CEFR Journal

Research and Practice

Japan Association for Language Teaching (JALT)

CEFR & Language Portfolio SIG (CEFR & LP SIG)

Volume 8 (February 2026)

ISSN 2434-849X

Title: CEFR Journal—Research and Practice

Type: Online Journal
URL: https://cefrjapan.net/publications/journal
Contact: journal@cefrjapan.net

Edited by: Japan Association for Language Teaching (JALT)
CEFR & Language Portfolio SIG
Fergus O'Dwyer (editor)
Dmitri Leontjev (editor)
Elif Kantarcıoğlu (editor)
Maria Gabriela Schmidt (president, editor)
Morten Hunke (liaison officer, editor)
Alexander Imig (treasurer, website editor)

ISSN: 2434-849X
DOI https://doi.org/10.37546/JALTSIG.CEFR

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Table of Contents

Mission statement

The CEFR Journal is an online, open-access, peer-to-peer journal for practitioners and researchers. Our editorial advisory board comprises stakeholders on a wide range of levels and from around the world. One aim of our journal is to create an open space for exchanging ideas on classroom practice and implementation related to the CEFR and/or other language frameworks, as well as sharing research findings and results on learning, teaching, and assessment-related topics. We are committed to a strong bottom-up approach and the free exchange of ideas. A journal by the people on the ground for the people on the ground with a strong commitment to extensive research and academic rigor. Learning and teaching languages in the 21st century, accommodating the 21st century learner and teacher. All contributions have undergone multiple double-blind peer reviews. We encourage you to submit your texts and volunteer yourself for reviewing. Thanks a million.

Aims, goals, and purposes

Our aim is to take a fresh look at the CEFR and other language frameworks from both a practitioner's and a researcher's perspective. We want the journal to be a platform for all to share best practice examples and ideas, as well as research. It should be globally accessible to the wider interested public, which is why we opted for an open online journal format.

The impact of the CEFR and now the CEFR Companion Volume (CEFR/CV) has been growing to previously wholly unforeseeable levels. Especially in Asia, there are several large-scale cases of adoption and adaptation of the CEFR to the needs and requirements on the ground. Such contexts often focus majorly on English language learning and teaching. However, there are other language frameworks, such as the ACTFL and the Canadian benchmarks, and the Chinese Standard of English (CSE). On the one hand there is a growing need for best practice examples in the form of case studies, and on the other hand practitioners are increasingly wanting to exchange their experiences and know-how. Our goal is to close the gap between research and practice in foreign language education related to the CEFR, CEFR/CV, and other language frameworks. Together, we hope to help address the challenges of 21st century foreign language learning and teaching on a global stage. In Europe, many take the CEFR and its implementation for granted, and not everyone reflects on its potential uses and benefits. Others are asking for case studies showing the effectiveness of the CEFR and the reality of its usage in everyday classroom teaching. In particular, large-scale implementation studies simply do not exist. Even in Europe, there is a center and a periphery of readiness for CEFR implementation. It is difficult to bring together the huge number of ongoing projects from the Council of Europe (CoE), the European Centre for Modern Languages (ECML), and the EU aiming to aid the implementation of the CEFR. This results in a perceived absence in the substance of research and direction. Outside Europe, the CEFR has been met with very different reactions and speeds of adaptation and implementation. Over the last few years, especially in Asia, the demand by teachers for reliable (case) studies has been growing.

For more than a decade, the people behind this journal—the Japan Association for Language Teaching (JALT) CEFR & Language Portfolio special interest group (CEFR & LP SIG)—have been working on a number of collaborative research projects, yielding several books and textbooks, as well as numerous newsletters. This is a not-for-profit initiative; there are no institutional ties or restraints in place. The journal aims to cooperate internationally with other individuals and/or peer groups of practitioners/researchers with similar interests. We intend to create an encouraging environment for professional, standard-oriented practice and state-of-the-art foreign language teaching and research, adapted to a variety of contexts.

Editorial advisory board

Journal editorial team

Editing and proofreading team

Layout

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Guest Editorial

David Little, Trinity College Dublin, Ireland
Neus Figueras, University of Barcelona, Spain
Lynda Taylor, University of Bedfordshire, Great Britain

https://doi.org/10.37546/JALTSIG.CEFR8-1

The ten articles published in this special issue of the *CEFR Journal* are based on papers given at the conference *Responding to the CEFR Alignment Handbook: Sharing experience of alignment activities and reflecting on lessons learned*, hosted by the GREDA Research Group on Education, Didactics and Learning at Blanquerna—Universitat Ramon Llull, Barcelona in October 2024 (for a report on the conference, see Figueras et al. 2025).[1] Discussion at the conference recognized

- the need for adaptation and localisation of the CEFR and CEFR/CV;
- the critical importance of context and its impact on the use of the CEFR and alignment processes in Europe and beyond;
- the importance of shared understanding, language and discourse;
- the CEFR and CEFR/CV as instruments of social justice;
- the role technology may play in the near future; and
- the need to encourage dissemination initiatives and collaboration.

These issues recur in the ten articles in a variety of ways; some of the articles also reflect on the usefulness of the *Alignment Handbook* (British Council et al. 2022).

Differing from one another in focus and scope, the articles fall into three broad categories: language testing and assessment; pedagogy and materials; policy and curricula.

1 Language testing and assessment

The five articles in this section are all concerned with high-stakes exams: Italian for immigrants in Italy, English for university students in Cuba, English in two international tests, and German for the workplace in Germany.

Diego Cortés Velásquez and Elena Nuzzo report on a project that partially aligned Certit (the certification of Italian as a second language developed at Roma Tre University) with the CEFR. Concerned with written production at level B1, they focus in particular on the test's construct validity, paying particular attention to the concept of tasks as it is elaborated in Chapter 7 of the 2001 CEFR. Having followed the step-by-step procedures recommended by the *Alignment Handbook*, they note the positive impact of training on

1. The conference presentations can be accessed here: https://ealta.eu/members-resources/#CEFR%20HANDBOOK%20Conference

the creation of assessment tasks that seek to reflect the CEFR's conception of language users/learners as social agents. Although the activities they describe are only part of a broader alignment process, they have led to professional awareness raising and improved test documentation, and provide a basis for ongoing evidence-based validation of Certit's writing component.

Nathaniel Owen and Oliver Bigland report on the design, development and CEFR-alignment of an innovative intertextual reading-into-writing summary task for the Oxford Test of English Advanced, targeting CEFR levels B2-C1. The task requires test takers to read two texts on the same topic and synthesise the information they contain into a 100-word summary. The authors used the *Alignment Handbook* to inform methodological decision-making throughout the development and alignment process, adopting an examinee-centred approach to validation. Analysis of data from a CEFR alignment panel and a pilot study indicated strong reliability for the summary task and demonstrated its effectiveness in distinguishing between B2 and C1 performances.

In 2015, the Ministry of Higher Education in Cuba adopted the CEFR as a proficiency framework and level B1 as the exit requirement for proficiency in English. Claudia Harsch and Yoan Martínez Márquez describe the first phase of a CEFR alignment project that entailed familiarization and training for 42 representatives of all centres of higher education in Cuba, the development of test specifications and tasks by the representatives, piloting of tasks on a small scale, and standardization and benchmarking of local examples.

Whether test developers first create a test and then align it with the CEFR or use the CEFR to guide test development, established approaches to alignment entail a complex and protracted process. In their article, Voula Kanistra and Jayanti Banerjee propose an alternative approach that streamlines alignment procedures. They amalgamated three standard-setting methods, the Dominant Profile Judgement method, the Item Descriptor Matching method, and the Body-of-Work method as a way of structuring and informing content creation and the preparation of standard-setting. They found that this process expedited panel alignment and contributed to panellist agreement, both within and between panels. The alignment process was carried out online. The authors argue that this is preferable because it allows novices to work individually at their own pace, and they cannot be influenced by more experienced peers.

In the last of the articles on high-stakes tests, Sibylle Plassmann reports on the alignment of workplace-oriented writing tasks with the CEFR at levels A2, B1, B2 and C1 in the context of the German Tests for Work (Deutsch-Tests für den Beruf, DTB) developed by telc for the German Federal Ministry of Labour and Social Affairs. The standardised exams serve as final assessments in vocational language courses. The article describes the process of defining learning objectives based on authentic workplace communication needs and the adaptation of CEFR descriptors to fit vocational contexts. It also discusses the design of writing tasks that reflect real-world professional communication, the establishment of rating criteria tailored to workplace requirements, and the standard-setting process.

2 Pedagogy and materials

The articles in this section are concerned with two very different contexts: the British Council's English courses for young learners, which are offered in 25 countries, and courses in Italian for foreigners at levels A2 and B1. Certification is attached to the Italian but not to the British Council courses, which nevertheless require assessment procedures that are transparent and yield results that are easy to interpret.

Carolyn Westbrook and Aidan Holland report on a project in which assessment researchers worked with teachers to standardise the approach to assessment in the British Council's global language programme for secondary-age learners. Teaching materials comprising 120 magazines were mapped to the CEFR following the procedures set out in the *Alignment Handbook*; standardised set-up notes were created for assessment tasks; assessment tools and training were developed; and the approach was

then implemented by teachers. The project was informed by the concept of the Comprehensive Learning System, so assessment was developed in close interaction with curriculum and delivery. Feedback showed that the new system improved objectivity and clarity in assessment, though challenges around feasibility and alignment with the CEFR/CV remained.

Published in 2020, the CEFR/CV assigns a new prominence to mediation and includes a large number of illustrative scales for different kinds of mediation activity. To date, however, little research has been published on aligning proficiency tests with mediation descriptors. In their article, Sabrina Machetti and Giulia Peri present the initial findings of a project to align CILS exams (Certification of Italian as a Foreign Language) with mediation descriptors at levels A2 and B1. The exams are aimed at foreign students learning Italian at school in Italy and abroad. The project has found that the exams have had a positive impact on syllabuses in some Italian high schools abroad, bringing them into alignment with the principles of learning-oriented assessment.

3 Policy and curricula

The three articles in this section deal with policy and curriculum reform, the first in the Netherlands, the second in Japan, and the third in Costa Rica, and they provide three very different approaches to alignment with the CEFR and CEFR/CV. The first two are founded on empirical work, the third less so.

Daniela Fasoglio describes a review of national learning objectives undertaken by the Netherlands Institute for Curriculum Development (SLO) on behalf of the Ministry of Education. Focusing on three key educational domains—qualification, socialisation and subjectification—the review sought to integrate the CEFR into the national learning targets while ensuring alignment with broader curriculum principles. A case study involving language teachers applied the methodology outlined in the *Alignment Handbook* to identify attainable proficiency levels for upper secondary education. As the process moves beyond the design phase, a key priority is to maintain curriculum quality, which depends on close collaboration between curriculum developers, school leaders, teachers, educational publishers, and test developers. Constructive alignment between learning goals, pedagogy and assessment is considered essential, and it is hoped that assessment will promote coherence between the CEFR's vision of language learning and language use and the goals of the national curriculum.

Masashi Negishi and Yukio Tono describe the influence of the CEFR on language teaching in Japan, comparing its impact on the teaching of English and Japanese. Drawing on a framework that operationalises key CEFR and CEFR/CV concepts—the action-oriented approach to teaching promoted by the CEFR/CV, the language user/learner as social agent, and the proficiency levels—their study analyses curriculum documents, textbooks and assessment tools. They show that there has been a marked difference in the adoption strategies: a cautious, laissez-faire approach in the teaching of English and a more top-down, mandated approach in the teaching of Japanese. The case of English reveals inconsistent alignment, with some progressive teachers and materials developers filling gaps left by national curricula, while the case of Japanese is characterized by a strong, though not yet widespread, alignment in accredited institutions. The authors conclude by discussing the inherent "power" of the CEFR, not as a prescriptive standard, but as a framework that can drive reform; they highlight the need for targeted training and support to achieve a broader and more uniform impact.

Ana C. González-Ramírez and Walter Araya-Garita report on the impact of the CEFR on Costa Rica's system of foreign language education. Adopted by the Ministry of Public Education in 2016, the CEFR has prompted a shift from traditional content-based teaching focused on grammar and vocabulary to a student-centred communicative approach. This has brought about improvements in students' proficiency levels and professional training for teachers, though challenges remain, including unequal resource distribution and lesson-time constraints. The article emphasises the need for systematic teacher training, the ongoing adjustment of policy, and implementation in higher education.

4 References

British Council, EALTA, UKALTA & ALTE. 2022. *Aligning language education with the CEFR: A handbook*. https://ealta.eu/documents/resources/CEFR%20alignment%20handbook.pdf (accessed 18 November 2025.

Figueras, Neus, David Little, Barry O'Sullivan, Nick Saville & Lynda Taylor. 2025. Responding to the CEFR Alignment Handbook: Sharing experience of alignment activities and reflecting on lessons learned. *CEFR Journal* 7. 112-116. https://doi.org/10.37546/JALTSIG.CEFR7-6.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Aligning Certit's written proficiency tests with CEFR standards

Diego Cortés Velásquez, Università degli Studi Roma Tre, Italy
Elena Nuzzo, Università degli Studi Roma Tre, Italy

https://doi.org/10.37546/JALTSIG.CEFR8-2

This article reports on a partial but significant step in the alignment of Certit (the certification of Italian as a second language developed at Roma Tre University) with the Common European Framework of Reference for Languages (CEFR). Focusing on the written proficiency test, the study describes an institutional process aimed at enhancing the test's construct validity and transparency through reference to the CEFR's action-oriented approach. Following the procedures outlined in the CEFR Alignment Handbook (British Council et al. 2022), a familiarization phase was conducted between March and November 2022 with the team of test designers to consolidate their shared understanding of the CEFR construct of task. This was followed by a systematic review of writing prompts, whose alignment with CEFR principles was analyzed and discussed collaboratively. The outcomes of this process include the refinement of prompt design and the identification of key task features. While representing a limited stage in a broader alignment effort, the process has fostered professional awareness, improved test documentation, and established the foundations for ongoing, evidence-based validation of Certit's writing component.

Keywords: CEFR alignment; language assessment; task; construct validity; Certit; writing proficiency; familiarization phase; action-oriented approach

1 Introduction

This study aims to assess the extent to which an awareness-raising initiative, conducted as part of Certit's CEFR alignment process (British Council, UKALTA, EALTA and ALTE 2022) has influenced the item writers' design of written production prompts.

Certit—one of the four official certifications of Italian as a foreign language—has recently undertaken a comprehensive effort to collect evidence and develop arguments demonstrating that its tests meet criteria derived from the CEFR (Council of Europe [CoE] 2001). The alignment process began with a focus on the B1 level, which, following Law No. 132/2018 (*Decreto Sicurezza*), is required for Italian citizenship applications and is therefore the level with the highest number of test takers. This article examines the alignment of the Certit B1 written proficiency test with the CEFR's action-oriented approach, which defines tasks as purposeful actions undertaken to achieve specific outcomes "in a specific environment and within a particular field of action" (CoE 2001: 9). To this end, written production prompts created and used before and after the awareness-raising initiative are compared and evaluated against a set of six analytical criteria derived from Chapter 7 of the CEFR (2001).

Section 2 presents the theoretical framework, namely the *task* construct as conceived within the CEFR's action-oriented approach. Section 3 outlines the research question, the methods of data collection

and analysis, and the results of the analysis. Section 4 discusses the findings, followed by concluding remarks in Section 5.

2 Theoretical Framework

The *task* represents the core of the language learning vision promoted by the CEFR (CoE 2001), which devotes a whole chapter (Chapter 7) to *Tasks and their role in language teaching*. According to the CEFR, a task is defined as

> any purposeful action considered by an individual as necessary in order to achieve a given result in the context of a problem to be solved, an obligation to fulfil or an objective to be achieved. This definition would cover a wide range of actions such as moving a wardrobe, writing a book, obtaining certain conditions in the negotiation of a contract, playing a game of cards, ordering a meal in a restaurant, translating a foreign language text or preparing a class newspaper through group work. (CoE 2001: 10)

This vision revolves around the simple observation that in everyday life we all carry out tasks of various kinds, some of which require the productive and/or receptive use of language. The centrality of the task, as defined in the CEFR (2001), has been maintained in the *CEFR Companion Volume* (CEFR/CV) (2020), which continues to frame language use in terms of purposeful, contextualized actions. As stated in the CEFR (2001: 157): "Communication is an integral part of tasks where participants engage in interaction, production, reception or mediation, or a combination of two or more of these."

This is the action-oriented view of language, which implies that we use language as a tool to bridge information gaps, and that our success in carrying out daily tasks also depends on how effectively we do so. Accordingly, learning an L2 essentially means learning to perform a growing range of tasks with progressively greater effectiveness.

So-called *real-life* tasks serve as core components of many syllabuses, textbooks, classroom activities, and assessments, though they are frequently adapted for instructional or testing purposes. These tasks are selected based on learners' needs beyond the classroom, whether in personal and public contexts or in connection with specific occupational or educational goals (on needs analysis, see, for example, Grote and Oliver 2022; Long 2005; Malicka et al. 2019). Other classroom tasks are distinctly pedagogical and only loosely connected to real-life situations or learners' practical needs. These may be derived from real-life tasks by simplifying them—usually by breaking them down into their component parts or sub-tasks (Long 2015). In fact, "a particular task may involve a greater or lesser number of steps or embedded sub-tasks" (CoE 2001: 157).

Whether modelled on real-life use or designed for instructional purposes, classroom tasks are communicative insofar as they require learners to understand, negotiate, and convey meaning to achieve a specific communicative goal. The presence of a clear communicative goal is therefore an essential feature of a task, and on this point all authors who have addressed the topic agree, even though they offer partly different definitions of a task in language teaching (see, among others, Ellis 2005; Nunan 2004; Willis and Willis 2013).

From what has just been said, it follows that any written output resulting from the accomplishment of a communicative task must have a purpose, even if fictional. From an assessment perspective, this has important implications: when there is a clear and defined objective, the assessor can evaluate whether—and how effectively—the student has achieved it. If, on the other hand, the objective is absent or vague, the assessor can only judge whether the learner's language output exhibits certain formal features, and to what degree of accuracy. However, "the primary focus [of a task] is on meaning as learners realize their communicative intentions" (CoE 2001: 158). Let us consider a concrete example.

If a student is asked to write a description of the last present they received, the communicative goal is rather vague, if not absent. Who is the learner supposed to address the description to? What is the

purpose of writing that description? Instead, we could have a task if a relevant context were set up, such as a fictional situation that requires the learner to write the description for a specific recipient and with a clear goal. For instance, the student could be asked to write a product description for a second-hand app, in order to sell the backpack they received as a gift because they already have one and want to give away the duplicate. In this second case, both the purpose and the recipient are clearly identifiable, making it possible to assess whether, and to what extent, the text produced is effective and appropriate.

In our view, only the second activity can be defined as a task, whereas the first one is simply a written composition exercise. It is only in the second case that one can assess whether the purpose has been fulfilled in a manner appropriate to the context and the recipient—that is, to evaluate what a learner *can do*, and how well, with the L2. A true alignment of productive tests, particularly written ones, with the CEFR requires that such tests be presented as communicative tasks rather than as traditional composition exercises.

The CEFR's conception of the communicative task as a purposeful action entails a set of essential characteristics. As noted, the focus should be on meaning rather than form, and tasks must lead to a tangible communicative outcome ("tasks [...] have identifiable (and possibly less immediately evident) outcomes"; CoE 2001: 158). A task also requires a recipient, without whom communication would not take place, and it should be situated in realistic or plausible contexts that facilitate authentic language use. Finally, in line with the action-oriented approach, learners should be able to make meaningful decisions about how to achieve the communicative goal. This highlights another key characteristic of tasks: they grant learners a degree of autonomy, reflecting the flexibility and unpredictability of real-world language use.

3 The study

In July 2022, the Certit staff began a process to review the task prompts in the Writing section of the B1 exam. As already mentioned, this is a particularly important level, as it is the one required by Italian law for applying for citizenship. The aim of the review was to ensure alignment with the CEFR and in particular with the construct of writing tasks as conceived in the document. As part of the familiarization phase recommended in the CEFR Alignment Handbook (British Council, UKALTA, EALTA and ALTE 2022), a series of meetings was organized with the test designers, during which exam prompts were reviewed and their alignment with the CEFR's action-oriented approach was collaboratively analyzed. This work formed part of a broader restructuring of the Certit system, the first outcome of which was the publication of the Guide to the Certit Certification (Di Salvo and Vitale 2023).

With this study, we aimed to assess the outcomes of this process of raising awareness among Certit test creators regarding the alignment of prompts with the CEFR approach. Specifically, we addressed the following research question:

> *To what extent did the review process, and particularly its awareness-raising function, affect the degree to which activities in the Certit B1 Writing section match the CEFR definition of communicative tasks?*

To answer this question, we analyzed separately the prompts used before and after the awareness-raising initiative. We hypothesized that the written production tests administered prior to 2022 largely consisted of open-ended composition exercises, whereas those introduced from 2022 onward would more closely reflect the CEFR's criteria for communicative tasks.

3.1 Data collection and analysis

We collected two datasets of B1-level written production prompts administered in the Certit examinations. The first dataset, labelled PRE-A, comprised 50 prompts created for use between 2014 and 2022, that is, prior to the alignment initiative. The second dataset, labelled POST-A, included 16 prompts created by the test designers after the review process and administered between 2022 and 2024.

To assess the alignment of Certit's B1 writing prompts with the CEFR's action-oriented approach, the two authors analyzed the dataset of prompts with reference to criteria grounded in Chapter 7 of the CEFR (CoE 2001). These criteria reflect the essential characteristics of a task identified in Section 2, namely:

- Communicative goal
- Focus on meaning
- Addressee
- Outcome
- Context
- Autonomy

Each author independently assigned a score of 0 if the prompt did not meet each criterion, 0.5 if it partially met it, and 1 if it fully met it. After the independent ratings, the authors met to discuss their scores. Overall agreement was high: the two raters scored all prompts identically on most criteria, making it impossible to apply interrater agreement measures because of the lack of variance. There were, however, minor differences in the assessment of some prompts with reference to Focus on meaning and Autonomy.

3.2 Results

3.2.1 The PRE-A dataset

The analysis of the PRE-A dataset suggested that the prompts fell into three main categories of activities: proper tasks, composition exercises (or pseudo-tasks), and non-tasks.

