020: 187). The panellists made four judgements per script, one for each assessment criterion, resulting in 900 judgements per round for the WOC task and 975 judgements for the WS task. Rasch Measurement Theory (RMT) was used to explore panel consistency and reliability. Table 1 presents a summary of the inter-panellist agreement and intra-panellist consistency results after the round 2 judgements; the full analysis is available in Kanistra (2025).

Table 1. ***Summary of inter-panellist agreement and intra-panellist consistency within RMT (n = 15)***

Index	Task 1	Task 2
Overall exact observed % agreement (expected %)	36% (34.9%)	46.8% (43.2%)
exact observed % agreement (expected %) minimum	26.6% (30.5%)	20.1% (27.9%)
exact observed % agreement (expected %) maximum	47.5% (37.9%)	58.2% (47.1%)
Mean Infit *Mnsq;* SD (*Zstd*)(Group)	0.84; 0.25 (-0.70)	0.91; 0.33 (-0.50)
Minimum Infit *Mnsq* (*Zstd*)	0.37 (-3.02)	0.37 (-2.07)
Maximum Infit *Mnsq* (*Zstd*)	1.32 (1.10)	1.40 (1.30)

The overall exact observed inter-panellist % agreement values were within ± 5 of the expected % agreement, indicating that the panellists acted as autonomous experts and exhibited an acceptable level of inter-panellist agreement. The mean Infit Mnsq values for all panellists were close to the ideal value of 1.00 (ranging 0.84 to 1.40 across tasks and judgement rounds). Additionally, the panellists' Infit measures fell within the acceptable Infit range (Infit mean ± 2SD) for both tasks (Pollitt and Hutchinson 1987). These analyses confirm that the panellists were consistent and reliable.

The recommended cut scores were evaluated post hoc for their precision and accuracy, and classification consistency and accuracy. Table 2 shows that the standard error of the mean of the panellists' judgements (SE) and standard deviation of their judgements (SD) by CEFR level were very small.

Additionally, the SE_j relative to the standard deviation of the population ($SE_j/SD_p \leq .33$ indicates that classification error had minimal influence on CEFR level assignment. Importantly, this also implies that the classifications of the written scripts used in the standard-setting workshop are robust. Note also that the SE_j of the script classifications was consistently lower than one-third of the standard error of measurement (*SEM*) for each cut score ($SE_j/SEM \leq 0.33$), as stipulated by Kaftandjieva (2010). Taken together, these findings offer validity evidence for the consistency-within-the-method aspect of evaluating standard-setting studies.

Table 2. ***Accuracy and precision of the writing cut scores (n = 5,014)***

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

Classification consistency and accuracy were evaluated using a classical test theory (CTT) method (Livingston and Lewis 1995) and an IRT-based method (Lee and Kolen 2008). The recommended cut scores were derived from performances identified as best representing the target CEFR levels. The CTT method used the test takers' raw scores, and the IRT-based method used test-taker ability estimates. This method also required item parameters to be included, so (for this study) the seven assessment criteria (three for the WOC task and four for the WS task) were treated as items. The dataset met the unidimensionality assumption. MFRM analysis was used to arrive at test-taker ability measures and scores for the module, accounting for measurement error due to raters.

For both methods, the decision accuracy [$DA(\gamma)$] and consistency [$DC(\varphi)$] measures at each CEFR level exceeded the recommended minimum criterion of 0.85 (Subkoviak 1988) for certification examinations, with the IRT-based method generally yielding higher indices (Kanistra forthcoming). Additionally, apart from the CEFR C2 cut score, where the κ value for the CTT method is 0.50, the κ values at each CEFR cut score for both methods exceeded the expected 0.60, surpassing 0.76 in the IRT-based method. The anomalous result for CEFR C2 is unsurprising since the cut score is very close to the maximum weighted raw score of 47. Subkoviak (1988) states that pchance (φc) increases for cut scores at the lower or upper end of the scale (in this case CEFR A1 and CEFR C2), and this bears out in the analysis. However, this is predictable since the least and most able test-takers tend to perform similarly regardless of test form. While κ values are affected by statistical edge effects, the A2-C1 cut scores provide the most informative results, with high κ values indicating that test-taker classification largely depends on their performance on the assigned tasks.

5 Discussion and reflections

This study has aligned the ISE Digital writing module to the CEFR through three stages: during the design phase, as part of the item writing and piloting cycle, and through standard setting using a combination of methods. Therefore, the module is aligned to the CEFR both qualitatively, in terms of content, and quantitatively, through standard setting. The standard-setting process took to heart Kaftandjieva's (2010) recommendation to use multiple, context-relevant methods. It also involved different panels and teams. Both methodological decisions supported the triangulation of cut score recommendations. Additionally, the study adopted innovations from Kollias (2023) and Kanistra (forthcoming), using online tools to create a flexible and rigorous workshop design that prioritized good decision making. This was confirmed by the panellist surveys. Most panellists "strongly agreed" or "agreed" that the standard-setting procedures enabled them to effectively map writing tasks and responses to the targeted CEFR levels. The facilitator's role was highly appreciated, ensuring inclusive and balanced discussions. Panellists also felt confident in their ratings and found other panellists' ratings helpful for advising their judgments. Additionally, the group-recommended CEFR classifications for Tasks 1 and 2 were widely endorsed as reflective of the minimum performance levels for the targeted CEFR standards.

As described previously, the empirical linking phase comprised two stages: a preparation stage and the standard-setting workshop, which applied different standard-setting methods. Table 3 presents the recommended cut scores for each stage (and method). It shows that the different groups (ISE Digital development team and standard-setting panellists) are very well aligned. This justifies the use of the Dominant Profile method to select the standard-setting performances. It also confirms that, if a test's alignment with the CEFR is critically revisited and adjusted throughout the development process, this promotes a strong alignment between internal and external judgements.