The prompts that could be considered as proper tasks were 25 in total and obtained a score of 4.5 (out of 6) or higher by both raters. The highest-scoring prompts shared a consistent profile characterized by a clear communicative goal, a well-defined addressee, and a concrete, recognizable outcome (see figure 1, shown here in English translation—as are all subsequent figures—due to space constraints). They were typically framed within plausible everyday or transactional contexts, such as requesting information from a service provider, arranging or declining social invitations, booking or cancelling accommodation, reporting a theft, or communicating with a doctor. Thematic domains were closely aligned with B1 "can do" descriptors, focusing on personal and social life, housing, travel, health, and official procedures. While the prompts specified essential content elements, they also allowed candidates autonomy in choosing details, thereby supporting purposeful language use in realistic scenarios. Most tasks required the production of short, self-contained written correspondence (e.g., emails, messages, forms) that closely replicated real-world communicative practices.

Yesterday you received this email. Reply.

You must write between 40 and 60 words.

Write your answer on the Answer Sheet.

"Dear Angelo,

I haven't heard from you for many months, how are you? What are you planning to do this summer? Why don't you come and visit me in Palermo? There are so many things we could do together... Write back soon.

Hugs,
Lucia"

Figure 1. PS011

Twenty-one prompts, all of which scored 2.5 or lower, were classified as "pseudo-tasks," due to the fact that, while superficially resembling tasks, they did not fully meet the requisite criteria. These were typically open-ended descriptive or narrative composition exercises, as shown in figure 2. While they often encouraged personalization and connected to learners' own experiences, they lacked the goal-oriented structure, defined addressee, and tangible outcome that characterize action-oriented tasks as described in the CEFR. Their communicative potential was therefore largely confined to sharing personal information or storytelling, with limited opportunity to simulate authentic social interactions or problem-solving situations.

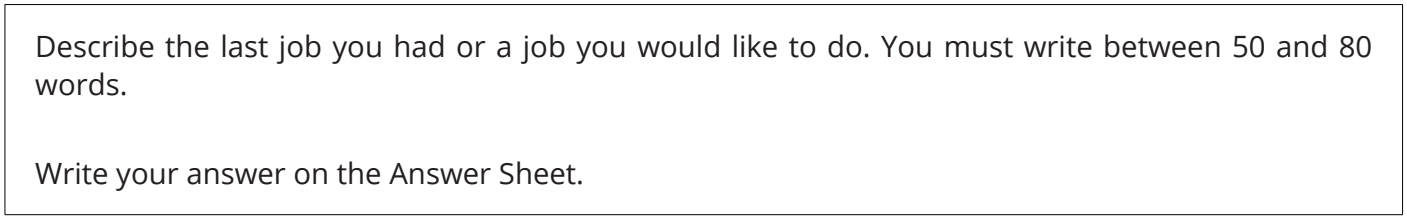

Figure 2. PS044

Within this category of pseudo-tasks, one prompt deserves particular attention, as it displays features that might suggest it belongs to the group of proper tasks (see figure 3). At first glance, the prompt appears to be communicatively oriented: it specifies a clear text type (an email) and identifies an addressee (a friend). However, on closer inspection, the task lacks a genuine communicative goal. The purpose of the exchange is not defined, as the only requirement is to "write an email and talk about a trip you would like to take soon". Without a concrete objective, such as persuading the friend to join, asking for advice, or making arrangements, the prompt encourages a general account of a personal plan.

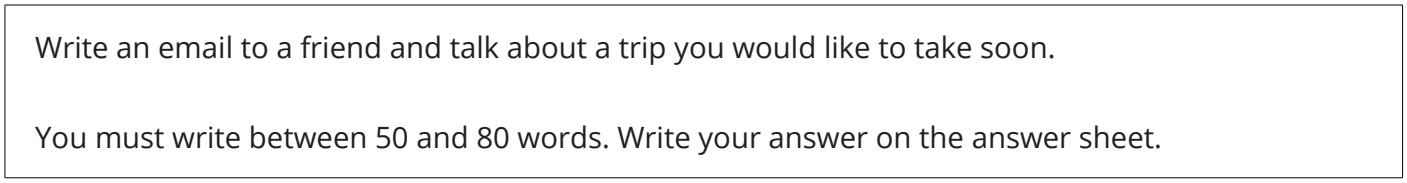

Figure 3. PS046

A final category in our dataset comprised three prompts requiring candidates to complete missing turns in a scripted dialogue based on given cues, as shown in figure 4. These items, which scored among the lowest, could not be considered genuine tasks in the CEFR sense and were labelled as non-tasks. While they involved inserting language into a communicative exchange, the interaction was heavily pre-determined: the context, turns, and communicative purpose were fixed, and candidates simply produced short utterances that fit the given prompts. Such exercises resembled controlled gap-filling activities rather than open-ended, goal-oriented tasks, as they offered minimal autonomy, no scope for negotiation of meaning, and no authentic outcome beyond completing the script. In addition, these prompts were framed as written tasks, yet they simulated a spoken dialogue, resulting in a modality mismatch that undermined their authenticity.

A. City of Rome, good morning. How can I help you?

B. _____
(Greets. Says that they need a residence permit.)

A. Do you live here in Rome?

B. _____
(Says yes and asks how long it takes to get the residence permit.)

A. Usually, the permit is issued within one month from the application. Are you in Italy for work?

B. _____
(Says no and explains that they are in Italy to study at university.)

A. All right, then you must go to the Municipality of your District with all your documents, fill out the application form, pay the fee, and submit everything to the clerk at the counter.

B. _____
(Asks which documents are needed to submit the application.)

A. You must bring with you: your passport, proof of payment of the university enrolment fee, two passport-sized photographs, and the rental contract for the house where you live.

B. _____
(Thanks and says goodbye.)

A. Have a good day.

Figure 4. PS010

In all the prompts analyzed, we found strict word limits (e.g., "write 70-80 words") which, while perhaps useful to provide a reference point for the test taker, may constrain the candidate's ability to fully engage with the task. In authentic communicative situations, the successful completion of a task is not defined by word count but by whether the communicative purpose is achieved. Thus, rigid word restrictions can introduce an artificial constraint that undermines the real-world nature of the task and limits the learner's agency in constructing an adequate, meaningful response.

3.2.2 The POST-A dataset

All prompts in the POST-A dataset, which scored between 5.5 and 6, met the six criteria (see Section 3.1). Each prompt specified a clear communicative goal and identified an addressee, whether an individual or a group. The focus was consistently on meaning, and the expected outcome was explicit (e.g., proposing an itinerary, organizing an event, making a complaint, requesting services). The context was well defined through situational framing, ensuring that tasks were anchored in realistic and recognizable settings. Finally, the prompts provided space for learner autonomy: even the indication of word length had been rephrased, moving from a strict requirement to an indicative guideline, so as to promote more authentic responses and reduce the risk of producing unnatural texts (see figure 5).

You are the first to learn that a friend from your group has just had a baby. Write a message in the WhatsApp group with your other friends to share the good news, propose some ideas for a gift, and organize a visit to their home.

For this text, a total length of between 50 and 80 words is expected.

Write your answer on the Answer Sheet.

Figure 5. PS070

The range of domains spanned both personal and public spheres, covering everyday life situations (e.g., neighbourhood matters, sports events, household issues) as well as transactions requiring information exchange or problem-solving. These prompts also varied between informal digital communication and more formal written genres, such as e-mails or advertisements, thereby encompassing a spectrum of registers and purposes while maintaining a tangible, real-world orientation

Compared with the PRE-A dataset, where almost half of the prompts took the form of decontextualized compositions or gap-filling exercises, the POST-A prompts represented a clear shift towards the ability to create fully contextualized, goal-oriented tasks that were closely aligned with the CEFR's action-oriented perspective.

4 Discussion

In our view, a recurring source of misunderstanding for test writers in the implementation of CEFR principles with regard to the design of prompts for production tests lies in the relationship between two intertwined key concepts: text and task. According to the CEFR, a text is defined as "any sequence or discourse (spoken and/or written) related to a specific domain and which in the course of carrying out a task becomes the occasion of a language activity, whether as a support or as a goal, as product or process" (CoE 2001: 10).

In this sense, the text plays a supporting yet pivotal role; it is central to the communicative event but is not equivalent to the event itself. The task, by contrast, is the purposeful action that frames and gives meaning to the use of the text within a specific context. Misinterpreting the text as the task can lead to assessment formats that privilege decontextualized textual production over authentic, goal-oriented communication.

This distinction has been a guiding principle in the recent revision of Certit's B1 writing tasks, aimed at ensuring that the assessment is not only text-based but also task-realistic and functionally aligned with the CEFR's action-oriented approach. The comparison between the PRE-A and POST-A datasets illustrates the practical consequences of this shift. In the PRE-A dataset, nearly half of the prompts were either decontextualized composition exercises or controlled gap-filling items. These formats tended to focus on producing a particular text type in isolation, without a clear communicative purpose, explicit audience, or tangible outcome. By contrast, the POST-A dataset shows a marked change in design philosophy. All prompts are situated within plausible, concrete scenarios—ranging from writing a WhatsApp message to neighbours about a community issue, to sending a formal email to request information or lodge a complaint. Each task specifies an identifiable addressee, a defined communicative goal, and an explicit expected outcome. The range of registers, from informal digital messages to formal letters, reflects the varied purposes and contexts in which language is used in real life. The shift from PRE-A to POST-A can therefore be interpreted as an alignment with the CEFR's action-oriented perspective: texts are no longer an end in themselves, but tools embedded within purposeful communicative actions. This not only

increases the validity of the tasks but also enhances their potential for eliciting authentic language use, promoting a better alignment between test performance and real-world communicative competence.

5 Concluding remarks

The alignment process presented in this article has several important implications. The first, and perhaps most obvious for a certification body, is that rating scales must be directly linked to the CEFR task construct discussed in this study. As a consequence of revising the B1 writing prompts, Certit also initiated a revision of its rating criteria. This led to the development of a new functional adequacy-based rating scale, designed to more accurately capture communicative task performance. Drawing on recent advances in task-based assessment (e.g., Kuiken and Vedder 2018), the scale integrates six dimensions: task fulfilment, content richness, comprehensibility, accuracy, cohesion, and lexico-grammatical range. These categories reflect the dual requirement of the CEFR's action-oriented approach: to assess not only the linguistic correctness of a text but also its effectiveness in achieving the intended communicative goal within a specific context. Validation of the scale is currently in progress through pilot studies with trained raters.

The introduction of this new scale also underscores the challenges of implementing CEFR-aligned writing tasks in high-stakes contexts. One difficulty lies in operationalizing descriptors that are sufficiently concrete for reliable rater use, while avoiding the over-simplification of complex constructs such as task fulfilment or comprehensibility. Another challenge concerns staff training, as both item writers and raters must internalize the CEFR's action-oriented vision and apply it consistently across diverse candidate populations. The familiarization initiative reported here represents a first step in this direction, but further iterative refinement will be required to ensure sustainable implementation.

From an institutional perspective, these changes enhance the fairness, transparency, and validity of Certit examinations—qualities that are particularly salient given the high-stakes nature of the B1 exam for citizenship applications. Candidates benefit from clearer task framing and more authentic prompts, which are likely to improve both the face validity of the test and the candidate experience, reducing perceptions of arbitrariness. For Certit as an institution, alignment with CEFR standards supports policy compliance at the national level and contributes to international credibility, reinforcing its role within the landscape of recognized Italian language certifications.

Looking ahead, future steps will not be limited to the empirical validation of the functional adequacy scale, the testing of interrater reliability, or the extension of the revised framework to other proficiency levels. In line with the CEFR Alignment Handbook (British Council, UKALTA, EALTA and ALTE 2022), Certit intends to pursue a comprehensive alignment process that encompasses all components of the certification system. This includes the systematic review of task design, rating scales, and standard setting procedures, as well as examiner training, administration practices, and score reporting. Mapping exercises are also underway to establish explicit links between revised prompts, CEFR descriptors, and performance standards, which will further consolidate the construct validity of the test and enhance transparency. In this way, alignment is understood not as a one-off initiative but as an ongoing institutional commitment.

6 References

British Council, UKALTA, EALTA, & ALTE. 2022. *Aligning language education with the CEFR: A handbook*. British Council, UKALTA, EALTA & ALTE. ISBN 9781739754419.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

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

Di Salvo, Francesca & Giuseppina Vitale. 2023. *Il manuale Certit. Guida alla certificazione dell'italiano L2 dell'Università degli Studi Roma Tre*. Roma TrE-Press. https://doi.org/10.13134/979-12-5977-203-9.

Ellis, Rod. 2005. Instructed language learning and task-based teaching. In Eli Hinkel (ed.), *Handbook of research in second language teaching and learning*, 713-728. Mahwah, NJ: Lawrence Erlbaum Associates.

Grote, Ellen & Rhonda Oliver. 2022. A task-based needs analysis framework for TBLT: Theory, purpose, and application. In Alessandro Benati & John W. Schwieter (eds.), *Second language acquisition theory: The legacy of Professor Michael H. Long*, 235-256. Amsterdam: John Benjamins.

Kuiken, Folkert & Ineke Vedder. 2018. Assessing functional adequacy of L2 performance in a task-based approach. In Naoko Taguchi & Youn-Hee Kim (eds.), *Task-based approaches to teaching and assessing pragmatics*, 265-285. Amsterdam: John Benjamins. https://doi.org/10.1075/tblt.10.11kui.

Long, Michael H. (ed.). 2005. *Second language needs analysis*. Cambridge: Cambridge University Press.

Long, Michael H. 2015. *Second language acquisition and task-based language teaching*. Chichester: Wiley-Blackwell.

Malicka, Ania, Roger Gilabert Guerrero & John M. Norris. 2019. From needs analysis to task design: Insights from an English for specific purposes context. *Language Teaching Research* 23(1). 78-106. https://doi.org/10.1177/1362168817714278.

Nunan, David. 2004. *Task-based language teaching*. Cambridge: Cambridge University Press.

Willis, Jane & David Willis. 2013. *Doing task-based teaching.* Oxford handbooks for language teachers. Oxford: Oxford University Press.

7 Biographies

Diego Cortés Velásquez is Associate Professor of Language Education at Roma Tre University, where he teaches courses on language assessment, plurilingual education, and task-based language teaching. His research interests include language testing and validation, intercomprehension among Romance languages, and plurilingual education policies. He is the scientific coordinator of Certit, the university's certification of Italian as a second language.

Elena Nuzzo is Associate Professor of Language Education at Roma Tre University. Her research focuses on speech act theory in second language learning, cross-cultural communication, and task-based language teaching. More recently, she has explored external manipulation in low-structured learning environments such as tandem and telecollaboration. She is co-editor-in-chief of *Instructed Second Language Acquisition* and serves on several editorial boards. Her publications include peer-reviewed articles, book chapters, monographs, and edited volumes.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Designing and validating an intertextual reading-into-writing summary task: A CEFR-aligned approach using the 2022 Handbook

Nathaniel Owen, Oxford University Press, Great Britain
Oliver Bigland, Oxford University Press, Great Britain

https://doi.org/10.37546/JALTSIG.CEFR8-3

This article reports on the design, development and CEFR-alignment of an innovative intertextual reading-into-writing summary task for the Oxford Test of English Advanced, targeting CEFR levels B2-C1. The summary task requires test takers to read two texts on the same topic (300 words total) and synthesize information into a 100-word summary. The study utilized the 2022 Aligning language education with the CEFR: A handbook (Handbook) to inform methodological decision-making throughout the development and alignment process, adopting an examinee-centred approach to validation. Data from a CEFR alignment panel (n = 7) and a larger-scale pilot study (n = 215) were analysed using many-facet Rasch measurement to investigate task performance, assessor reliability, and concurrent validity with a traditional essay task. Results indicate strong reliability (ICC = .87), equivalent to a traditional essay writing task, and demonstrate the task's effectiveness in distinguishing B2 from C1-level performances. The study provides evidence for the utility of the 2022 Handbook in guiding CEFR alignment methodology and provides evidence that summary writing tasks are a viable alternative to traditional essay writing assessments in high-stakes contexts.

Keywords: CEFR alignment, summary writing, integrated skills, reading-into-writing, mediation, Handbook 2022

1 Introduction

Language testing has historically been predicated on a "four skills" approach, treating listening, speaking, reading, writing as isolated competencies (Lado 1961). However, this traditional approach fails to reflect the integrated nature of language use in real-world contexts, particularly in academic and professional settings where skills are routinely combined to complete specific tasks (Gebril and Plakans 2014; Plakans and Gebril 2012; Yu 2013). The *Common European Framework of Reference for Languages* (CEFR) (Council of Europe [CoE] 2001) advocates for "modes of communication" rather than individual "skills" and has evolved to encompass concepts such as "interaction" and "mediation" within the *CEFR Companion Volume* (CEFR/CV; CoE 2020) which emphasize combining skills for describing overall language proficiency.

Against this backdrop, Oxford University Press has developed an intertextual reading-into-writing summary task for the *Oxford Test of English Advanced*, designed to measure CEFR levels B2-C1, using the *Companion Volume* to inform task design. This task requires test takers to read two texts, select, paraphrase and synthesize information into a single piece of writing, closely resembling the kinds of language use tasks examinees encounter in academic and professional contexts (Sawaki et al. 2009).

As this task was being developed, the publication of the 2022 Handbook *Aligning Language Education*

with the CEFR (British Council, UKALTA, EALTA and ALTE 2022, hereafter referred to as the Handbook) marked a significant development in CEFR alignment methodology, providing updated guidance for aligning language tests, curricula, and course content to the CEFR. The Handbook emphasizes the importance of the Comprehensive Learning System (CLS) approach, which advocates for close alignment of curriculum, delivery, and assessment elements (O'Sullivan 2020).

This article reports on the development and validation of this summary task, demonstrating how the 2022 Handbook informed methodological decisions throughout the alignment process. Specifically, the Handbook distinguishes between examinee-centred and test-centred approaches to validation. The former approach requires the collection of test taker performance samples and having them evaluated by independent, external participants. The latter approach focuses more specifically on the evaluation of test design and content (Handbook: 49). This study provides evidence for the utility of the Handbook's recommendations in guiding CEFR alignment for mediation-related test tasks.

2 Literature review and theoretical framework

2.1 Mediation in the CEFR and the Handbook

The concept of mediation, central to the CEFR Companion Volume, refers to "a social and cultural process of creating conditions for communication and cooperation" (Council of Europe 2020: 106) and includes both cross-linguistic mediation and mediation within a target language, chiefly concerned with facilitating the communicative needs of others. Mediation often occurs across modalities, where written output may involve processing and relaying the message of a spoken text or synthesizing multiple sources. The original CEFR framework (CoE 2001) established a theoretical foundation for mediation in language assessment, while the later CEFR/CV (CoE 2020) expanded the concept to better reflect contemporary language use in electronic and multilingual contexts. The 2020 CEFR/CV and 2022 Handbook also emphasize mediation as a crucial component of CEFR alignment, particularly for integrated-skills tasks. Figueras et al. (2022) provide a comprehensive overview of the Handbook's development and its significance for CEFR alignment methodology.

2.2 The intertextual reading-into-writing construct

The purpose of developing an integrated reading-into-writing task is to better represent the kinds of activities that are fundamental to academic and professional language proficiency domains. When reading for writing, language users adopt appropriate reading strategies to construct models of text structure, construct textual and intertextual representations that allow them to select, evaluate, and use information according to the writing purpose (Weigle et al. 2013). This process represents a form of discourse synthesis (Nelson and King 2022), which refers to operations such as organizing, selecting, and connecting content from multiple sources on the same topic. For summary writing tasks, evaluation criteria should examine content transformation and degree of source in addition to traditional criteria content such as organization, grammar or vocabulary. Higher scores should be awarded for making explicit links across sources, especially where such links may only be implied in the original source texts.

2.3 Designing intertextual reading-into-writing summary tasks

The integrated nature of a summary task means that test developers are required to address multiple design considerations, including input (what test takers are required to read), output (what test takers are required to do), and how responses will be scored. Regarding input, Li (2014) found that source text genre has a significant impact on test taker performance in summary tasks. Narrative and expository texts pose different challenges and elicit different strategies from students. Students performed better when summarizing expository texts compared to narrative texts, as expository texts contain more

explicit topic sentences and hierarchical structures compared to narrative texts. In traditional writing tasks such as essays, text length is often the strongest predictor of test taker performance (Crossley et al. 2023). However, summary tasks require test takers to select information, meaning lengthier responses may show less evidence of test takers' ability to discriminate between main and supporting information. Summary tasks may therefore benefit from an *upper* word count rather than a minimum word count to ensure idea selection and synthesis rather than text reproduction and paraphrasing.

Regarding scoring, analytic rating scales are generally preferred over holistic scales due to the complexity of cognitive processing involved in task completion. Developers must decide whether source use is a separate scale or integrated into descriptors for other rating scale components. Lestari and Brunfaut (2023) compared the use of a scale with a separate criterion for reading/source use with a scale which integrated source use with writing criteria, finding that both scales functioned well, but the separate criterion offered greater transparency to raters.

To date, existing tests of English used for university admission or professional purposes have largely eschewed intertextual reading (Owen 2016; Weir and Chan 2019) and do not ask test takers to synthesize information from multiple texts into a single piece of writing. The brief discussion here demonstrates that developing such a task represents a significant challenge, complicated by the requirement of developing carefully controlled input, identifying explicit task requirements (i.e., what test takers are required to do with the input), and how to offer support to assessors who must navigate between task requirements, input texts and the test taker's response. This complexity likely explains the dearth of such tasks. However, this omission has resulted in the proliferation of English language tests which inadequately reflect real-world language use in academic and professional domains.

2.4 The 2022 Handbook's contribution to CEFR alignment

The 2022 Handbook provides updated guidance on aligning language tests to the CEFR, building on *Relating Language Examinations to the Common European Framework of Reference for Languages: A Manual* (Council of Europe 2009, hereafter referred to as the Manual) but incorporating developments from the 2020 CEFR Companion Volume. The Handbook emphasizes the importance of the examinee-centred approach for integrated skills assessment, which involves collecting test taker performance samples and scoring them using established systems by external participants. The Handbook's five-stage alignment process (familiarization, specification, standardization, standard setting, and validation) provides a structured approach to ensuring CEFR alignment, consistent with that found in the Manual. The examinee-centred approach is particularly relevant for integrated skills tasks, as it allows for the collection of authentic performance data to investigate the complex task completion processing involved.

3 Methodology

3.1 Research design and research questions

This study employed a mixed-methods approach combining quantitative analysis of test performance data with qualitative assessment of task design and validation procedures. The research design followed the five-stage alignment process outlined in the 2022 Handbook: familiarization, specification, standardization, standard setting, and validation. The data reported in this study represent the standard setting and validation phases of alignment.

The study addressed the following research questions:

RQ1: To what extent does the summary task demonstrate equivalent reliability to traditional essay tasks in measuring writing proficiency at CEFR levels B2-C1?

RQ2: How effectively do assessors score summary task responses using an analytic rating scale compared to traditional essay tasks?

RQ3: What is the concurrent validity between summary task performance and traditional essay task performance?

RQ4: To what extent does the summary task discriminate between B2 and C1 level performances according to CEFR standards?

RQ5: How does the difficulty level of the summary task compare to traditional essay tasks, and what factors contribute to any differences?

Data from the online CEFR alignment validation panel (n = 7) and subsequent large-scale pilot study (n = 215) were used to address the research questions. RQ1 and RQ2 were addressed using pilot study data, RQ3 was addressed using data from test takers who completed both a summary and an essay task in the pilot study, and RQ4 and RQ5 were addressed using CEFR alignment validation data from the panel.

3.2 Task design

The summary task was designed using CEFR mediation descriptors at B2 and C1 in addition to detailed domain analysis and the recommendations of the 2022 Handbook. Test takers are presented with two texts on the same topic (300 words total) and required to synthesize information into a summary. The task parameters and features include:

- Two source texts of different genres, one textbook extract and one lecture transcript (approximately 150 words each)
- Clear instructions emphasizing synthesis rather than reproduction
- A 100-word limit to ensure idea selection and transformation
- A glossary of low-frequency lexis to support comprehension
- 20-minute time limit including both reading and writing time

For more detail regarding task design and scoring, please see the online test specifications (Oxford University Press 2025).

3.3 Participants

A total of 665 test takers participated in the pilot study. A total of eight assessors marked the Speaking and Writing test scripts. 314 of the test takers received scores from a minimum of two assessors. However, of these 314, only 215 received marks for both essay and summary writing tasks from a minimum of two assessors. As a result, assessor data presented in the findings (RQ2) is the output of analysis of the 314 test takers. Concurrent validity and reliability research questions (RQ1, RQ3 and RQ5) are addressed using data from the 215 test takers for direct comparability. Piloting took place in Spring 2024. The pilot study collected data across multiple test administrations in Spring 2024, with detailed findings reported in the Oxford University Press pilot study report (Owen 2024b).

As part of ethical approval, participants were not required to complete biodata entries as a condition of participation. As a result, biodata collection was partial. Of the 215 test takers reported for reliability and concurrent validity research questions, 185 responded to the biodata questions. The first language reported by the cohort is dominated by Turkish (86) and Spanish (Castilian) (57), followed by Italian (16), Arabic (9), German (4), Portuguese (3) and one each for Catalan-Valencian, Portuguese, Czech, Kurdish, Luxembourgish, Dutch-Flemish and English (one teacher participant). Gender distribution shows 104 females, 78 males, and 3 “prefer not to say”. Ages range from 7 to 66 years (mean ≈ 22.8), with most participants between 16 and 25 years. Test takers were recruited through test centres and selected based on expert judgment of their B2–C1 level proficiency carried out by teachers at their respective test centres.

For the CEFR alignment validation panel, seven assessors participated in the study. The panel consisted of five male and two female assessors, all English L1 speakers with extensive experience in language assessment. Three assessors held PhDs in language testing, while the remaining four possessed extensive experience in English language teaching, item writing, and materials development. All standard setting activities were undertaken by assessors independently and all assessors received the same standard setting materials. Samples were selected from pretesting, which had occurred in summer 2022. Panel members judged 48 samples of writing (24 essay and 24 summary responses) across four pretests, with each pretest containing an essay and a summary task (six responses per pretest). The CEFR alignment panel and pilot study used the same analytic rating scale with four components: Task fulfillment, Organization, Grammar, and Lexis. Each response received four scores (maximum score = 28). Standard setting data is used to address RQ4. The CEFR alignment validation procedures and results are documented in detail in the CEFR alignment report for *Oxford Test of English Advanced* Writing and Speaking modules (Owen 2024a).

3.4 Data analysis

Data for both the pilot study and the CEFR alignment were analysed using many-facet Rasch measurement (MFRM) using a Rasch-Masters partial credit model within the program FACETS v3.84 (Linacre 2023). A five-facet model was adopted (assessors, test takers, task, pretest, component) with an eight-point rating scale (0-7). Reliability was assessed using the intraclass correlation coefficients, rater agreement, separation and strata output from the FACETS analysis.

4 Findings

4.1 RQ1: Task performance and reliability

The summary task demonstrated strong reliability across both pilot study and CEFR alignment validation. The summary task achieved a reliability value of .87, exactly equal to the traditional essay task (.87) (n = 215), indicating that the integrated nature of the task does not compromise measurement reliability, supporting the viability of intertextual reading-into-writing tasks in high-stakes assessment contexts. Inter-rater agreement was 35.1% for exact agreements, which is consistent with expected levels for subjective assessments and identical to the 35.1% agreement achieved for essay tasks. The separation and strata indices (8.41 and 11.54 respectively) for the summary task indicated strong proficiency differentiation by assessors while maintaining consistency in rating standards. These values are comparable to those achieved for essay tasks (12.40 and 16.87), suggesting that assessors applied rating criteria to integrated skills tasks with similar consistency to traditional essay writing tasks.

4.2 RQ2: Assessor performance

Many-facet Rasch analysis revealed strong assessor performance for the summary task. Table 1 presents the assessor performance statistics for Writing Script 2 (Summary task) from the pilot study data.

Table 1. *Assessor performance statistics (N.B. data for n = 314 test takers)*

Assessor	T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	OutfitMS	PtMea	Discrim
1	2800	1256	2.23	2.21	.92	.04	.89	.88	.81	1.11
2	2769	820	3.38	3.15	-.72	.04	1.36	1.36	.84	.61
3	2240	852	2.63	2.43	.48	.05	.94	.93	.83	1.06
4	1708	428	3.99	3.33	-.99	.06	1.09	1.25	.87	.71
5	1867	600	3.11	2.69	.02	.05	.91	1	.86	1.06
6	2829	1164	2.43	2.42	.51	.04	.90	.93	.84	1.07
7	147	68	2.16	2.51	.33	.17	.73	.71	.78	1.29
8	1809	528	3.43	3.04	-.55	.05	.81	.79	.86	1.23
Model, Populn: RMSE .08 Adj (True) S.D. .63 Separation 8.41 Strata 11.54 Reliability (not inter-rater) .99										
Model, Sample: RMSE .08 Adj (True) S.D. .68 Separation 8.99 Strata 12.33 Reliability (not inter-rater) .99										
Model, Fixed (all same) chi-squared: 1436.4 d.f.: 7 significance (probability): .00										
Model, Random (normal) chi-squared: 7.0 d.f.: 6 significance (probability): .33										
Inter-Rater agreement opportunities: 12920 Exact agreements: 4530 = 35.1% Expected: 4685.8 = 36.3%										

The analysis shows that assessors were able to score summary responses effectively using the analytic rating scale. Most assessors demonstrated good fit statistics, with Infit and Outfit mean square values close to the ideal range of 0.7-1.3. Assessor 2 showed slightly higher fit values (1.36 for both Infit and Outfit), indicating some inconsistency in rating patterns, while

Assessor 7 demonstrated excellent fit (0.73 and 0.71 respectively) despite having a smaller sample size. Point-measure correlations ranged from 0.78 to 0.87, indicating strong correlation between assessors' ratings and expected ratings. Discrimination values ranged from 0.61 to 1.29, with most assessors achieving values above 1.0, indicating effective differentiation between different levels of performance.

4.3 RQ3: Concurrent validity

To assess concurrent validity, we compared performance on the summary task with traditional essay tasks. Figure 1 shows the relationship between summary and essay scores for test takers who completed both tasks and received scores from more than one assessor. Total scores represented are summed from fair mean averages output from Rasch-Masters partial credit FACETS analysis of the analytically scored dataset.

The scatter plot shows the relationship between scores on Writing Script 1 (Essay) and Writing Script 2 (Summary) for n = 215 test takers in the pilot study who completed both tasks. The r-squared value indicates that approximately 64% of the variance in Essay scores can be explained by the variance in Summary scores. A correlation of .80 suggests that the summary task measures similar underlying writing ability to traditional essay tasks, while also capturing additional skills related to reading comprehension and information synthesis. Test takers generally received slightly lower scores for the summary task compared to essay tasks. This pattern is visible in the scatter plot, with the regression slope and most data points falling below the ideal x = y line (marked in red), indicating that test takers typically scored higher on essay tasks than summary tasks. This difference is attributed to the additional

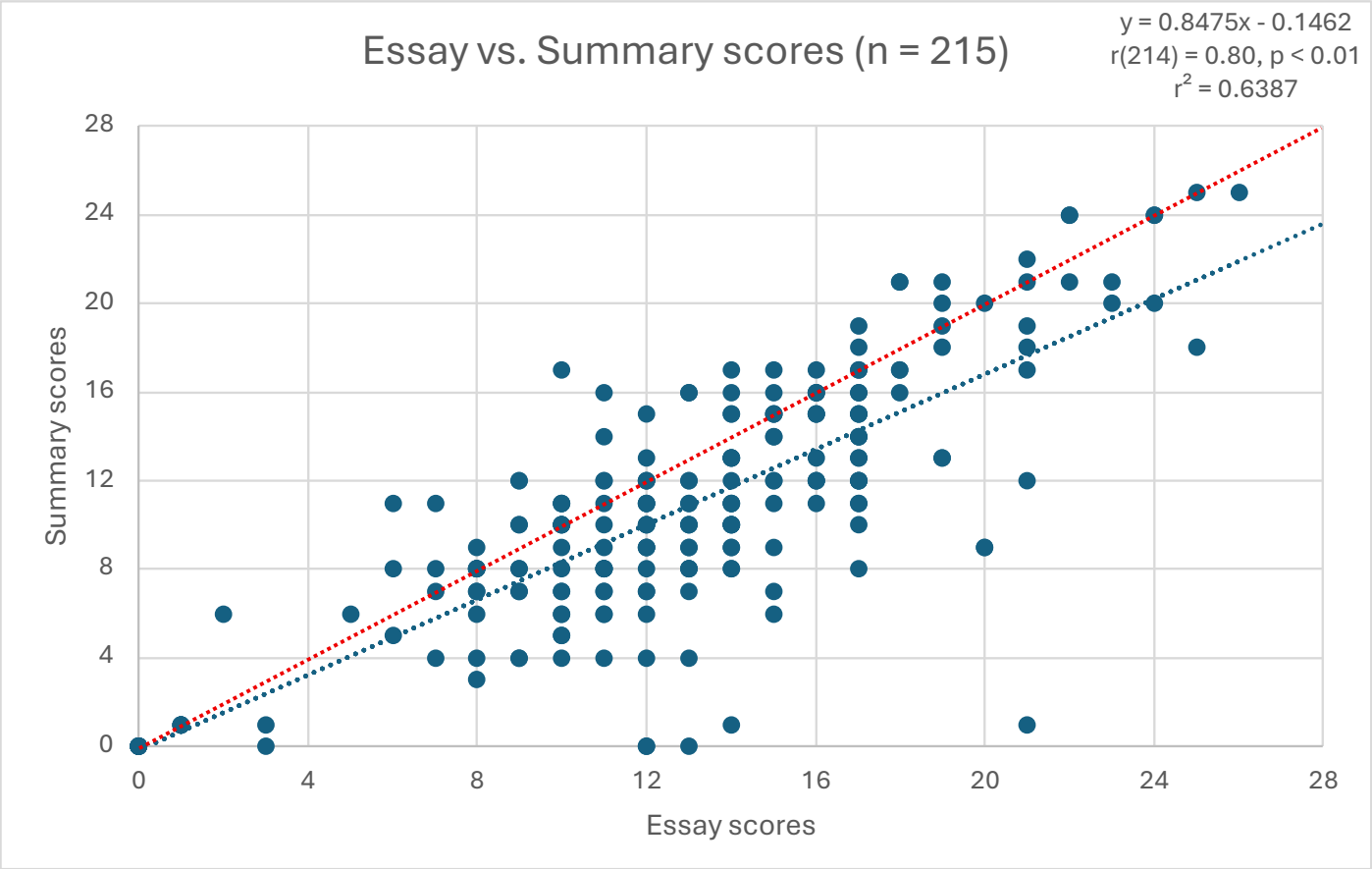

Figure 1. Essay vs. Summary scores for pilot participants (n = 215)

cognitive demands of processing multiple sources and synthesising information, as well as test takers' relative unfamiliarity with the integrated task type.

4.4 RQ4: CEFR level discrimination

Data from the CEFR alignment panel activities showed that the summary task effectively discriminated between B2 and C1 level performances. Forty-eight sample performances (24 essay and 24 summary) which had been marked internally were used in the CEFR standard setting activity. Eighteen out of 24 summary samples were awarded the same CEFR band by panel assessors as awarded during internal pretesting, demonstrating strong alignment with CEFR standards. The correlation between internal and external assessor scores was .94 for the summary task, indicating very strong agreement between different rating panels. The overall level of agreement for awarding CEFR bands was .77. The CEFR alignment validation showed that the summary task can be used to effectively distinguish between levels B2 and C1. The findings suggest that the task successfully measures the intended CEFR levels and that assessors can consistently apply CEFR standards to summary task responses.

4.5 RQ5: Task difficulty analysis

Rasch analysis revealed that the summary task was approximately one logit more challenging than traditional essay tasks. The difficulty measures for the two tasks were 0.43 logits for the essay task and -0.43 logits for the summary task, representing a difference of approximately 0.86 logits. This difference is statistically significant ($\chi^2 = 212.3$, d.f. = 1, $p < .001$). The increased difficulty of the summary task is attributed to the additional cognitive demands of processing multiple sources and synthesizing information into a single test. Also, test takers' relative unfamiliarity with the integrated task type may

contribute to the increased difficulty, as they lack experience with this type of assessment format. The difficulty difference was consistent across different administrations and assessor panels, suggesting that it reflects a genuine difference in task complexity rather than measurement error. This finding supports the validity of the summary task as a measure of advanced language proficiency, as it successfully differentiates between different levels of ability within the B2-C1 range.

5 Discussion and conclusions

5.1 Implications for test design

The success of the summary task demonstrates that integrated skills assessment can achieve reliability levels equivalent to traditional essay-style tasks. The task's effectiveness in discriminating between B2 and C1 levels supports its use in CEFR-aligned assessment contexts. The additional cognitive demands of the task, reflected in its higher difficulty level, may provide better measurement of advanced language proficiency by requiring test takers to demonstrate higher-order processing skills. The success of the analytic rating scale with integrated source use criteria suggests that separate source use scales may not be necessary for integrated tasks. This finding aligns with the recommendations of Lestari and Ho (2023) and provides practical guidance for rating scale development.

5.2 Contribution to CEFR alignment methodology

This study demonstrates the utility of the 2022 Handbook in guiding CEFR alignment methodology, particularly in the context of online alignment activities. The Handbook's examinee-centred approach proved particularly valuable for validating integrated skills tasks, as it allowed for the collection of authentic performance data that reflects the complex nature of real-world language use.

The five-stage alignment process outlined in the Handbook (15-17) provided a structured approach to ensuring CEFR alignment that was well-suited to fully online implementation. The Handbook's concise format, compared to the more extensive 2009 Manual, made the alignment process more navigable and manageable for participants. The Handbook's excellent organization and cross-referencing with the Companion Volume and Manual (8-9) proved particularly valuable during the alignment process. When participants required additional information about specific stages or procedures, they could easily locate relevant sections in the supporting documents, ensuring that the alignment process remained comprehensive despite the Handbook's more concise format.

5.3 Online alignment implementation

All CEFR alignment activities in this study were conducted online through Microsoft Teams (Microsoft Corporation n.d.) webinars, representing a significant departure from traditional face-to-face alignment procedures. This online approach offered several advantages that enhanced the alignment process. The Teams platform facilitated easy collection and distribution of materials through well-organized Teams folders, ensuring that all participants had consistent access to necessary documents and resources throughout the alignment process.

The online format also mitigated potential group dynamics issues that can occur in face-to-face settings. Individual participants working remotely were less likely to encounter pressure from more experienced team members and simply defer to their expertise, i.e., conformity bias or groupthink (Janis 1972). This reduced the risk of dominant personalities influencing group decisions and ensured that all participants could contribute equally to the alignment process. The asynchronous nature of some online activities allowed participants to work independently on rating tasks before coming together for discussion, further reducing the potential for group pressure to influence individual judgments.

5.4 Limitations of the Handbook and future directions

While noting that the layout of the Handbook lends itself well to online alignment processes, the 2022 Handbook has several limitations that became apparent during the online alignment process. The Handbook provides limited information about conducting online alignment procedures and how these will or should differ from face-to-face alignment panels, despite the increasing prevalence of remote work and virtual collaboration in language assessment. This gap in guidance may prove challenging organizations seeking to conduct future alignment activities in online environments. Future editions of the Handbook should seek to expand information around online alignment processes.

The Handbook also offers limited information about the impact of technology on alignment processes. New approaches such as AI or LLM-informed alignment and adaptive comparative judgment methods are emerging as potential alternatives to traditional alignment panels. The Handbook would benefit from addressing technological developments and their implications for future CEFR alignment methodology. Additionally, the Handbook provides limited technical support for those seeking to align to the CEFR. The establishment of the CEFR as an online API could significantly enhance alignment processes by providing standardized access to CEFR descriptors and facilitating automated alignment procedures. Such an API could enable real-time validation of alignment decisions and provide immediate access to relevant CEFR documentation, potentially streamlining the alignment process and reducing the time required for comprehensive alignment activities.

6 References

British Council, UKALTA, EALTA, & ALTE. 2022. *Aligning language education with the CEFR: A handbook*. In Neus Figueras, Barry O'Sullivan, Nick Saville, Lynda Taylor, & David Little (eds.). http://www.ealta.eu.org/documents/resources/CEFR%20alignment%20Handbook.pdf (accessed 20 Nov 2025).

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. 2009. *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual*. Strasbourg: Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr (accessed 20 Nov 2025).

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

Crossley, Scott A., Qian Wan, Laura K. Allen & Danielle S. McNamara. 2023. Source inclusion in synthesis writing: An NLP approach to understanding argumentation, sourcing, and essay quality. *Reading and Writing* 36. 105-1083. https://doi.org/10.1007/s11145-021-10221-x.

Figueras, Neus, David Little & Barry O'Sullivan. 2022. Aligning language education with the CEFR: A handbook. *CEFR Journal, 5,* 1-10. https://doi.org/10.37546/JALTSIG.CEFR5-1

Gebril, Atta & Lia Plakans. 2014. Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing* 21. 56-73. https://doi.org/10.1016/j.asw.2014.03.002.

Janis, Irving L. 1972. *Victims of groupthink: A psychological study of foreign-policy decisions and fiascos.* Boston: Houghton Mifflin.

Lado, Robert. 1961. *Language testing: The construction and use of foreign language tests.* New York: McGraw-Hill.

Lestari, Santi Budi & Tineke Brunfaut. 2023. Operationalizing the reading-into-writing construct in analytic rating scales: Effects of different approaches on rating. *Language Testing* 40(3). 684-722. https://doi.org/10.1177/02655322231155561.

Li, Jiuliang. 2014. Examining genre effects on test takers' summary writing performance. *Assessing Writing* 22. 75-90. https://doi.org/10.1016/j.asw.2014.08.003.

Linacre, John M. 2023. *Facets* (Version 3.71.4) [Computer software]. https://www.winsteps.com/facets.htm (accessed 20 Nov 2025).

Microsoft Corporation. n.d. *Microsoft Teams* [Computer software]. Retrieved from https://www.microsoft.com/en-us/microsoft-teams/group-chat-software (accessed 20 Nov 2025).

Nelson, Nancy & James R. King. 2022. Discourse synthesis: Textual transformations in writing from sources. *Reading and Writing* 35. 769-808. https://doi.org/10.1007/s11145-021-10243-5.

O'Sullivan, Barry. 2020. *The Comprehensive Learning System.* London: British Council.

Owen, Nathaniel. 2016. *An evidence-centred approach to reverse engineering: Comparative analysis of IELTS and TOEFL iBT reading sections*. University of Leicester PhD thesis.

Owen, Nathaniel. 2024a. *Oxford Test of English Advanced CEFR alignment report: Speaking and Writing.* https://fdslive.oup.com/www.oup.com/elt/general_content/global/ote/4-ref-0011-cefr-alignment-report-sandw-for-website.pdf (accessed 20 Nov 2025).

Owen, Nathaniel. 2024b. *Oxford Test of English Advanced pilot study report.* https://fdslive.oup.com/www.oup.com/elt/general_content/global/ote/4-ref-0030-pilot-study-report-2024-for-website.pdf (accessed 20 Nov 2025).

Oxford University Press. 2025. *Oxford Test of English Advanced test specifications.* Oxford: Oxford University Press. https://fdslive.oup.com/www.oup.com/elt/general_content/global/ote/oxford-test-of-english-advanced-test-specifications.pdf (accessed 20 Nov 2025).

Plakans, Lia & Atta Gebril. 2012. A close investigation into source use in integrated second language writing tasks. *Assessing Writing* 17(1). 18-34. https://doi.org/10.1016/j.asw.2011.09.002.

Sawaki, Yasuyo, Lawrence J. Stricker & Andreas H. Oranje. 2009. Factor structure of the TOEFL Internet-based test. *Language Testing* 26(1). 5-30. https://doi.org/10.1177/0265532208097335.

Weigle, Sara Cushing, Weiwei Yang & Megan Montee. 2013. Exploring reading processes in an academic reading test using short-answer questions. *Language Assessment Quarterly, 10(1),* 28-48. https://doi.org/10.1080/15434303.2012.750659.

Weir, Cyril J. & Sathena Hiu Chong Chan. 2019. Trends in language assessment research and practice: The view from *Language Testing* 1986–2016. *Language Testing* 36(3). 349-363. https://doi.org/10.1177/0265532219826396.

Yu, Guoxing. 2013. From integrative to integrated language assessment: Are we there yet? *Language Assessment Quarterly* 10(1). 110-114. https://doi.org/10.1080/15434303.2013.766744.

7 Biographies

Nathaniel Owen is Senior Research and Analysis Manager at Oxford University Press. He holds a PhD in language testing from the University of Leicester specializing in L2 reading processes. His research interests and publications include the interface of language testing and technology, developing integrated-skills tasks, big data analytics, the use of language tests in English-medium instruction contexts, research methods and widening participation in higher education.