Table 3. *Recommended raw score ranges for each CEFR level (by standard-setting method)*

	Dominant Profile method	**ID Matching method**
CEFR Level	(Preparation Stage)	(Standard-setting Workshop)
A1	6-11	6-11
A2	12-19	12-19
B1	20-26	20-24
B2	27-30	25-32
C1	31-34	33-34
C2	35-36	35-36

Importantly, each CEFR calibration cycle during the test development process and the standard-setting methods applied in the linking process were incorporated into the normal test design, development, and standard-setting activities. As such, they did not present an increase in burden during any of the stages. That the different panels (the test development team and the standard-setting panel) arrived at such closely aligned recommendations, even though they used different methods and had different characteristics (especially their relative familiarity with the exam), is excellent evidence for the promise of the structure of the alignment process. Online tools, when used systematically and rationally, maximize panellist engagement, their understanding of the CEFR and the standard-setting methodology, and their confidence in their recommendations.

6 References

British Council, UKALTA, EALTA & ALTE. 2022. *Aligning language education with the CEFR: A handbook.* http://www.ealta.eu.org/documents/resources/CEFR%20alignment%20handbook.pdf. (accessed 28 January 2026).

Cicchetti, Domenic V. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardised assessment instruments in psychology. *Psychological Assessment* 6(4). 284-290. https://doi.org/10.1037/1040-3590.6.4.284.

Cicchetti, Domenic V. & Sara Sparrow. 1981. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency* 86(2). 127-137.

Cizek, Gregory J. 2012. The forms and functions of evaluations in the standard setting process. In Gregory J. Cizek (ed.), *Setting performance standards: Foundations, methods, and innovations*, 164-178. New York: Routledge.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion Volume*. Strasbourg: Council of Europe.

Ferrara, Steve & Daniel M. Lewis. 2012. The Item-Descriptor (ID) Matching method. In Gregory J. Cizek (ed.), *Setting performance standards: Foundations, methods, and innovations*, 255-282. New York: Routledge.

Ferrara, Steve, Emily Lai, Amy Reilly, & Paul D. Nichols. 2017. Principled approaches to assessment design, development, and implementation. In André. A. Rupp & Jacqueline P. Leighton (eds.), *The Handbook of cognition and assessment: Frameworks, methodologies, and applications* (First Edition), 41-74. Chichester: John Wiley & Sons, Inc.

Griffiths, Mark. 2023. *Linking ISE Digital to the CEFR: A claim by specification*. Trinity Research Report 2023-01. London: Trinity College London.

Harsch, Claudia & Voula Paraskevi Kanistra. 2020. Using an innovative standard setting approach to align integrated and independent writing tasks to the CEFR. *Language Assessment Quarterly* 17(3). 262-281. https://doi.org/10.1080/15434303.2020.1754828.

Kaftandjieva, Felianka 2010. *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: CiTO. https://ealta.eu/documents/resources/FK_second_doctorate.pdf (accessed 15 August 2025).

Kanistra, Paraskevi. 2025. *Linking ISE Digital to the CEFR: Setting cut scores and performance standards*. Trinity Research Report 2024-01. London: Trinity College London.

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

Kingston, Neal M. & Gail C. Tiemann. 2012. Setting performance standards on complex assessments: The Body of Work method. In Gregory J. Cizek (ed.), *Setting performance standards: foundations, methods, and innovations*, 201-223. New York: Routledge.

Kollias, Charalambos. 2023. *Virtual standard setting: Setting cut scores*. Berlin: Peter Lang.

Lee, Won-Chan & Michael J. Kolen. 2008. *IRT-CLASS: IRT classification consistency and accuracy v 2.0*. University of Iowa. https://education.uiowa.edu/casma/computer-programs (accessed 28 September 2025).

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

Pollitt, Alastair & Carolyn Hutchinson. Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing* 4(1). 72-92.

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

Raymond, Mark R. & Jerry B. Reid. 2001. Who made thee a judge? Selecting and training participants for standard setting. In Gregory J. Cizek (ed.), *Standard setting: Concepts, methods, and perspectives,* 119-157. Mahwah: Lawrence Erlbaum.

Subkoviak, Michael J. 1988. A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement* 25(1). 47-55.

Wyse, Adam E. & Ben Babcock. 2020. It's not just Angoff: Misperceptions of hard and easy items in bookmark-type ratings. *Educational Measurement: Issues and Practice* 39(1). 22-29. https://doi.org/10.1111/emip.12315.

7 Biographies

Paraskevi (Voula) Kanistra holds a PhD in language testing (University of Bremen) and is Associate Director/Senior Researcher at Trinity College London. She is a highly experienced assessment specialist with experience in all aspects of language test design development including item writing, assessor training, and post-hoc statistical analyses of test data. She has particular expertise in (virtual) standard setting, (CEFR) alignment projects, measurement analysis (Classical Test Theory and Rasch Measurement Theory), quantitative and qualitative research, mixed-method research, and validation studies. She has presented her research at international conferences in Europe and Asia and has published in *Language Assessment Quarterly* and *Assessing Writing*.

Jayanti Banerjee holds a PhD in Applied Linguistics (Lancaster University) and is a language assessment professional and researcher with experience as a teacher, university lecturer, and assessment designer and researcher. She has led projects to develop new language tests and advised on strategies for test development, product improvement and assessment techniques. She has also developed and managed research grant programmes and has published articles in leading journals, including the *Annual Review of Applied Linguistics*, *Language Testing*, and *Assessing Writing*. She is particularly interested in research into innovative task designs, rating scale validation, and equality, diversity and inclusion in language assessments.