Oliver Bigland holds an MA in Applied Linguistics from the University of Birmingham. His research interests include the design and evaluation of integrated skills tasks, the role of functional language in speaking assessments, and the identification and mitigation of bias in language testing. He is also interested in the practical application of Rasch measurement theory and computational methodologies, particularly Python-based data analysis, in the context of language assessment.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

The CEFR in Cuba: Alignment endeavours for English certification in Cuban higher education

Claudia Harsch, University of Bremen
Yoan Martínez Márquez, University of Informatics Sciences, Cuba

https://doi.org/10.37546/JALTSIG.CEFR8-4

We report CEFR alignment endeavours for the certification of English in Cuban higher education. In 2015, the Ministry of Higher Education (MES) introduced a new language policy, employing the CEFR as a proficiency framework, and setting CEFR level B1 as the exit standard. The MES wanted to introduce a CEFR-aligned certification system to teach and assess students' English proficiency in order to achieve international recognition.

We present achievements and outcomes from the project's first phase, which encompassed a) familiarization and training for 42 representatives of all HE language centres in Cuba, b) developing test specifications and tasks by the trained representatives, c) piloting the tasks on a small scale, followed by d) standardization and benchmarking of local examples. All available information has been reported in a local handbook, and all data have been compiled in a database to support the piloting and the formal standard setting, which will be conducted in the second project phase, recently confirmed by the MES.

1 The educational context in Cuba

Since 2015, the Ministry of Higher Education (MES) in Cuba has been promoting a paradigmatic change in the teaching and learning of the English language. The CEFR was introduced as a globally accepted proficiency framework, with CEFR-level B1 being set by the MES as the exit standard for all non-major BA university programs. The CEFR was adopted as a proficiency framework in order to achieve international recognition of the planned certification, and level B1 was chosen from a pragmatic perspective, as an achievable level within a BA study program.

This reform required changes to the curriculum, teaching and assessment practices. In 2017, a project was launched to develop a certification system for teaching and assessing students' English proficiency. Partners in this international endeavour are MES, Universidad de las Ciencias Informáticas (UCI) and the University of Bremen, Germany with support from the British Council, the International Language Testing Association (ILTA), and the Belgian VLIROUS network.[1]

Seeking international recognition of the new certification system via alignment to the CEFR, training needs in the following areas were identified: familiarization with the CEFR, curricula development in line with the CEFR, and high-stakes assessment aligned to the CEFR. The new *Teaching and certification system in EAP in Cuban HE for BA students* is composed of three main stages, which integrate teaching, formative and summative assessment, and final certification. Taking a placement test in the first year of their major, students are assigned to suitable courses. If they obtain level B1 in the placement test,

1. The Belgian capacity building network VLIROUS is "the leading funding body for scholarships for and partnerships between academics from Flanders and partner institutions in Africa, Latin America, and Asia, focused on global sustainable development" (www.vliruos.be/en).

students can directly take the certification exam. A mentoring process is opened for those students who request it or who fail the final certification.

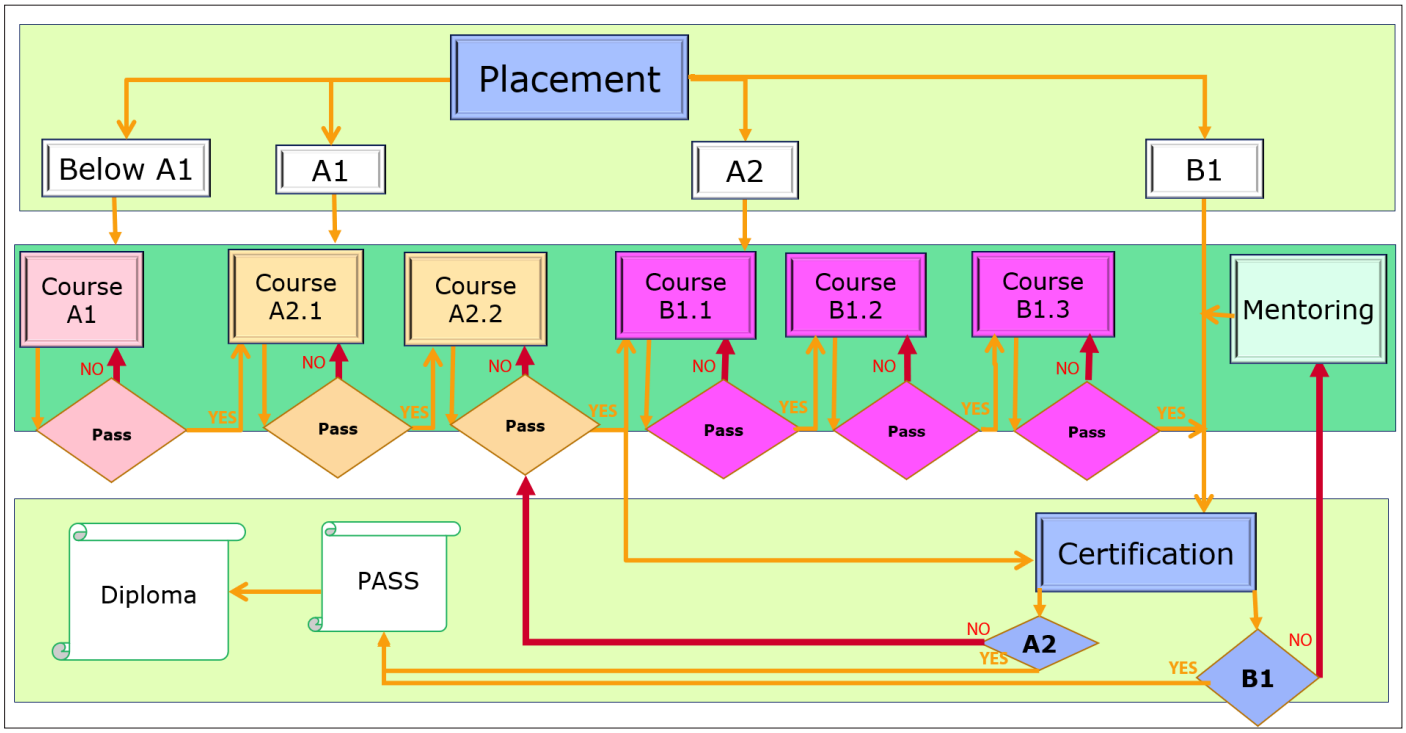

Figure 1. Design of the Teaching and Certification System in EAP in Cuban HE for BA students

The teaching process was designed to operationalize the targeted CEFR-levels from A1 to B1, and was intended to offer maximum flexibility to students, to match their individual progress. At this point in time, the certification system is kept open for an additional certification of the A2 level for those students who may not reach level B1. The final decision on whether to certify both levels or only the B1 level is yet to be taken. In the Cuban context, the teaching system is organized into 90-hour courses over one semester. CEFR level A1 can be achieved in one course, CEFR level A2 can be achieved in two courses, and CEFR level B1 in three courses. The learning outcomes for each course are based on selected and adapted CEFR/CV descriptors that were considered relevant for the project context, and the learning and teaching approaches follow the learner-centred, action-oriented approach of the CEFR. Through mentoring, at any time of the learning process, teachers can assess whether students would benefit from attending another course level or whether they are ready to take the final certification exam. This flexibility is one of the features of the new MES policy that teachers and students value the most.

2 Development of the new proficiency exam

In this article, we focus on developing the new proficiency exam, which is placed at the end of the learning and teaching process. We report on the first phase of the exam development project, which took place between 2017 and 2022, and which encompassed 1) familiarization and training for teachers representing all HE language centres in Cuba, 2) developing test specifications and tasks, 3) piloting the tasks on a small scale, followed by 4) standardization and benchmarking of local examples.

All available information is reported in a local handbook (Collada et al. 2023), and all data collected so far are compiled in a database to support the piloting and the formal standard setting, which will be conducted in the second project phase, recently confirmed by the MES. Table 1 provides the overall plan of the project.

Table 1. *The overall plan of the project in Cuban higher education*

Steps	Timeline
Planning	2016–2017
Training	2017–ongoing
Developing test specifications, tasks, rating scales	2017–2019
Feedback in the team, external feedback	2018–2020
Pre-testing and benchmarking	2020–2023 => task revision
Compiling an item bank	2020–ongoing
Piloting, IRT scaling	[2026]
Alignment to CEFR	[2027]
Implement proficiency exam	[tbd]
Monitor the impact on teaching and learning	[tbd]

2.1 Project design: the CEFR in Cuban HE

The project to develop a CEFR-aligned exam was designed cooperatively between the former project lead in Cuba (current project lead: co-author) and the external advisor (author) to encompass three synergetic strands, combining teacher training, exam development, and research. The training and exam development process was a collaborative and responsive project that was evaluated on an ongoing basis and adjusted according to the feedback and needs of the local participants (Harsch et al. 2021, on the design and its evaluation). One of the aims was to train teachers from all language centres across Cuba, to enable them to become local trainers, mediating the outcomes and skills to their peers in the local language centres.

The following local training needs were identified before the first workshop. Participants needed to be familiarized with the CEFR, its approaches and conceptualizations, as well as its scale system; as assessment literacy was not part of formal teacher education, the training should also encompass an introduction to different assessment purposes and suitable instruments for different purposes, such as placement testing, diagnostic assessment, assessment for achievement purposes and proficiency exams. Furthermore, an introduction to the theory and practice of language assessment, constructs of relevant communicative skills, the role of linguistic competences, communicative task formats, and basics of constructive alignment was considered a necessary part of the introductory training. Building on this foundation, the training then aimed at developing practical skills in designing test specifications, selecting suitable test formats and developing communicative test tasks.

The training was designed in such a way that one-week-long workshops were delivered with participants attending face-to-face meetings in Habana (during the pandemic, we had to switch to online delivery). The workshops were followed by working in groups, where teachers in their local regions collaborated on practical task development, giving each other feedback. The tasks were then cooperatively evaluated at the beginning of the next workshop. We collected feedback after each workshop to design the following one based on participants' needs.

In the first workshop, the teachers were familiarized with the CEFR, its approaches, the philosophy behind it and its scale system, while the external advisor (author) was familiarized with the Cuban context, the teaching and learning system, the local constructs and teaching/assessment task formats, the students' characteristics, and the general HE system. In this phase, it became transparent that in Cuba, the four skills of listening, reading, speaking and writing are traditionally taught in the language classroom, and the new certification system should focus on targeting these four skills.

Overall, nine workshops and ensuing working group phases were conducted, covering the following aspects:

- theories and practice of assessing the four skills
- developing test specifications for the four skills and targeted levels A2 and B1 (the test development process should cover these two levels, to be able to create exams for both levels, and to also certify students who may not reach B1)
- developing suitable task formats targeting the different levels and skills
- developing rating scales for (interactive and productive) speaking and writing, based on selected CEFR/CV scales (Harsch et al. 2020)
- trialling and benchmarking local performances
- basics of item analysis and reporting
- planning, conducting, and analysing a pilot study
- task and item revisions based on statistical analyses.

We aimed to constructively align curriculum, classroom practice and assessment. To reach this alignment, teacher participants brought their practical classroom perspective and experience into the development of test specifications, the selection of suitable task formats, as well as the decisions on suitable topics and inputs for the assessment. Not only did we take classroom expertise into account when selecting and adapting CEFR/CV descriptors, we also took into account the newly revised curriculum for higher education, which was also adopting the CEFR and its descriptors, while building on existent traditions. While the curriculum development took place in a different project, we had regular contact between the two projects, with some participants being active members of both project teams. The curriculum was finalized in September 2019 (Casar Espino et al. 2019), but even during the development phase of the curriculum, we used drafts of this revised curriculum as a complementary source next to the CEFR/CV descriptors for the test design and development.

We will now outline the achievements of the first project phase with regard to test development, piloting and benchmarking.

2.2 Project outcomes of phase 1

The project team developed test specifications for the Cuban standardized national exam in all four skills (reading, listening, speaking, and writing). The skills targeted in this exam are based on selected CEFR/CV scales and descriptors, on the recently revised Cuban University Curriculum (that also takes the CEFR into account), and the Pearson Global Scale of English that explicitly targets the academic context (Pearson Education 2022), and the rating scales of the IELTS academic exam, as it reflects relevant academic speaking and writing aspects, and its bands are aligned to the CEFR (IELTS 2013, IELTS 2016, IELTS 2018). Table 2 lists the general exam specifications.

Table 2. *Test specifications, general exam purpose*

Purpose	To serve as a certification of English language proficiency, a prerequisite for university graduation in the Cuban context.
Age groups	Mostly 18-24
Expected L1s	Mainly Spanish and Portuguese
Possible targeted situations	Four language skills in general and academic contexts Interaction English for international communication

Test structure	Four equally weighted sections (reading, listening, speaking, writing)
Targeted CEFR levels	A2 and B1
Topic areas	Mostly general, professional, or academic, accessible to a general audience; from concrete to mostly concrete; distressing topics avoided

The following excerpt from the reading test specifications for level B1 shows how the CEFR descriptors were adapted, and how the other sources such as the curriculum and the GSE were integrated to formulate the specific purpose of the exam:

The students can

1. *understand the gist of straightforward factual texts related to topics students are familiar with* [CEFR scale OVERALL READING COMPREHENSION, level B1; Cuban curriculum]
2. *scan longer general as well as professional/academic texts accessible to a general audience to locate information to solve specific tasks* [CEFR scale READING FOR ORIENTATION, level B1; Cuban curriculum]
3. *understand main ideas and supporting details in general, professional, or academic texts accessible to a general audience* [GSE, Cuban curriculum]
4. *...*

Furthermore, the project team developed item writer guidelines for the four skills, interlocutor guides for conducting the speaking exam, a set of standardized instructions for reading and listening tasks, as well as 150 test tasks in total (reading 52, listening 21, writing 47, speaking 30). All tasks were specified with a task specifier that includes the targeted competences (from the specific purpose in the test specifications), characteristics of input, expected output, task formats, and expected duration, along with the answer key.

For speaking and writing, rating scales were developed that were also based on the CEFR and its Companion Volume. The development and the accompanying challenges are reported in Harsch et al. (2020). The scales were revised in several rounds and finalized after standardization and rater training within the project group. For this endeavour, the group collected performances in small-scale trials, which also served as a basis for selecting benchmarks to illustrate all criteria and levels.

Table 3 illustrates the criteria for the rating scale for assessing writing, showing the defining descriptors for level B1; note that the colours indicate the sources underlying the descriptor wording; red indicates wording from the CEFR/CV descriptors, blue indicates the revisions after the first trial, green indicates the IELTS writing scale (IELTS, 2013, 2016).

The rating scales, along with the task types used in the exam, have been introduced to regular classroom teaching and have been applied for more than two years in teaching and assessment practice all over the country, in order to familiarize teachers and students with the approaches well before the actual exam is introduced.

In May 2022, a pre-pilot study with a selection of tasks for all four skills took place to gauge the possibility of a representative pilot study with participants from across Cuba. This also served to pilot feasible procedures for data collection, data coding, rating, and quantitative and qualitative data analyses. Moreover, we introduced basic statistics to the teachers and trained the group in interpreting item and task analyses, with a view to revising tasks and items in preparation for the representative pilot study that is to take place in Phase 2 of the project.

Table 3. *Rating scale for writing, level B1 (taken from Harsch et al. 2020: 96)*

	Task Fulfilment	Coherence / cohesion	Vocabulary (range & appropriateness)	Grammar (range & accuracy)	Orthography (spelling & mechanics)
B1	The message is generally clearly conveyed. (CLAN) The ideas/content are generally relevant to the topic of the task. (CLAN) Performs most of the language functions required by the task (e.g., comparing, describing, explaining, etc.) (Test specifications: 8 and adapted from CEFR/CV: 138). Mostly follows the conventions of the text type/format required by the task (CLAN), *but the format may be inappropriate in places* (IELTS band 5). Shows awareness of the required register, *but may still be inconsistent in tone* (IELTS band 6). Generally follows salient politeness conventions, but not always appropriately (adapted from CEFR/CV: 138)	Mostly organizes ideas into a meaningful sequence, with adequate topic progression (TS, GE). Makes simple, logical paragraph breaks if required by task. (adapted from CEFR/CV: 142) Links a series of shorter, discrete simple elements into a connected, linear sequence of points by using a limited number of cohesive devices (adapted from CEFR/CV: 142)	Uses sufficient topic-specific vocabulary to express themselves on familiar topics. (adapted CEFR/CV: 132) Shows appropriate use of a wide range of basic, frequent vocabulary. (adapted from CEFR/CV: 134) *Major errors may still occur when expressing more complex thoughts*. (adapted from CEFR/CV: 134) *May use circumlocution and occasionally unclear expressions.* (adapted from CEFR/CV: 131, 174)	Uses a range of simple grammatical features and sentence structures with reasonable accuracy. (adapted from CEFR/CV: 133) Attempts a limited range of complex sentence structures or complex grammatical features, *though they may usually be incorrect*. (adapted from IELTS band 5) *In general, the reader can interpret the errors correctly based on the context*. (adapted from CEFR/CV: 174)	Produces generally intelligible spelling for most common words, *mother tongue influence is likely with less common words*. Punctuation is enough to be followed most of the time, but *mother tongue is likely to influence punctuation*. (adapted from CEFR/CV: 137)

3 Constructive alignment and next steps

Over the whole process, we aimed at constructively aligning curriculum, classroom practice and assessment, following Little and Erickson (2015). The learning and teaching objectives and the competencies to be assessed were CEFR-informed, and we took both the CEFR and the learning objectives depicted in the curriculum into account when developing the test specifications. Concerning implementing the new curriculum and assessment approaches in classroom activities, the trained teachers are acting as mediators and are delivering training on assessment literacy, assessment procedures, new task formats and the use of the new rating scales on an ongoing basis, so that all Cuban teachers are gradually familiarized with new approaches and task formats, and can explore the tasks and rating scales with their students in an informal way, well before the certification exam is in place.

With regard to the formal alignment process stated in the CEFR alignment handbook (British Council et al. 2022: 14), we have covered the stages of familiarization, specification and standardization. Familiarization was ensured through the aforementioned workshops, specification was achieved via the test and task specifications, while standardization was targeted via several standardization and benchmarking workshops that also served as rater training, to ensure a common understanding of the CEFR levels and to benchmark local performance samples to relevant CEFR levels. All available data have been entered in a database, statisticians are currently trained, and all tasks are currently pre-trialled in classrooms to finalize them for the pilot study.

In Phase 2 of the project, we aim to conduct a formal standard setting, to align the exam, its tasks, performances and results to the CEFR. This will involve both local stakeholders, such as policy decision-makers, language centre directors and teachers, and international CEFR experts. Furthermore, we will investigate the validity of the exam and the standard-setting procedure from internal and external perspectives. Several PhD projects in this realm are currently in planning within the Cuban project team.

4 Conclusion

We would like to conclude this contribution with insights that we gained from the project experiences. We will outline challenges when adapting CEFR/CV descriptors and how we dealt with them, wider implications for adapting the CEFR, and general recommendations when planning such an endeavour.

We exemplify the challenges when adapting CEFR/CV descriptors for our rating scale development, where we chose a descriptor-based approach, as reported in Harsch et al. (2020). We selected relevant descriptors from the CEFR/CV and other CEFR-related scales suitable for the HE context. In an initial intuitive approach, the teachers re-sorted the selected descriptors, and in ensuing empirical approaches, we undertook benchmarking exercises in several rounds. We encountered the following challenges, as outlined in depth in Harsch et al. (2020): there was an overwhelming abundance of scales at different places (e.g., the writing assessment grid is presented separately from the writing scales) in the CEFR/CV, which was challenging at a time when the searchable Excel-file containing all CEFR/CV descriptors was not yet available. We also found different categorisations in the CEFR/CV and our context; for instance, the writing assessment grid in the CEFR/CV contains the criteria of range, coherence, accuracy, description and argument, while the local assessment criteria combined range and accuracy for vocabulary and grammar. Furthermore, descriptors for the "plus" levels were not always provided. Finally, we encountered inconsistent wording across scales and/or across levels; for example, the nature and impact of errors is described in different ways in different scales, even when looking at the same level (for a detailed analysis, Harsch et al. 2020). These aspects are a challenge when the aim is to systematise selected CEFR/CV descriptors into a distinct locally shaped assessment criteria grid.

Our solution consisted of the following four main approaches, which are also described in detail in Harsch et al. (2020):

1. Reorganizing CEFR/CV descriptors into the local assessment criteria, to adapt CEFR/CV categorisations to the local context.
2. Adapting CEFR/CV descriptors (i.e., changing wording), to overcome the aforementioned inconsistencies in wording and make levels more coherent.
3. Adding descriptors from other sources, particularly for the plus levels, as the CEFR/CV has some gaps for the plus levels which we needed to fill for the rating scales.
4. Adding and adapting descriptors to account for the local context, both for criterion levels and plus levels.

With regard to implications when aiming to adapt the CEFR, we would recommend allowing all participants to get familiar with the CEFR and other relevant materials in their own ways and at their own pace; here, a collaborative approach with hands-on activities for familiarization seems most feasible. We also found that pre-selecting relevant scales for certain activities helps simplify complex endeavours such as rating scale development. We would also recommend making use of other sources relevant to the context in a complementary way and documenting all steps transparently.

To ensure constructive alignment, it was helpful to co-develop learning goals, curricula, and assessment goals across the different project teams; to simultaneously consider learning activities, teaching tasks, and assessment tasks; and to make use of the alignment handbook (British Council et al., 2022), its helpful explanations, activities, tools, guidelines for reporting, and practical suggestions.

With regard to overall project management, when aligning an exam to the CEFR, we would recommend (based on Harsch et al. 2021) familiarizing all participants with the context, with institutional conditions and constraints, and with already existing professional development initiatives. It is important to include all relevant stakeholder groups in all phases, from design to implementation, thus enabling them to bring their expertise to the table and to form communities of practice. We can only underscore the importance of scheduling sufficient time for changes to take place in teaching, learning, and assessment practices, so that all participants can get on board. It was helpful in our context to facilitate collaboration by providing sufficient resources and shared spaces online as well as face to face, and to combine hands-on experiential workshops (with a focus on practical outcomes and applicability) with seminars and lectures where theoretical underpinning can be provided as needed.

5 References

British Council, EALTA, UKALTA & 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).

Casar Espino, Liliana, Pedro Castro Álvarez, Cecilia Clemencia González Gonzáles & Lissette Rubio Mederos. 2019. *Guía para la estructuración por niveles de competencia comunicativa en inglés en la educación superior cubana, Niveles Básico uno, Básico 2 e intermedio* [Guide for the instruction towards levels of communicative competence in English in Cuban higher education, level basic one, basic 2 and intermediate]. Regulations of the Ministerio Educación Superior de Cuba, internal unpublished document.

Collada Peña, Ivonne, Yoan Martínez Márquez & Gillermo Manuel Negrín Ortiz (eds.). 2023. *Handbook for standardised proficiency test development in Cuban higher education*. Cuban Language Assessment Network (CLAN). Preliminary Version, 1st edn. Habana, Cuba: Ediciones Futuro.

Harsch, Claudia, Ivonne de la Caridad Collada Peña, Tamara Gutiérrez Baffil, Pedro Castro Álvarez & Ioni García Fernández. 2020. Interpretation of the CEFR Companion Volume for developing rating scales in Cuban higher education. *CEFR Journal* 3. 87-97.

Harsch, Claudia, Sibylle Seyferth & Salomé Villa Larenas. 2021. Evaluating a collaborative and responsive project to develop language assessment literacy. *Language Learning in Higher Education* 11(2). 311-342. https://doi.org/10.1515/cercles-2021-2020.

IELTS. 2013. IELTS TASK 1 Writing band descriptors (public version). The official and updated (May 2023) band descriptors for writing tasks 1 and 2 are now available at: https://ielts.org/cdn/ielts-guides/ielts-writing-band-descriptors.pdf (accessed 28 January 2026).

IELTS. 2016. IELTS TASK 2 Writing band descriptors (public version).The official and updated (May 2023) band descriptors for writing tasks 1 and 2 are now available at: https://ielts.org/cdn/ielts-guides/ielts-writing-band-descriptors.pdf (accessed 28 January 2026).

IELTS. 2018. IELTS Speaking: band descriptors (public version). https://ielts.org/cdn/ielts-guides/ielts-speaking-band-descriptors.pdf (accessed 28 January 2026).

Little, David & Gudrun Erickson. 2015. Learner identity, learner agency, and the assessment of language proficiency: Some reflections prompted by the *Common European Framework of Reference for Languages*. *Annual Review of Applied Linguistics* 35. 120-139.

Pearson Education. 2022. Global Scale of English Learning Objectives for Academic English. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/pearson-languages/en-gb/pdfs/gse/gse-resources/gse-learning-objectives-adult-academic-english.pdf (accessed 28 Jan 2026).

6 Biographies

Claudia Harsch is a professor at the University of Bremen, specializing in language learning, teaching, and assessment. She has worked in Germany and in the UK, and is active in teacher training worldwide. Her research interests focus on areas such as language assessment, language and migration, the development of language assessment literacy, and the implementation of the CEFR. Claudia is currently the immediate past president of the International Language Testing Association (president from 2023-2024), and was president of the European Association of Language Testing and Assessment from 2016 to 2019.

Yoan Martínez is a professor at the University of Informatics Sciences, Cuba. His research fields are English language teaching and assessment, and ICT in education. He has participated in training programs, internships, and the ongoing assessment literacy training series developed by Professor Claudia Harsh. His research interests focus on the sustainability of language learning assessment and the localization of international language standards in the Cuban education system. Yoan is currently the leader of the Cuban Assessment Project for Higher Education, and the quality assurance of the British Council Cuba Academic Committee in Higher Education for the project "InglésParaelDesarrollo".

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.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Making it work: On the alignment of work-oriented writing tasks with the CEFR

Sibylle Plassmann, telc GmbH, Germany

https://doi.org/10.37546/JALTSIG.CEFR8-6

This article explores the challenges and solutions involved in aligning workplace-oriented writing tasks with the Common European Framework of Reference for Languages (CEFR) within the context of the German Tests for Work (Deutsch-Tests für den Beruf, DTB) at levels A2, B1, B2, and C1. Developed by telc GmbH for the German Federal Ministry of Labour and Social Affairs, these standardized exams serve as final assessments in vocational language courses. The article details the process of defining learning objectives based on authentic workplace communication needs and the adaptation of CEFR descriptors to fit vocational contexts. It discusses the design of writing tasks that reflect real-world professional communication—emphasizing brevity, appropriateness, and mediation—and the establishment of rating criteria tailored to workplace requirements. The article also examines the standard-setting process, including expert workshops and calibration, and presents findings from a 2024 re-rating study that demonstrate improved alignment and fairness in assessment outcomes. The case study concludes with reflections on CEFR alignment in vocational language assessment, highlighting the need for ongoing standard validation, authentic task formats, and continuous rater training to ensure valid and reliable certification of workplace language competence.

1 Context

This article examines the challenges of aligning workplace-focused writing tasks with CEFR levels in standardized German examinations, the four German Tests for Work (Deutsch-Tests für den Beruf—DTB) A2, B1, B2 and C1. The German Tests for Work are an exam suite that serves as a final assessment of proficiency in vocational language courses with language level goals of A2 to C1. The exams were developed by telc GmbH on behalf of the German Federal Ministry of Labour and Social Affairs and under the technical guidance of the Federal Office for Migration and Refugees (BAMF). The four exams represent an integral component of the German federal programme of vocational language courses (*Berufssprachkurse*) providing learners with the opportunity to develop and refine their job-oriented language skills in German.

The programme of vocational language courses is based on course concepts (BAMF 2021a, 2021b, 20212c, 2021d), a catalogue of learning objectives (Bärenfänger et al. 2019), dedicated qualifications for teachers in these courses (telc 2020e), a choice of fit-for-purpose course materials, and the exams presented here (telc 2020a, 2020b, 2020c, 2020d). These are specially developed components designed to ensure both appropriate target group orientation and the high quality of the programme. Constructive alignment of all these components was one of the major aims in the development phase.

The test development process is fully documented in the Examination Handbook (Plassmann et al. 2021). The focus of this article will not be on the very complex project to develop the exam suite as a whole, but on the writing tasks in the German Tests for Work. The development of the writing tasks exemplifies general challenges and solutions found in test development and administration of CEFR-related tests for a vocational purpose. Figure 1 shows the project's general timeline.

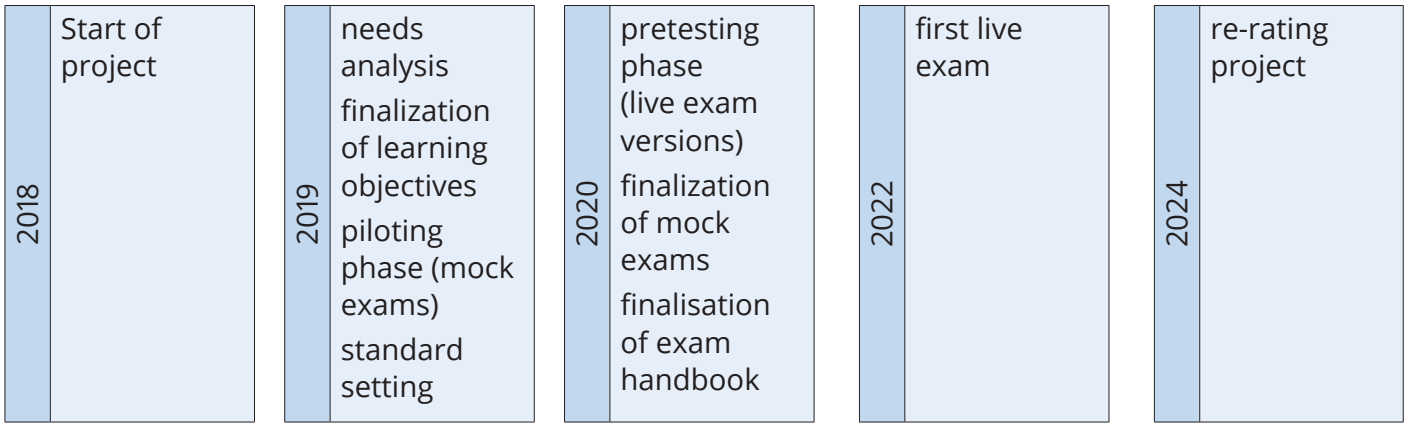

Figure 1. Timeline of test development and re-rating project

2 Learning objectives

The CEFR served as a comprehensive reference system for the project, with particular emphasis on the most relevant levels—A2, B1, B2 and C1. In particular, the CEFR was used as the overall framework for the German Tests for Work.[1] It was important to align them to the four levels (A2 to C1) as these levels were defined as goals for the respective language courses. From the start of the project, however, it was clear that CEFR descriptors would be helpful but not sufficient, as they do not reflect language use in the workplace in the necessary depth. Therefore, the first steps of test development were a needs analysis and the development of a Catalogue of Learning Objectives *A2-C1*. This catalogue provides a model of language competence in the workplace, based on an empirical analysis of language use and aiming at the description of real-world language use. Authentic communicative situations in the workplace are described and grouped along the employee life cycle (Bärenfänger et al. 2019: 20). The 344 detailed learning objectives are structured into eleven domains and 60 broad learning objectives. From the perspective of CEFR alignment the microstructure is most interesting: each of the 344 learning objectives has language functions from A2 to C1 (if applicable). These language functions provide descriptors which break down the multiple learning objectives into different competence levels, according to the CEFR.

For example, learning objective 30.1[2] reads: "Can handle complaints and respond to them appropriately." (Bärenfänger et al. 2019: 120) This descriptor shows what is expected at the workplace. It is clear, however, that learners at lower CEFR levels cannot fully achieve this objective. In order to provide guidance as to the degree of communicative competence with regard to this situation at the relevant levels, language functions are defined as follows:

C1: *[…] Can respond to complaints fluently, correctly, and convincingly; if mistakes occur, they are hardly noticeable; the degree of formality is appropriate to the circumstances.*

B2: *[…] Can adequately respond to complaints relatively spontaneously and fluently, practically without giving the impression of being constrained in what he/she wants to say; the degree of formality is appropriate to the circumstances.*

B1: *[…] Can respond to complaints in simple, coherent statements; even if there may be problems with the wording, they can usually be understood without difficulty.*

1. The CEFR itself in its German version (Council of Europe [CoE] 2001) was used together with the Companion Volume (CEFR/CV; CoE 2020) in the English version as the latter had not yet been translated into German. Note that the Companion Volume was not yet available for the 2019 Standard Setting described in this article.
2. This is one of the detailed learning objectives under the broader objective 30: *Respond appropriately to error or fault messages from others and offer assistance* (Bärenfänger et al. 2019: 120).

A2: *[…] Can respond to complaints in short sentences and simple phrases, even if they may falter and have to restart.*

3 Writing tasks and rating criteria

The German Tests for Work have four writing tasks: Reading and Writing, Listening and Writing (both mixed-skill tasks, including mediation), an argumentative Writing task, and a gap-fill task. This range of tasks was developed because it corresponds to the needs analysis. In professional settings, individuals tend to produce written communication that is generally shorter in length when compared to that produced in an educational environment or other contexts. Texts are generally concise, avoiding extensive descriptions or arguments, and often exhibit less variation in linguistic devices. There was much discussion about how this analysis of written language use could be reconciled with the CEFR descriptors which demand a certain complexity, in-depth argumentation and a range of linguistic means and strategies, particularly with regard to the higher competence levels from B2 upwards. This was solved by creating three writing tasks instead of only one. Writing three very different texts gives test takers the opportunity to more fully demonstrate the range of their linguistic resources, while adhering to the writing conventions of many workplaces with respect to conciseness and brevity. The three open-ended tasks are designed to elicit different registers, a variety of work-related conventions and thus a range of linguistic means.

Four rating criteria are used: Task Management, Communicative Design, Accuracy, and Linguistic Range. These rating criteria were developed using CEFR descriptors with some modifications. The original CEFR descriptors were shortened and combined for better handling in the day-to-day rating process. An example of this is the criterion *Linguistic Range* at level B2:

> *Setzt ein hinreichend breites Spektrum sprachlicher Mittel ein, um sich auch zu komplexeren Sachverhalten zu äußern. Variierte Formulierungen; Lücken im Wortschatz können dennoch zu Umschreibungen führen. Verwendet einige komplexe Satzstrukturen.* (telc 2020c: 40-41)
>
> [Has a sufficient range of language to be able to communicate topics of some complexity. Can vary formulation, but lexical gaps can still cause circumlocution. Is using some complex sentence forms.]

This criterion is derived from the following original CEFR descriptors:

> General linguistic range, B2: *Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words/signs, using some complex sentence forms to do so.*
>
> Vocabulary range, B2: *Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.*

When rating the test takers' output, it should be noted that the three texts demonstrate different facets of writing competence. None of the tasks elicits educational language. Consequently, the expectations of some raters, which may have been influenced by their educational background, whether at school, university or in their current teaching context, were not fully met. Rating should reward texts that are succinct, readily comprehensible and written in a way that is appropriate for the specific target audience in the workplace, i.e., colleagues, superiors and clients. The traditional discussion paper, reminiscent of those encountered in academic settings, is ill-suited to this context.

The B2 Reading and Writing task can serve as an example for this discussion point. The task uses two emails as input: the first is a customer complaint, and the second is guidance from your superior on how to respond to the complaint. The test taker must write a short email to the customer apologizing and outlining the next steps. In this situation, neither very original answers nor elaborate ones are expected (indeed, the space on the answer sheet is limited accordingly). This means that parts of the answer may

consist of typical phrases that any customer service staff member would use. While the test taker must respond to the specific situation, generic phrases such as "I sincerely apologize for the inconvenience." and "We're committed to finding a solution." are appropriate and acceptable. The rating focuses on appropriateness in customer relations situations and the mediation required to take the superior's email into account, rather than any personal originality on the part of the employee.

4 Setting standards

During the test development phase, alignment to the CEFR played an integral part. It was explicitly stipulated that the four German Tests for Work would be utilized as end-of-course assessments within a course system that is structured according to CEFR levels. The expectation was that each learning phase would enable learners to progress to the subsequent level. Consequently, alternative approaches, such as defining uneven profiles or varying goals according to different professional aims, were not viable. All in all, the utilization of the CEFR and its four relevant levels (A2 to C1) as a widely understood anchor has facilitated acceptance and rendered the interpretation of results more straightforward than it would have been with a concept requiring more explanation.

The use and usefulness of the CEFR levels for test development was questioned, however, and the CEFR as the default reference framework challenged. The CEFR had to be applied in a way that provides fair learning and assessment opportunities for test takers as well as transparent certification of the competence level reached for all stakeholders. To facilitate the range of perspectives and ensure sufficient scope for discussion, a large number of experts and stakeholders were invited to contribute to the conception of the German Tests for Work, including the definition of learning objectives, via surveys, discussion rounds and interviews. (Plassmann et al. 2021: 13-14). Specifications and exam models were developed and pretested with more than 1,000 learners.

The next step was a standard setting event which took place from 10 to 12 October 2019 with 80 experts. On 10 October, an introduction to the mission, needs analysis, learning objectives and exam formats was given in plenary session. On 11 and 12 October, the experts worked in four groups, each dealing with one skill as the exams provide for a separate partial result for each skill (i.e., Reading, Listening, Speaking and Writing). The event aimed to document and prove that the four exams were aligned to their respective CEFR levels. This involved discussing and judging the provisional cut scores and rating criteria, considering their ability to produce exam results firmly rooted at a specific CEFR level. The general suitability of the tasks was also discussed providing the test development team with extra qualitative feedback. Methods were taken mainly from the *Manual* (CoE 2009) for relating languages exams to the CEFR.

The workshop group focusing on writing skills consisted of 19 independent experts plus two experts who had taken part in the previous phases of the test development project and conducted the workshop. They were mainly professional raters with many years of experience in rating written exam tasks as well as rater and examiner trainers. One expert came from a test provider, one from an academic context, and one represented the Federal Office who had commissioned the tests. It was particularly helpful that the experts offered such different perspectives. The fact that not all of them were very experienced raters led to relevant questions that might otherwise not have been raised.

In the familiarization phase, the group first dealt with the CEFR level system. All participating experts brought in-depth knowledge of the CEFR so that this phase could be kept relatively short at 90 minutes. Sorting exercises and a discussion of salient features of each relevant CEFR level were especially helpful to focus the experts' discussion (CoE 2009: 20-21, 123).

For calibration, writing samples from existing exams were rated. These samples had been rated through intensive calibration work in previous expert workshops and thus provided a benchmark to

which the group could orientate itself.[3] In each case, a global judgement was first given at one of the six CEFR levels. The results were recorded and projected onto the wall as a graph. A strong majority opinion emerged in all cases; deviations were discussed intensively. This was followed by a second global rating in conjunction with an analytical rating process using the relevant CEFR scales for Writing. This second round of rating was discussed in detail as well, which led to further agreement. Finally, 40 writing samples taken from the new exams' pilot version were rated without further discussion.

Before starting the actual benchmarking, familiarization with the learning objectives and writing tasks of the German Tests for Work was necessary. At this point, it was especially important to understand the examinations' concept of offering several relatively short tasks which have to be rated as a whole. Assessing the output to one task only would in many cases—at least for the higher levels B2 and C1—lead to lower marks. A short writing sample provides only a limited opportunity to show the expected range of linguistic means. Also, errors in grammar and spelling tend to gain more weight if there are fewer correct passages to outweigh slips in correctness. It was therefore crucial that raters take *all* the writing tasks into consideration and not judge on first impressions.

In the German Tests for Work the rating process is managed by asking the raters to provide their marks for task management for each written text, but to reserve their judgement on linguistic criteria till the end. During the benchmarking session, however, experts tended to consider each text sample separately. In some cases, rating a rather short text at a time led to lower marks than expected, as short texts might not demonstrate the full range of linguistic abilities of the writer. It was only when all written productions were judged together that the full writing competence emerged. This was discussed in depth and a common understanding about the all-encompassing view necessary for these exams was reached. However, it became clear that this concept would take time to be grasped, as traditional writing tasks tend to consist of one long text, which in turn leads to a slightly different rating process. Some experts were not able to implement this rating across text boundaries fully during the workshop.

Following this discussion, the rating criteria were revised in terms of wording to strengthen the connection to the CEFR and in terms of layout to provide guidance on the rating process. Some of the tasks were modified, as a few were considered too difficult for their respective CEFR level. Extensive training and calibration were planned for the actual implementation of the examinations.

5 Dealing with short texts and mediation

The discussion about short written texts and their rateability can be illustrated by the B2 Reading and Writing task. As described already, a short answer to a customer complaint is required. This is an example[4] discussed at the benchmarking event:

> Sehr geehrter Herr Stemmler,
>
> da Sie schon lange Zeiz unsere Kunde sind, möchte ich uns entschuldigen, dass die Qualität unseres Arbeits nicht mehr gut ist. Im Moment haben wir die Mangel von Mittarbeitern, weil viele krankgeschrieben sind. Ich möchte Ihnen verprechen dass wir schon ab nächstes Mal die Reinigung Ihres Büros ordentlich machen werden.
>
> Mit freundlichen Grüßen, Xxx Xxx

In English translation, the text reads as follows (linguistic errors not included):

3. Form C2 of the *Manual* was used to record these activities (CoE 2009: 182) and the Written Assessment Criteria Grid in Table C4 was used for rating (CoE 2009: 187).
4. Source: internal documentation; the text was not published.

Dear Mr. Stemmler,

As you have been a customer of ours for a long time, I would like to apologize for the fact that the quality of our work is no longer satisfactory. We are currently short-staffed because many of our employees are on sick leave. I would like to promise you that we will clean your office properly from next time onwards.

Kind regards, Xxx Xxx

Discussion of this text revealed that raters found it difficult to weigh up the linguistic strengths and weaknesses in view of the short length of this writing sample, and to deal with the fact that only the third sentence touches on the actual problem at hand, with the other two sentences being generic. The first solution to these rating issues was to consider the test taker's other written work, in order to gain an overall picture of their writing competence. However, the issue of stock phrases and the lack of direct reference to the situation at hand was not fully resolved. Two main arguments were raised in the discussion:

1. Test construct: We want to assess whether the test taker can write at B2 level, i.perhapse. with some confidence, in a clear and detailed way, demonstrating a certain degree of fluency and spontaneity. For the type of text expected in this task, however, it is possible to rely on prepared phrases and text conventions, which some of the experts at the benchmarking did not find sufficient to prove B2 proficiency.
2. Authenticity: Long, complex texts are rare in a work environment. It would also be inappropriate to deviate from the conventional scope of linguistic means in this highly formal occupational situation.

In addition, this is a mediation task. Mediation gained more attention only when the CEFR/CV was published in 2020 and therefore raters were not as confident about this skill as about the more traditional approaches of testing written production and interaction. What can be expected when mediation is asked for? Certainly not just a repetition of everything the input text contains. Some of the raters raised the question of whether ignoring some parts of a superior's email should not lead to lower marks. This rather strict view leads to counting content points rather than evaluating if the text as such fulfils the communicative aim. Actually, some of the best output texts add a personal perspective to the superior's view. This is exactly what mediation as defined by the CEFR/CV should be, and from a very broad perspective it reflects our work environment, where employees are allowed and expected to have their own opinion and perspective.

An intensive discussion about these questions arose during the benchmarking and continued later in regular calibration sessions. The exam provider held the view that deviation from the input email was not to be seen as avoidance of difficult aspects of the task and therefore as an indicator of lower competence, but as an individual approach to task management in line with the learning objectives and the CEFR.

6 Re-rating 2024

The original standard setting was not the end of discussion. Rating criteria were modified and illustrated by examples, their translation into points and cut scores finalized, and a general consensus on the rating process was reached. As rating written (or spoken) output is a highly qualified task, constant training and calibration was provided. So, raters became more confident over time.

Two years after the German Tests for Work were first introduced in 2022, the test provider decided to conduct a small-scale study on rating to check whether the inter-rater reliability observed in calibration sessions meant that the controversial issues from the benchmarking in 2019 had been resolved. To achieve this, the twelve most controversial writing samples from the original benchmarking were chosen and re-rated by ten of the experts who were present in 2019 and were still active raters for the live examinations.

In 2019, a consensus had been reached, but not everybody had been fully convinced. The more skeptical views on the test construct asking for short texts and mediation had a certain influence on the group which may have led to a more severe outcome than intended. Such severe judgments were later overcome through calibration and training. At least, this was the claim of the exam provider, which was to be proved by the re-rating. The study therefore aimed to establish whether the overly severe initial application of the rating criteria, in view of the work-related test construct, had diminished, particularly for B2/C1 texts, which were often compared to tasks in academic language assessment contexts.

The results of the re-rating demonstrate exactly this. As the rating criteria had been modified in their wording in order to provide the best possible CEFR alignment and the underlying points had been altered in order to achieve the best possible balance in the final scoring, the focus of the comparison between 2019 and 2024 is on whether each sample (consisting of three texts for each test taker) was judged as a pass or fail with respect to the intended CEFR level of the task. The following table clearly shows that the pass rate for B2 and C1 writing samples has risen significantly.

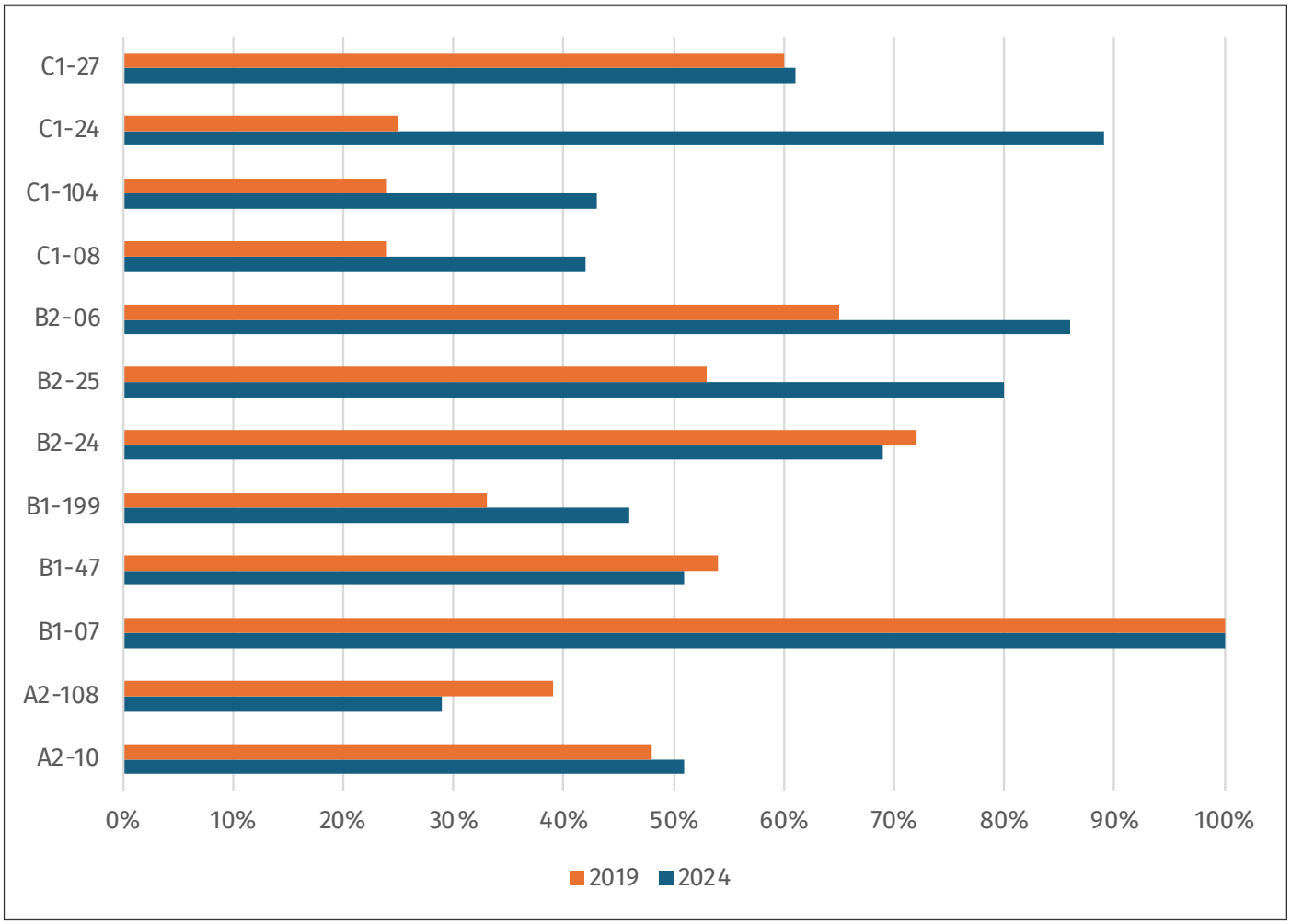

Figure 2. Rating of writing samples in 2019 and 2024 (percentage of raters who placed the text at the respective target level)

The C1 samples 24, 104, and 08 show a far better pass rate in 2024. Sample 24 is regarded as adequate to the target level by almost 90 percent of raters in 2024, compared to only 25 percent in 2019. Samples 104 and 08 are almost 20 percentage points above their initial ratings. B2 samples 06 and 25 as well as B1 sample 199 also show rating outcomes significantly closer to the target level than initially awarded. The other samples show little deviation or in one case a stricter rating. In these cases, other factors than the original discussion about text length and content were dominant in rating.

Thus, questions about the B1/B2/C1 worthiness of these tasks are resolved, as reflected in the higher pass rate, which corresponds to the levels achieved in other skills. Today's rating practice reflects the original intention of the test developers and provides a fair outcome for test takers.

7 Implications for CEFR alignment in vocational language assessment

This case study offers several key insights into linking examinations with the CEFR.

Firstly, it is advisable to establish a strong foundation by clearly defining learning objectives and aligning them with CEFR descriptors. When it comes to work-related exams or teaching and learning materials, using CEFR scales in their original form is not entirely appropriate. Raters, teachers and other stakeholders must apply CEFR descriptors from the personal, public or educational domain to the occupational context, as there are not enough work-related illustrative scales. This demands a high degree of abstraction and is therefore less intuitive than desired.

Secondly, a single standardization workshop is not enough. Defined standards must be validated in practice and reinforced continuously, even after the initial development phase, when tasks, rating criteria and scoring are well defined. The need for continuous refinement and the consistent application of established standards is particularly pertinent in the assessment of free text production, where human judgement is required. Ongoing calibration, discussion and shared understanding are indispensable in this context.

Thirdly, the project demonstrates that task formats which deviate from traditional expectations, such as short or highly formalized texts, can be successfully implemented. Although developing such tasks and ensuring valid assessment requires additional effort, the increase in the authenticity of language use is significant.

8 References

Bärenfänger, Olaf, Nadja Nitsche & Sibylle Plassmann. 2019. *Lernziele. Spezialberufssprachkurse A2 und B1. Basisberufssprachkurse B2 und C1*. Frankfurt am Main: telc GmbH. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/lernzielkatalog-spezial-und-basisberufssprachkurse.pdf?__blob=publicationFile&v=7 (accessed 6 Oct 2025).

BAMF. 2021a. *Konzept für einen Spezialkurs A2 im Rahmen der bundesweiten berufsbezogenen Deutschsprachförderung nach § 45a Aufenthaltsgesetz*. Nürnberg: BAMF. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/spezialmodul-a2.pdf?__blob=publicationFile&v=8 (accessed 6 Oct 2025).

BAMF. 2021b. *Konzept für einen Spezialkurs B1 im Rahmen der bundesweiten berufsbezogenen Deutschsprachförderung nach § 45a Aufenthaltsgesetz*. Nürnberg: BAMF. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/spezialmodul-b1.pdf?__blob=publicationFile&v=7 (accessed 6 Oct 2025).

BAMF. 2021c. *Konzept für einen Basiskurs B2 im Rahmen der bundesweiten berufsbezogenen Deutschsprachförderung nach § 45a Aufenthaltsgesetz*. Nürnberg: BAMF. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/kurskonzept-b2.pdf?__blob=publicationFile&v=11 (accessed 6 Oct 2025).

BAMF. 2021d. *Konzept für einen Basiskurs C1 im Rahmen der bundesweiten berufsbezogenen Deutschsprachförderung nach § 45a Aufenthaltsgesetz*. Nürnberg: BAMF. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/kurskonzept-c1.pdf?__blob=publicationFile&v=14 (accessed 6 Oct 2025).

Council of Europe. 2001. *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Klett.

Council of Europe. 2009. *Relating language examinations to the 'Common European Framework of Reference for Languages: Learning, teaching, assessment' (CEFR). A Manual*. Strasbourg: Council of Europe. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d (accessed 6 Oct 2025).

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

Plassmann, Sibylle, Hannah Blumöhr-Giuri, Mustafa Cikar & Magdalena Igiel. 2021. *Prüfungshandbuch. Deutsch-Tests für den Beruf A2, B1, B2 und C1*. Frankfurt am Main: telc GmbH. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Integrationskurse/Lehrkraefte/pruefungshandbuch-deutsch-tests-beruf.pdf?__blob=publicationFile&v=6 (accessed 6 Oct 2025).

telc. 2020a. *Übungstest 1. Deutsch-Test für den Beruf A2*. Prüfungsvorbereitung. Frankfurt am Main: telc GmbH. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/a2-modelltest-bsk.pdf?__blob=publicationFile&v=14 (accessed 6 Oct 2025).

telc. 2020b. *Übungstest 1. Deutsch-Test für den Beruf B1*. Prüfungsvorbereitung. Frankfurt am Main: telc GmbH. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/b1-modelltest-bsk.pdf?__blob=publicationFile&v=14 (accessed 6 Oct 2025).

telc. 2020c. *Übungstest 1. Deutsch-Test für den Beruf B2*. Prüfungsvorbereitung. Frankfurt am Main: telc GmbH. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/b2-modelltest-bsk.pdf?__blob=publicationFile&v=14 (accessed 6 Oct 2025).

telc. 2020d. *Übungstest 1. Deutsch-Test für den Beruf C1*. Prüfungsvorbereitung. Frankfurt am Main: telc GmbH. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Berufsbezsprachf-ESF-BAMF/BSK-Konzepte/c1-modelltest-bsk.pdf?__blob=publicationFile&v=14 (accessed 6 Oct 2025).

telc. 2020e. *Additive Zusatzqualifizierung für Lehrkräfte in Berufssprachkursen. Konzeption mit einem Kompetenz- und Anforderungsprofil für Lehrkräfte*. Frankfurt am Main: telc GmbH. https://www.bamf.de/SharedDocs/Anlagen/DE/Integration/Integrationskurse/Lehrkraefte/konzeption-fuer-die-zusatzqualifikation-von-lehrkraeften-bsk-pdf.pdf?__blob=publicationFile&v=7 (accessed 6 Oct 2025).

9 Biography

Dr. Sibylle Plassmann is Head of Tests and Training at telc GmbH, specializing in language testing for academic purposes, integration and in vocational contexts. With extensive experience in language teaching, teacher training and assessment, she has contributed to national and international projects on language proficiency and integration. She also works on advancing quality standards in language testing, ensuring validity of tests as well as robust procedures in test administration.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

"Every teacher was an island": Teacher perceptions of a CEFR alignment project to implement a standardized approach to assessment

Carolyn Westbrook, British Council, Great Britain
Aidan Holland, British Council, Great Britain

https://doi.org/10.37546/JALTSIG.CEFR8-7

Implementing fair and valid assessment is a fundamental part of a teacher's job as assessments enable learners to demonstrate progress as well as highlighting strengths and weaknesses (Rahman 2018). Results should be reported in an accessible way so learners and other stakeholders understand the outcomes and learners' progress. This article reports on a collaboration between assessment researchers and teachers to standardize the approach to assessment in a global secondary language programme. Teaching materials were mapped to the CEFR, standardized set-up notes were created for assessment tasks, assessment tools and training were developed, and teachers implemented the approach in a practical way. Feedback showed the system improved objectivity and clarity in assessment, though challenges around feasibility and alignment with the CEFR Companion Volume (CEFR/CV) (Council of Europe [CoE] 2020) remained.

Keywords: CEFR alignment, classroom-based language assessment, teacher assessment, teacher-researcher collaboration, standardized approach to assessment

Acknowledgements

The authors would like to thank Harpreet Kaur, Samantha Lewis, Johnathan Cruise, Cristina Barry and Howard Cheung for their invaluable contributions to the project, Richard Spiby and the late Dr Jamie Dunlea for their input in the early stages of the project, and the British Council teachers involved in the project for their time and feedback.

1 Introduction

Assessment is a key part of teaching and learning. As Race et al. (2005: xi) note, "[n]othing we do to, or for our students is more important than our assessment of their work and the feedback we give them on it". Effective assessment highlights learners' strengths and weaknesses, enables them to demonstrate progress (Rahman 2018), and shapes classroom practice (Bachman and Palmer 2010). To be meaningful, assessments must be valid, reliable, and practical for teachers to use, while results should be reported in accessible ways that promote transparency and support learner progress.

In 2016, the British Council began standardising its secondary teaching materials across 50 countries. Courses were based around CEFR levels, yet the teaching materials were not formally aligned to the

CEFR. The teaching materials are based around topic modules, each of which is presented in the form of a magazine, culminating in a task-based project. Each CEFR (sub-)level comprises 10 magazines. However, assessment remained inconsistent: there were no shared criteria, limited guidance on evaluating projects, and no explicit alignment of tasks to the CEFR, leaving teachers uncertain about standards, and stakeholders unclear about learner progress.

This article reports on a collaborative project between researchers and teachers to create a CEFR-aligned assessment system for the British Council's *Secondary Plus* courses, delivered to learners aged 11-17 worldwide. The project involved mapping course materials to the CEFR, developing a standardized approach to assessment, developing assessment and reporting tools, training teachers, and collecting feedback on implementation. The article presents teachers' perceptions of the new system and the challenges encountered, particularly regarding use of descriptors from the CEFR/CV (CoE 2020).

2 Literature review

Traditionally, curriculum, delivery (i.e., the operationalization of the curriculum in a specific context) and assessment have been seen as separate entities: teachers focus on teaching and learning but curricula are designed by publishers or education boards while assessment experts design assessments (O'Sullivan 2021). However, Bunch (2012: 1) argues that "[a] key component of educational achievement test validation is alignment of the test to both curriculum and instruction". Similarly, O'Sullivan (2021) posits that curriculum, delivery and assessment must be inextricably linked for a learning system to function:

> Within the system, the three core elements (curriculum, delivery, assessment) must be based on a single philosophy of learning, supported by clearly defined models of language ability and progression, and underpinned by a measurement model. Failure to ensure that all three are fully in harmony is likely to lead to the failure of the system. (O'Sullivan 2021: 2)

The philosophy of learning in the CEFR is the action-oriented approach. Learners are seen as 'social agents' who have tasks to accomplish in a particular context (CoE 2001: 9). The CEFR provides "a common basis for the explicit description of objectives, content and methods" (CoE 2001: 1), thereby enabling curricula, teaching materials and assessments to be aligned to external standards such as the CEFR.

The document *Aligning Language Education with the CEFR: a handbook* (British Council et al. 2022) outlines five stages for aligning materials to the CEFR:

1. Familiarization—participants are familiarized with the CEFR descriptors so they can apply them accurately during specification.
2. Specification—descriptors that match the skills and competences targeted by materials are identified.
3. Standardization—experts analyse materials and establish benchmark samples that exemplify performance at each CEFR level.
4. Standard setting—minimum standards for each level are determined.
5. Validation—gathers evidence to support the alignment claims.

Although traditional assessment has focussed on assessing the four skills plus grammar and vocabulary, the (2001) CEFR and its *Companion Volume* highlight four modes of communication: reception, production, interaction and mediation. These reflect authentic language use (Plakans 2020). Technological advances have also enabled multimodal assessments, which enhance motivation, promote critical thinking (Varaporn and Sitthitikul 2019) and support inclusivity for learners with Special Educational Needs and Disabilities (SEND) (Ellis 2024).

3 Project phases

The project progressed through four main phases. These are summarized in Figure 1.

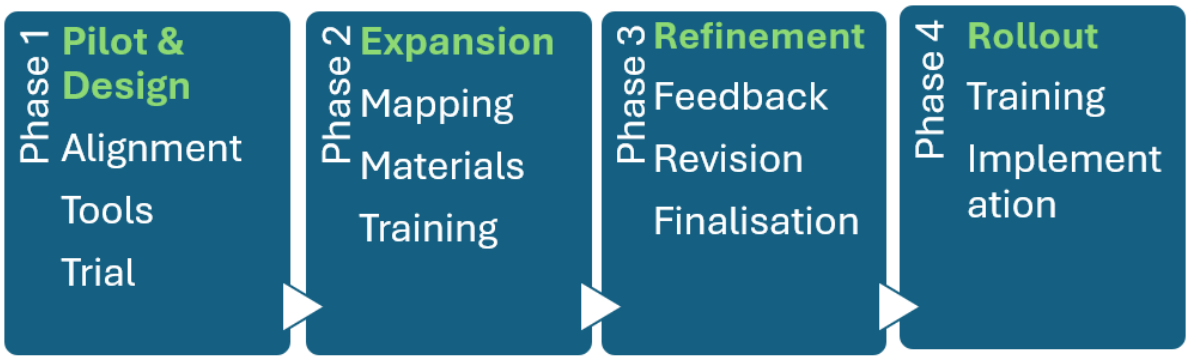

Figure 1. The four project phases

3.1 Phase 1

Phase 1 was a pilot phase which involved mapping one set of ten magazines at the B1.1 level to the CEFR using the British Council's internal CEFR database, the British Council Curriculum Framework (BCCF), and the CEFR/CVfollowing the steps outlined in the Handbook (British Council et al. 2022). The BCCF was created before the publication of the CEFR/CV so descriptors for some scales were missing, hence the need to use the CEFR/CV too. Since each magazine culminated in a multimodal project, these projects were used as the basis for the assessment and amended as necessary to reflect the content of the magazine, the specific learning outcome and the level of the magazine. Phase 1 also involved creating standardized project set-up notes for teachers, assessment and reporting tools using CEFR Can Do descriptors and specially developed Performance Indicators (PIs) which needed to be feasible for teachers with limited time for marking. The Performance Indicators are included in the BCCF and provide additional information about performance at the different CEFR levels, broken down into six areas: focus, fluency, range, accuracy, discourse, and appropriateness. This phase culminated in a small-scale trial using the projects and the assessment and reporting tools produced.

3.2 Phase 2

In the second phase, the CEFR mapping was extended to all Secondary Plus levels from A1 to C1. In total, 12 levels, each with 10 magazines per level, were mapped, resulting in 120 magazines being mapped to the CEFR. Standardized set-up notes for teachers, assessment and reporting tools for all 120 projects were created as well as a training pack for teachers including a handbook and a training presentation. These materials were then trialled, and samples were collected for standardization and benchmarking. An expert panel of six teachers and researchers led the standardization and standard setting phases, producing benchmark samples to support future training and ensure consistency across levels.

3.3 Phase 3

The third phase involved finalising the assessment sheets and teacher training materials based on the feedback received.

3.4 Phase 4

In this final phase, rollout commenced. This involved training the teachers to implement the new assessment approach using the teacher training materials developed previously. Training was delivered

to 600 teachers and academic leads, and the approach has been implemented in 25 countries around the world. The rollout is currently on hold due to ongoing transformation within the organization.

4 Results

The teachers who were trained as part of the initial rollout were asked to complete a brief pre-training questionnaire to gauge their level of teaching experience and the amount of prior assessment training. After undertaking the training, they were asked to complete a post-training questionnaire. A selection of the results from both questionnaires is reported below.

4.1 Pre-training questionnaires

The pre-training questionnaire was completed by 195 teachers. Questions 1-5 and 16-17 collected consent and background data from the respondents. Due to space limitations, only a selection of key results is presented here.

Question 6 asked about teaching experience. Sixty-six percent of respondents had between 11 and 30 years' experience teaching, 8% had over 31 years' experience, and the remainder (26%) had been teaching for 10 years or less.

Despite all respondents confirming that they assessed their learners regularly, 23% had not had any training in assessment (Q.7). Of the remaining 77% who had had some training in assessment, this was often only short sessions of up to two days (Figure 2).

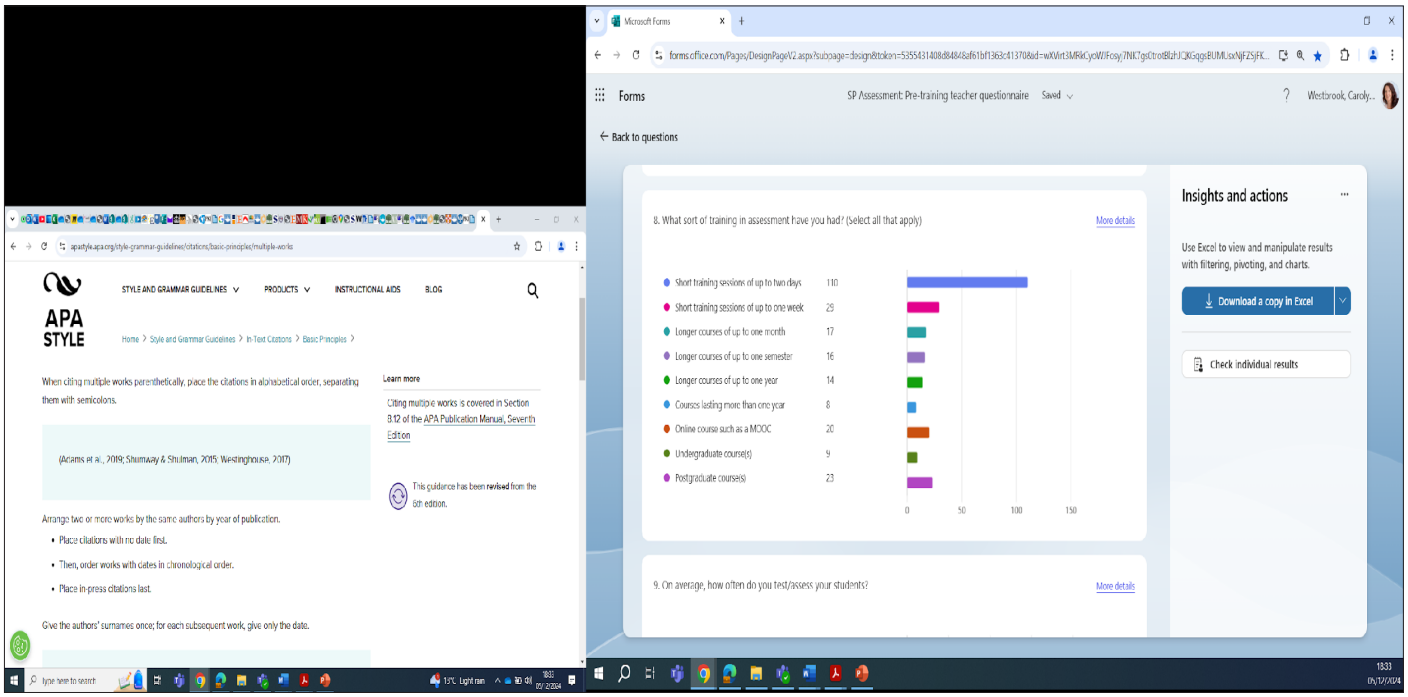

Figure 2. Responses to "What sort of training in assessment have you had?"

Question 13 asked if the assessments helped teachers teach; 79% responded yes. Only 5% of participants replied that this was not the case, while 16% were unsure. Question 14 expanded on the previous question by asking how the assessments helped teachers. Many respondents stated that the assessments provided information about learners' weaknesses, which helped teachers to plan for future lessons and to see progress:

I can see how well the ss [sic] have understood the material covered and what they need for the future. So it helps me mold the course to the needs of my ss [sic]. (T18)

[Assessments] help to see if my learners are making progress … (T119)

Finally, question 15 looked at teachers' satisfaction with assessment up to that point (Figure 3). 50% were either very satisfied or satisfied, while 41% were neutral. The remaining 9% were either dissatisfied or very dissatisfied.

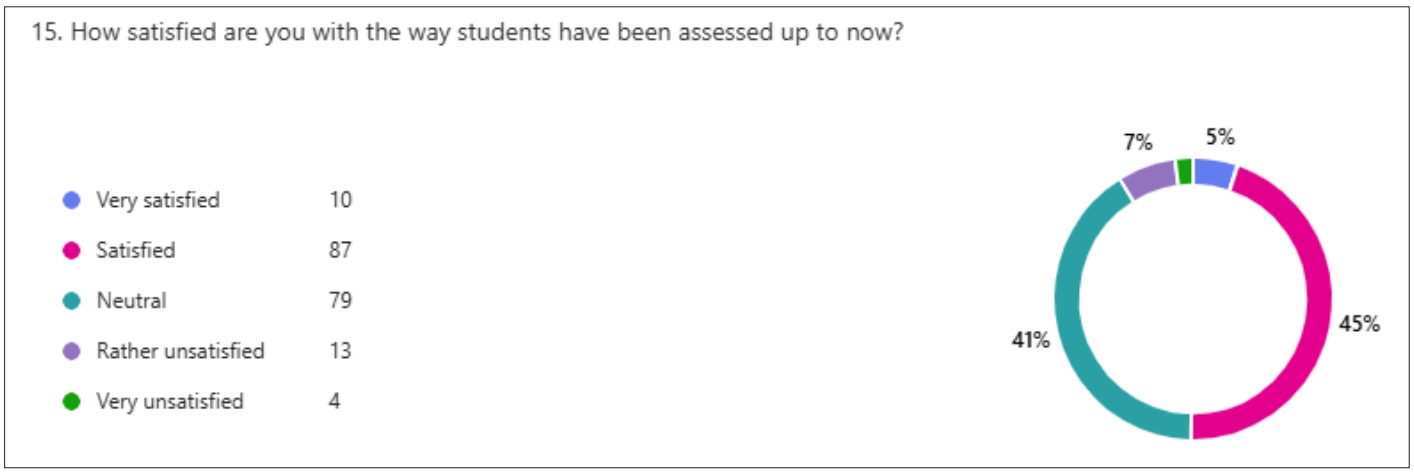

Figure 3. Responses to "How satisfied are you with the way students have been assessed up to now?"

4.2 Post-training questionnaires

Ninety-four teachers completed the post-training questionnaire (59 of them had completed the pre-training questionnaire), providing feedback on their experiences with the new assessment approach.

Questions 1-3 and 9-10 collected consent and background information. Questions 4-8 investigated teachers' perceptions of the new approach. Figure 4 shows that 72% of respondents agreed or strongly agreed that the new assessment system would enable them to assess more objectively (question 4).

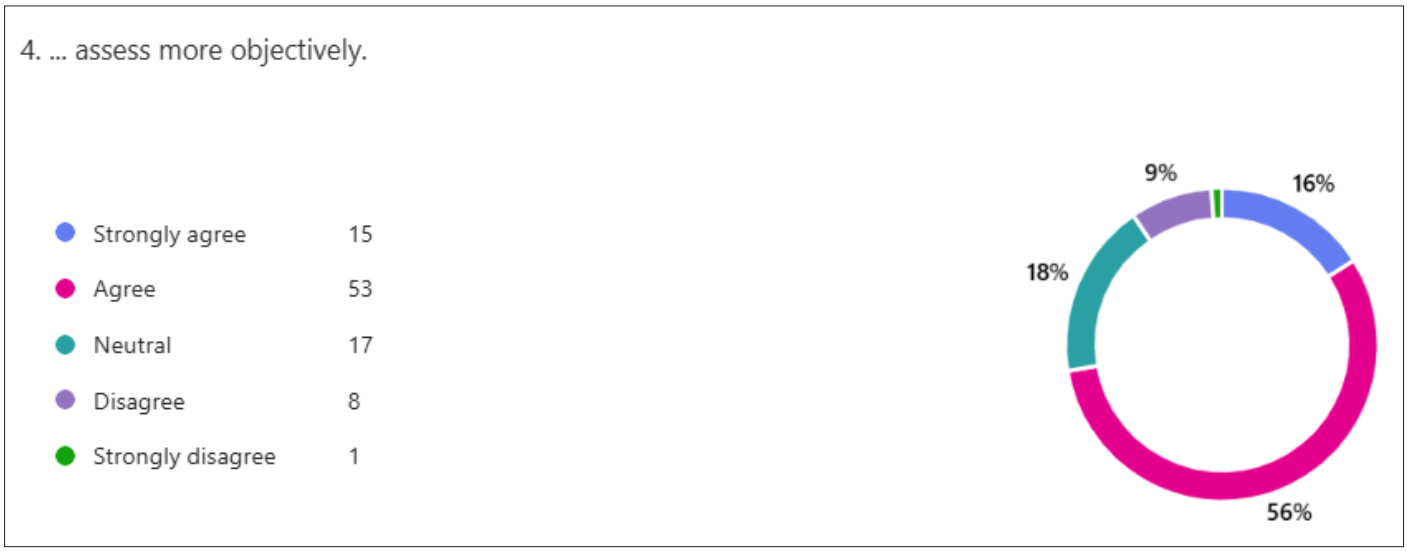

Figure 4. Responses to "The new assessment system will enable me to assess more objectively"

For question 5, 62% either agreed or strongly agreed that the new approach would improve student outcomes/results while 30% were neutral (see Figure 5).

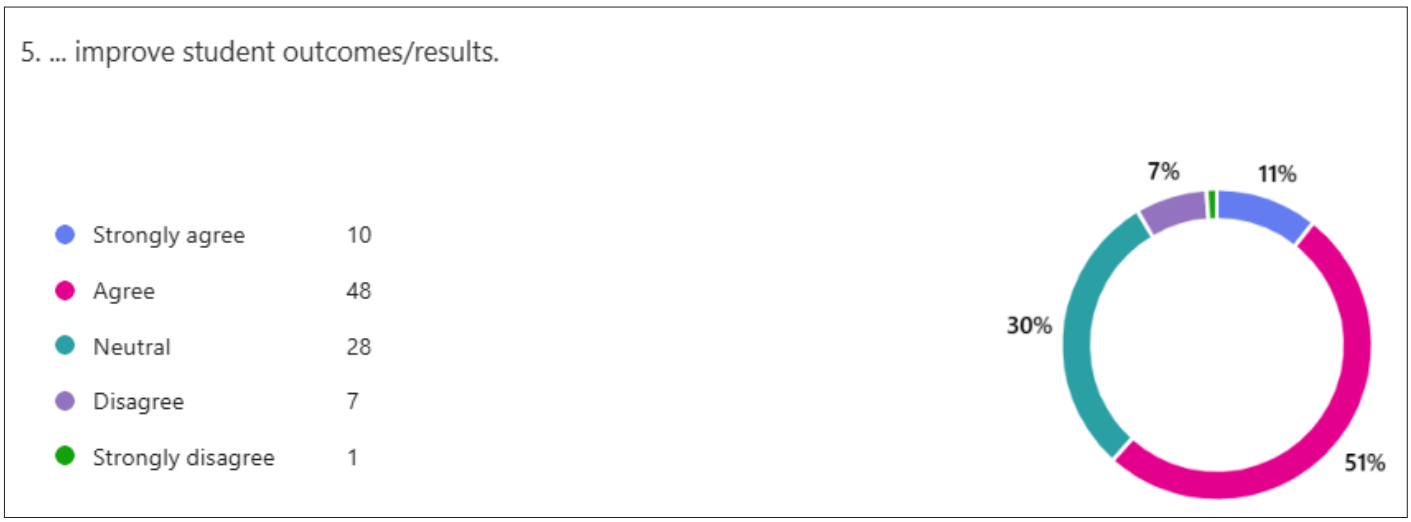

Figure 5. Responses to "The new assessment system will enable me to improve student outcomes / results"

Question 6 (Figure 6) asked whether the new approach would enable teachers to provide better feedback to students and parents. 78% either agreed or strongly agreed that this was the case.

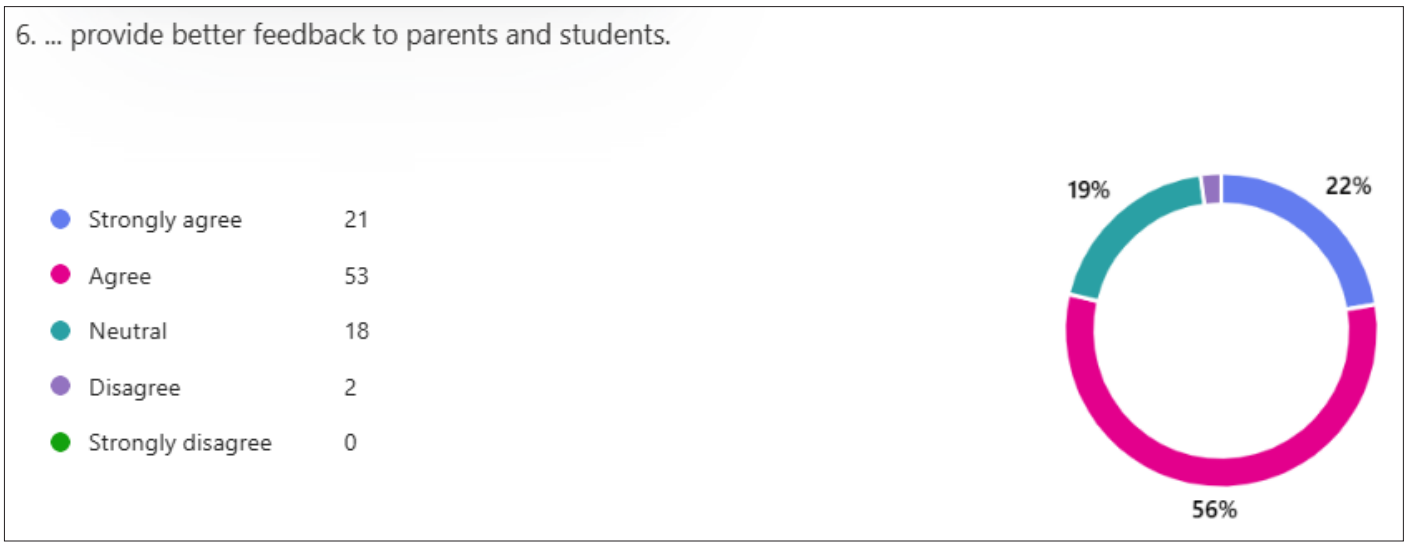

Figure 6. Responses to "The new assessment system will enable me to provide better feedback to parents and students"

Similarly, 78% either agreed or strongly agreed that the new approach would enable them to better prepare their students for the project/task (question 7, Figure 7).

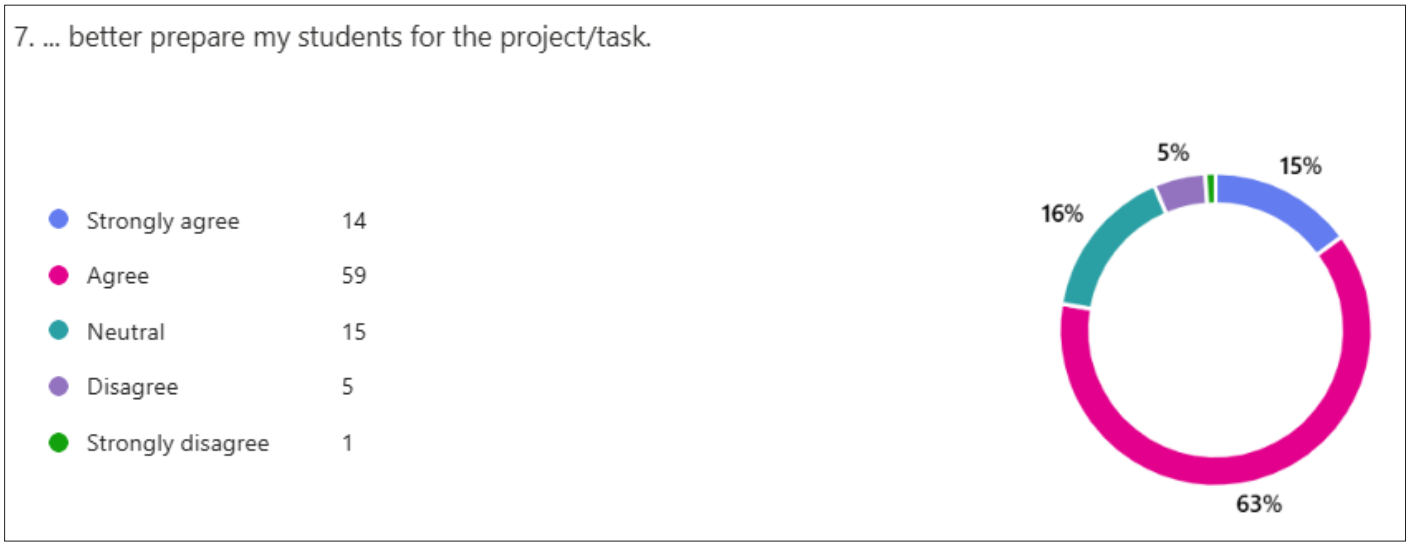

Figure 7. Responses to "The new assessment system will enable me to better prepare my students for the project/task"

For question 8, 54% either agreed or strongly agreed that the new approach would enhance how they teach; 32% were neutral (Figure 8).

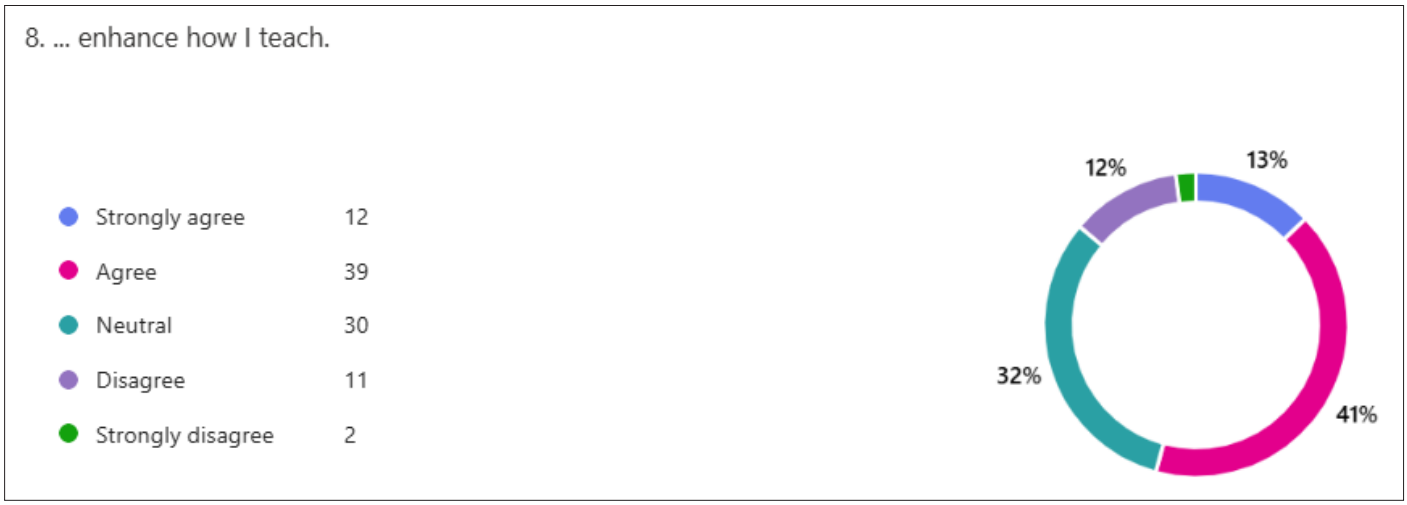

Figure 8. Responses to "The new assessment system will enable me to enhance how I teach"

For questions 5 and 8, 30% and 32% of respondents respectively were neutral. This may have been because they had not yet implemented the new approach so they were unsure of its impact.

4.3 Qualitative analysis of feedback from focus groups

Two focus groups and two interviews were carried out with teachers who had implemented the approach with two or more magazines to collect their feedback about standardization in the new approach, the number of assessments, and the ease of use of the Can Dos and the PIs.

4.3.1 Standardization in the new assessment approach

Some participants reported issues with the previous approach which led to bias and inconsistency, and welcomed the standardization in the new approach:

> *[The previous assessment approach] had its pluses and minuses. Obviously, standardization was the biggest minus because we all had our different ideas and … we know our students well. So … we already have an idea of what grade to give our students. I think there's a bit of bias based on their past performance … (I1M)*

> *… [previously] everybody was their own island … that is the beauty of this new assessment system, because with my class, it was just me and what I saw fit for my students, … and there was inconsistency, right, between … me and other teachers*. (I2G)

> *… the parents will want it to be standardized because [that's] what they're used to [from school] …* (FG3H)

Others were less happy about the uniform approach, feeling like they were expected to "become robots" (P1-1) or that it took away some of the fun of teaching the Secondary Plus courses:

> *… when students come to do the First Certificate exam, … they might not enjoy it, but they know they're preparing [sic] the specific certificate exam [whereas] often with the [Secondary Plus] magazines, it's [sic] viewed … by teachers and also maybe by students that [they] should be … enjoyable …* (FG4J)

4.3.2 Number of assessments

The number of assessments that would be feasible in the new approach was discussed repeatedly during the project. There were differing opinions within the project team and among the teachers in the pilot phases. In the end, the decision was made to have only one standardized assessment per module. However, opinions among focus group and interview participants would suggest that more than one would be appropriate:

> *… I would say … test them on everything …* (FG3H)

> *I really don't think that parents would just wear one assessment per module*. (FG3K)

> *I don't think it's too tough to do two assessments*. (I1M)

4.3.3 Ease of use of Can Dos and PIs

For a standardized assessment approach to function, teachers need to feel confident that they understand and can apply the Can Dos and PIs when marking learners' work. Participants were generally positive about the grade bands, the Can Dos and PIs. They felt they were useful and easy to apply:

> *Just three categories makes it really, really simple. And I think it's quite clear as well*. (I1M)

> *I think they're fair … they're clear … they're achievable*. (I2G)

But many participants agreed that hitting all six PIs was challenging for learners:

> *… hitting all of them is really hard*. (I1M)

5 Discussion

The pre-training questionnaires demonstrated that most teachers had had some assessment literacy training before doing the training for the new assessment approach. Teachers assess their learners regularly and they see the value of assessments in identifying strengths and weaknesses (Rahman 2018) and helping to shape classroom practice (Bachman and Palmer 2010) by providing information to teachers regarding areas that learners struggled with so teachers can go over these again.

The post-training questionnaire responses and the comments from focus group participants suggest that teachers recognized the strengths of the new approach while at the same time noting challenges with the previous approach. One of the weaknesses of the latter lay in the lack of standardization and reliance on subjective interpretations whereas the new approach allows teachers to evaluate more objectively and provide clearer feedback to learners and carers.

Despite the perceived benefits of the new approach, some teachers felt constrained by the new system. The data suggests that this perception might have several causes. Firstly, the new assessment system was seen as potentially stunting opportunities for teacher creativity, akin to the effect which preparing for large-scale exams might have. Furthermore, the limited assessment literacy (Stiggins 2014) among teachers may have been another obstacle to accepting the new approach. Consequently, any resistance could be due to a lack of understanding of the importance of standardized, reliable marking. Overall, though, the results from the pre- and post-training questionnaires and the focus groups suggest that the system contributes to greater clarity and purpose in teachers' assessment processes and demonstrates the system's potential to positively impact teaching while underscoring the need for continued support and training.

The feedback also highlights the tension between achieving construct validity and practicality. While the assessment researchers emphasized the importance of multiple assessments to mitigate the risk of construct under-representation (Messick 1989), practical constraints led to a compromise of one standardized project assessment per module, carried out by teachers, with the option of an additional assessment for each module based on context. This underscores the challenge of balancing best practices in assessment with the realities of teaching contexts, particularly where workloads and contractual limitations are significant factors.

Teachers reported feeling comfortable and confident in using and applying the Can Dos and PIs. However, the general consensus that hitting all six would be a challenge for learners is something that could be reviewed for future academic cycles. Alternatively, additional information could be provided to better prepare learners for the assessments, which may occur organically as teachers become more accustomed to the new approach.

6 Challenges

As with any project, there were challenges along the way. The challenges were two-fold, relating to internal issues on the one hand and the use of the BCCF and CEFR/CV on the other.

6.1 Internal challenges

One challenge was that teachers frequently adapt Secondary Plus materials to suit their classes and their contexts due to cultural differences and interests. However, if the materials are amended too much, the learners may not have covered everything necessary to successfully complete the project, thereby leading to a mismatch between delivery and assessment (O'Sullivan 2021). To avoid this mismatch, the project team provided detailed notes so teachers would know what aspects to focus on during the magazine even if input was amended.

The next challenge was the debate around how many assessments would be feasible for teachers. Due to contractual constraints, many teachers are unable to spend much time marking outside of class,

so a key focus of the approach had to be on practicality. Further to many discussions within the project team, the decision was made to have only one standardized assessment per module. Some teachers considered this sufficient while others wanted more than one assessment, often because that is what parents in those contexts expect. As a result, a second assessment task was created for any centres requiring an additional assessment.

Finally, many of the projects originally culminated in presentations, so some projects were amended to have a different output. This may have been because the materials were not empirically aligned to the CEFR when they were designed, thus resulting in a lack of construct coverage across the possible CEFR scales. However, since the alignment exercise highlighted this construct under-representation, this can be addressed when revising and updating the materials.

6.2 BCCF & CEFR/CV challenges

When aligning the materials to the CEFR using the BCCF, there were issues with missing descriptors since the BCCF had been created before the CEFR/CV was published. As a result, the new scales and descriptors, most notably the mediation scales, were missing so we had to refer to the CEFR/CV directly. However, the CEFR/CV was not designed with young learners in mind so, in line with the recommendation in the CEFR/CV to "select … [and] adapt the formulation of [descriptors] … to better suit the specific context" (CoE 2020: 42), some descriptors were amended slightly.

Another challenge was that the CEFR/CV was also missing descriptors in some scales. For example, the *Explaining data* scale includes descriptors for the plus levels at A2 and B1 but not at B2. Therefore, differentiating between a B2.1 performance and one at B2.2 is challenging. Given that the Secondary Plus courses are broken down into B2.1 and B2.2, we adopted the B2 descriptors and tried to differentiate between levels using other descriptors which were also appropriate for the particular task.

Similarly, some scales had descriptors for one side of an interaction but not the other. For example, in the *Interviewing and being interviewed* scale, there are no descriptors for the interviewer role below B1.1 nor at B2.1. Conversely, there is no interviewee role at B2.2. Where the descriptors for certain roles were missing, we amended the assessment task so that the assessed role was the one for which there was a corresponding descriptor.

Finally, in both the BCCF and the CEFR/CV, there are often several descriptors for one CEFR level or descriptors with multiple examples, which are not all relevant. Again, the project team had to decide which descriptors to adopt and which to adapt.

7 Conclusion

This project contributed to fairer and more consistent assessment by mapping teaching materials to the CEFR to ensure the materials targeted the correct CEFR level, and developing a standardized approach to assessment in which teachers assess learners using the project at the end of each magazine, using standardized set-up notes, criteria, and reporting tools. Teachers were trained in the assessment approach and standardized using CEFR-aligned benchmarks. The project created stronger collaboration between researchers and teachers, integrating classroom insights into the design process. Despite challenges such as balancing workload with assessment demands and addressing gaps in the BCCF and CEFR CV, the results of the mapping exercise and the alignment of the assessments to the CEFR confirmed that the Secondary Plus materials broadly target the intended CEFR levels. Furthermore, the new tools should make learning more tangible for learners and promote objectivity in teachers' assessments.

8 Applications

Beyond the immediate context, the project illustrates how collaborative alignment initiatives can strengthen assessment literacy, improve transparency for stakeholders and build consistency in applying marking standards. It also shows that engaging teachers throughout the project increases the feasibility and acceptance of new approaches.

9 Future directions

Moving forward, the project highlights the need to update and extend frameworks such as the BCCF and CEFR/CV, including descriptors tailored for young learners. It also points to the value of continued teacher training and support, both to address perceptions of constraint and to sustain consistent application across contexts. More broadly, similar alignment projects could inform assessment practices in other large educational programmes, supporting fairer, more transparent learning systems internationally.

10 References

Bachman, Lyle F. & Adrian S. Palmer 2010. *Language assessment in practice.* Oxford: Oxford University Press.

British Council, UKALTA, EALTA & ALTE. 2022. *Aligning language education with the CEFR: A handbook.* https://www.britishcouncil.org/sites/default/files/cefr_alignment_handbook_layout.pdf (accessed 28 November 2024).

Bunch, Michael B. 2012. Aligning curriculum, instruction, and assessment. https://www.measurementinc.com/sites/default/files/AligningCurriculumAssessmentandInstruction.pdf (accessed 10 Sept 2025).

Council of Europe. 2020. *The Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion Volume.* https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4 (accessed 29 Nov 2024).

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press. https://rm.coe.int/1680459f97 (accessed 31 Oct 2025).

Ellis, Paul. 2024. Cultivating inclusion: Strategies for embracing diverse learners. Cambridge University Press and Assessment, Partnership for Education. https://www.cambridge.org/partnership/cultivating-inclusion-strategies-for-embracing-diverse-learners (accessed 25 Nov 2024).

Messick, Samuel. 1989. Validity. In Robert L. Linn (ed.), *Educational measurement*, 3rd edn, 13103. American Council on Education; Macmillan.

O'Sullivan, Barry. 2021. *The Comprehensive Learning System.* London: British Council. https://www.britishcouncil.org/sites/default/files/the_comprehensive_learning_system_new_layout.pdf (accessed 8 Sept 2025).

Plakans, Lia. 2020. Assessment of integrated skills. In Carol A. Chapelle (ed.), *The concise encyclopedia of applied linguistics*. Chichester: Wiley Blackwell. DOI: 10.1002/9781405198431.wbeal0046.pub2.

Race, Phil, Sally Brown & Brenda Smith. 2005. *500 Tips on assessment*, 2nd edn. London: Routledge.

Rahman, Md. Mehadi. 2018. Exploring science teachers' perception of classroom assessment in secondary schools of Bangladesh. *European Journal of Education Studies* 4(9). 139-160. DOI: 10.5281/zenodo.1296835.

Stiggins, Rick J. 2014. Improve assessment literacy outside of schools, too. *Phi Delta Kappan* 96(2). 67-72. DOI: 10.1177/0031721714553413.

Varaporn, Savika & Pragasit Sitthitikul. 2019. Effects of multimodal tasks on students' critical reading ability and perceptions. *Reading in a Foreign Language* 31(1). 81-108. https://files.eric.ed.gov/fulltext/EJ1212804.pdf (accessed 25 Nov 2024).

11 Biographies

Carolyn Westbrook is a test development researcher at the British Council. Formerly an associate professor in EFL, she is a senior fellow of the Higher Education Academy (now Advance HE) in the UK. She delivers assessment literacy training for teachers and lecturers around the world and has worked on a number of assessment development projects. Her research interests are the assessment of integrated skills, EAP, ESP and EMI. She has worked as a teacher and teacher trainer for over 30 years.

Aidan Holland works as Global Assessment Solutions Manager, Wider Europe. He has been working for the British Council in a variety of roles since 2009. He has over 20 years' international experience in teaching and teacher training and almost a decade of experience in academic management positions. He holds an MA in TEFL/SL from the University of Birmingham and an MA in Language Testing from the University of Lancaster.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

The alignment process as good practice in Italy for linking learning and assessment: A case study

Sabrina Machetti, University for Foreigners of Siena, Italy
Giulia Peri, University for Foreigners of Siena, Italy

https://doi.org/10.37546/JALTSIG.CEFR8-8

Since the publication of the CEFR Companion Volume (CEFR/CV) (Council of Europe [CoE] 2020), research on aligning proficiency tests with new mediation descriptors remains scarce in Italy. However, the four Italian language certifying bodies (CLIQ Consortium), are developing projects that emphasize mediation in L2 Italian teaching and assessment. One such initiative, "Je le sais faire en Italien", is currently being piloted in high schools in Marseille and Aix-en-Provence in collaboration with the French Ministry of Education. Building on this project, this study presents the initial findings on aligning mediation descriptors with CILS exams (Certification of Italian as a Foreign Language) at A2 and B1 levels. These exams are aimed at foreign students studying Italian in school contexts, in Italy and abroad. This study analyses data collected to gather evidence on assessment impact (Saville and Khalifa 2017). Notably in some Italian high schools abroad, this approach has positively influenced syllabi, aligning with the principles of learning-oriented assessment (Purpura and Turner 2018) principles.

Keywords: mediation, L2 Italian, proficiency test, learning-oriented assessment, CILS exams

1 Introduction

This article presents the initial results of a study aligning CILS exams (Certification of Italian as a Foreign Language) with the CEFR descriptors dedicated to mediation (Council of Europe 2020) at levels A2 and B1. These exams are intended for candidates of foreign origin who use the Italian language for different purposes and in various contexts, including educational settings, both in Italy and abroad. Given this characteristic of the CILS exams and also considering the complex nature of communication and therefore users'/learners' needs, mediation appears to be pivotal for the various communicative dynamics in which users/learners are required to use their mediation skills in order to perform their daily tasks and achieve communicative success. The article discusses the data collected during the first steps of the alignment process, with particular attention to the familiarization and specification phases. It supports and contributes to the process of aligning the CILS exams with the CEFR/CV. This process follows the one initiated through the use of the *Manual for relating language examinations to the CEFR* (CoE 2009), documented and disseminated both through internal reports and publications (Barni et al. 2010; Bagna et al. 2012). The alignment process includes among its various advantages achieving systemic coherence and transparency, establishing a basis for principled comparison, and monitoring for purposes of quality assurance (British Council et al. 2022: 11). In addition, for the CILS exams it also constitutes the starting point for the revision and updating of their construct.

2 Research background

2.1 Mediation in the CEFR (2001) and CEFR/CV (2020)

In the field of linguistic sciences, the conceptualization of mediation as a semiotic activity—as a process of meaning-making through linguistic and other symbolic resources—has only emerged relatively recently (North and Piccardo 2016; Machetti and Siebetcheu 2017). This focus is central to the CEFR (CoE 2001) and to the more recent CEFR *Companion Volume* (CoE 2020). Both the CEFR and the CEFR/CV consider mediation as an essential and irreplaceable component in the processes of constructing and negotiating meaning and sense, and they discuss its characteristics, which go beyond a merely instrumental function. This entails that the use and learning of a language—through continuous interaction with other systems of signs and in situations of contact with other languages—requires each user/learner to engage in the mediation of texts, a process that "involves passing on to another person the content of a text to which they do not have access, often because of linguistic, cultural, semantic or technical barriers"; in the mediation of concepts, which consists of "facilitating access to knowledge and concepts for others, particularly if they may be unable to access this directly on their own"; and in the mediation of communication, which is essential to "facilitate understanding and to shape successful communication between users/learners who may have individual, sociocultural, sociolinguistic or intellectual differences in standpoint" (CoE 2020: 91). The mediation of texts, concepts and communication constitutes a set of essential processes that underpin communication across the diverse contexts in which it occurs. Without these processes, the level of communicative effectiveness—which relies both on contextual and situational factors and on the communicators' ability to construct and interpret meaning—is at risk of breaking down, potentially resulting in misunderstanding or even a complete communication breakdown.

When viewed in this way, mediation emerges as a "normal"—that is, inherent and integral—process within communication, and indeed as virtually synonymous with it, insofar as its function lies in the production, transmission and negotiation of meanings and sense. As Machetti and Siebetcheu (2017) argue, to communicate is to mediate, and this holds true whether communication occurs among users/learners who share the same language and culture or among those with different linguistic and cultural repertoires.

This perspective, moreover, aligns closely with what is proposed in the CEFR/CV, which conceptualizes mediation as a set of activities and strategies that vary according to the user/learner's level of linguistic-communicative competence and their specific communicative needs. This diversification follows the CEFR model and thus develops along a vertical dimension—a sequence of levels describing the learner's competence—and a horizontal dimension, aimed at identifying the domains and areas of use, communicative contexts, skills, and text types involved (Vedovelli 2010: 64). This framing conceptualizes mediation as one of the four modes in which the CEFR model organizes communication (reception, production, interaction, mediation). Therefore, mediation is a component that can be taught, learned, and assessed.

2.2 Mediation in language testing and assessment. The Italian context

Five years since the CEFR/CV was published, research aligning proficiency tests with mediation descriptors remains scarce in Italy. This scarcity is not limited to Italian language proficiency tests but extends to the broader spectrum of Italian as a second or foreign language (S/FL) assessments. More generally, it impacts the entire teaching and learning process of S/FL Italian. In certain respects, this process remains anchored to a traditional model, in which the explicit teaching of grammar continues to occupy a central role. In other instances, it reflects a communicative approach, where the different skills interact, yet in practice they continue to be taught and presented separately.

With regard to the assessment of L2 Italian, it can be stated that attention to mediation is at present almost entirely absent. For instance, in the Matura exams (Italian high school final exams), assessing Italian consists of writing an essay on literary, historical, philosophical or topical subjects. The assessment takes into account grammatical and lexical accuracy, the coherence and cohesion of the text, the argumentation capacity and the originality of the essay, without any reference to mediation. The same occurs in the Italian school system, when students are required to take national standardized tests designed to measure learning outcomes. These tests are administered by the *Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione* (INVALSI), and their respective components have been elaborated following a specific reference framework for Italian which provides the necessary specifications according to school grade. According to this framework, the aspects of language proficiency that are assessed consist of reading comprehension, reflection on language and lexical competence, also in this case without any reference to mediation.

However, and going back to proficiency tests, the four Italian language certifying bodies that constitute the CLIQ Association (www.associazionecliq.it), with one ALTE full member and three ALTE affiliates, are actively developing projects that are targeting mediation in L2 Italian teaching and assessment processes.[1] Among these is *Je le sais faire en Italien*, a project developed in collaboration with the French Ministry of Education, which has reached its second year of implementation. In this project, which involves high school students in their second and fifth years in Marseille, Aix-en-Provence and Nice, as stated in the reference syllabus, the writing skill is also intended as the ability to interact with written texts, precisely to highlight the focus on the changes proposed by the CEFR/CV.

Notably in some Italian high schools abroad, an assessment where mediation is explicitly or implicitly taken into account has positively influenced teaching syllabi, aligning with the principles of learning-oriented assessment (LOA; Turner and Purpura 2016). In this context, and in cooperation with the Scenario-Based Language Assessment (SBLA) Lab at Teachers College, Columbia University, the CILS Centre developed the Italian SBA project, which explores the development of SBA for L2 Italian. Grounded in the theoretical framework of LOA, the Italian SBA emphasizes collaborative problem-solving, integrates learning and assessment, and focuses on the development of topical knowledge (Peri 2025). Mediation is fully embedded in the SBA tasks, particularly in activities of synthesis, re-elaboration, virtual interaction and argumentative production (Peri in preparation). The project, which was also piloted in the Turkish school context, was conceived within a broader research initiative launched in 2019 between Teachers College and the CILS Centre at the University for Foreigners of Siena.

3 The study context

Our study focuses on CILS exams. The Certification of Italian as a Foreign Language (Certificazione di Italiano come Lingua Straniera or CILS) is an Italian language qualification offered by the CILS Centre of the University for Foreigners of Siena (Centro CILS 2019; Machetti and Vedovelli 2024).

The exams are standardized tests of Italian measuring general language proficiency. The CILS exams are premised on the six CEFR levels—CILS A1, CILS A2, CILS UNO-B1, CILS DUE-B2, CILS TRE-C1 and CILS

1. Since 1993 Italy has recognized four official certifications for Italian as a foreign language: CELI (Certificato di Conoscenza della Lingua Italiana), awarded by the Università per Stranieri di Perugia; CILS (Certificazione di Italiano come Lingua Straniera), awarded by the Università per Stranieri di Siena; IT (Italiano), awarded by the Università degli Studi Roma Tre; and PLIDA (Progetto Lingua Italiana Dante Alighieri), awarded by the Società Dante Alighieri. These institutions operate synergistically, with the common objective of promoting the study of Italian worldwide through certifications suited to various learning goals and professional needs; introducing and systematizing language testing practices in the Italian context; adopting assessment principles aligned with international standards; developing expertise in language testing and fostering international collaboration; supporting teacher training in this field through dedicated courses; advancing research on language assessment in the Italian context.

QUATTRO-C2—and are aimed at a general adult population considering an average cultural level of Italian language users studying for educational, professional or general cultural purposes. However, CILS is also available for specific public groups: for young learners at A1 and A2; for teenagers from A1 to B1; for migration purposes at A1 and A2; for citizenship at B1.

CILS exams are recognized by universities, employers, and institutions in Italy and worldwide. CILS exams were launched in 1993. In 2023, CILS exams were taken by 46101 test-takers. Since 1993, the CILS exams have been administered in Italy and in around 90 foreign countries. All CILS exams are paper-based and therefore delivered in person in CILS Test Centres in Italy and abroad (Machetti 2022).

4 Methodology

The methodology followed in this study adheres to the guidelines proposed by the handbook (British Council et al. 2022), thereby replicating its various phases. In the phase reported here, the study involved eight experts in language teaching and assessment (teachers, item writers, raters), along with two post-doctoral research fellows, all coordinated by a full professor who also served as head of the CILS Centre.

All participants had between five and fifteen years of experience in the field of L2 Italian teaching and assessment and were familiar with the procedures outlined in the Handbook. This familiarity stems from the fact that similar procedures had already been implemented as proposed by the manual for relating language examinations to the CEFR (CoE 2009). In addition, several of the experts involved in this study regularly used the manual in the context of CILS raters' continuing training and in the ongoing process of exam validation. The alignment process itself, consistent with the procedures set out in the manual, can be summarized in four principal stages. First, in the familiarization stage, professionals involved in the project developed a shared and detailed understanding of the CEFR scales and descriptors, as well as the content of the CILS exams. Second, in the specification stage, CEFR levels were assigned to each CILS exam task. Third, the standardization stage was implemented, consisting of training with calibrated CILS examples, benchmarking through judges' evaluation of sample performances, and standard setting to establish the CEFR level corresponding to different scores. Finally, in the validation stage, the internal validity of the procedures and the consistency of expert judgments were verified, complemented by qualitative analyses that included candidate feedback on exam content and administration.

In this article, we provide some evidence from the familiarization phase with the mediation descriptors, in addition to providing and analysing the main data from the specification phase. According to the handbook, "familiarization is designed to ensure that those involved in an alignment project have an appropriate knowledge of the CEFR and share a common understanding of the purpose of the project" (British Council et al. 2022: 19); and "specification analyses the content of any resource, existing or new, in terms of approach and coverage in relation to the categories presented in the CEFR" (British Council et al. 2022: 28).

In our study, the type of familiarization adopted was both generic and specific and so was the specification procedure. This means that familiarization took place in a single session and on an individual basis, but at the same time it also required group work, distributed across multiple sessions and guided by a coordinator. Specification involved a broad linking of the structure and content of the CILS exams to the content and level descriptors of the CEFR related to mediation; the specific procedure applied the same principles but at a much more detailed and fine-grained level. In addition, the method used for specification was the one indicated by the handbook, though it was used in a flexible manner. More precisely, the method involved an initial bottom-up phase in which analysis of existing exams (A2 and B1 levels) provided evidence of what they already covered on mediation (content analysis). Then, a top-down phase was conducted using the CEFR/CV as the basis for integrating the existing exams with mediation strategies and activities (needs analysis).

5 Data analysis and discussion

5.1 Familiarization

For the familiarization phase, forms 2.1, 2.2 and 2.3 of the handbook were used. Among the most interesting data that emerged from the use of form 2.1 are those of the majority of experts with reference to the relevance of the scales in the CILS exams context. The three mediation activities (mediating a text, mediating concepts, mediating communication) were in fact all judged important, but for different reasons.

Mediating a text, as suggested in the CEFR/CV, namely "mediating a text for oneself (for example in taking notes during a lecture) or in expressing reactions to texts, particularly creative and literary ones" (CoE 2020: 106), constitutes a fundamental set of scales because it represents what every test-taker does/should do when taking the CILS exam. As one expert observed, "every item/task involves the management of a text, and not only when it comes to creative and literary texts, because the development of linguistic-communicative competence is nothing but a process of text management, from text, about text, to text" (Vedovelli 2010).

In the A2 and B1 CILS exams, mediating a text constitutes the initial and essential step for comprehending the oral passages presented in the Listening Test. This process, which essentially entails taking notes on the main ideas or key words of an oral text, is indispensable for the completion of the subsequent task (multiple-choice or true/false). Without it, performance relies solely on memory, rendering the task either unfeasible or considerably more difficult.

Mediating concepts is a process which "involves two complementary aspects: on the one hand constructing and elaborating meaning and on the other hand facilitating and stimulating conditions that are conducive to conceptual exchange and development" (CoE 2020: 91) and refers to another important group of scales, because it makes the language test a tool for constructing and elaborating meaning and, at the same time, a tool that facilitates and encourages the elaboration and sharing of concepts. In this regard five participants emphasized how these scales reminded them of the LOA approach as applied in the Italian SBA, on which all the experts participating in this project have gained some experience (Purpura 2021; Peri 2025). LOA assigns a central role to the socio-interactional dimension, including such elements as turn-taking, repair strategies, feedback and the construction of social identity through interaction (Purpura and Turner 2018).

Finally, mediating communication, whose aim "is to facilitate understanding and to shape successful communication between users/learners who may have individual, sociocultural, sociolinguistic or intellectual differences in standpoint" (CoE 2020: 91) represents, according to all participants, a set of scales that should be one of the objectives of any language test that should apply every time a test-taker takes a test.

In the A2 and B1 CILS exams, this activity proves to be decisive for the completion of Task 1 of the Oral Production Test, which consists of a face-to-face conversation between the candidate and the interlocutor. As highlighted by Masillo and Machetti (2023), in cases where the interlocutor does not act to facilitate and make communication effective with candidates who typically present different individual, sociocultural, sociolinguistic, or intellectual characteristics, the administration of the test as a whole is seriously compromised.

In sum, the evidence provided by the experts in the familiarization phase sees mediation as a fundamental process of the test/within the test, which could even be thought of as a mediation tool, capable of creating "the space and conditions for communicating and/or learning, collaborating to construct new meaning, encouraging others to construct or understand new meaning, and passing on new information in an appropriate form" (CoE 2020: 103).

In relation to mediation strategies, experts repeatedly commented on them being described as "the techniques employed to clarify meaning and facilitate understanding" (CoE 2020: 126). Specifically,

one participant underscored that "taking a language test always requires clarifying meaning, even to oneself", e.g., when performing a task you are linking to previous knowledge, therefore you are adapting language and breaking down complicated information. This is a fairly common strategy adopted by candidates in certification exams. In some cases, however, it may be considered inadmissible. By definition, a language test should not provide evidence of knowledge and skills outside the construct under assessment, namely the language itself. Nevertheless, also with reference to the CEFR model of linguistic-communicative competence, a connection is recognized between this competence and extra-linguistic knowledge and skills. These, in turn, are essential for the development of pragmatic and sociolinguistic competence.

This evidence therefore suggests that experts were well aware of mediation strategies and their centrality, even if these strategies are not all explicitly and necessarily present in the current CILS exam construct.

5.2 Specification

The specification phase started from the analysis of existing exams (A2 and B2 levels) to collect evidence of what they already provide on mediation activities and strategies (content analysis). The procedure started by using Form 3.1, followed by the analysis of the test specification (in our case, CILS Guidelines for the A2 and B1 exams) using Forms 3.2 and 3.5 for which we used sample test items and tasks. The result for mediation activities, based as mentioned on the judgment of 11 participants, indicated that the mediation scales for which the current exams already provide evidence are the following:

a. in Writing and Speaking Tasks—mediating texts: Expressing a personal response to creative texts; Analysis and criticism of creative texts (e.g., *l'ultimo libro che hai letto; un film che ti è piaciuto particolarmente* [the last book you read; a film you particularly enjoyed])
b. in Writing and Speaking Tasks—mediating concepts: no evidence
c. in Writing and Speaking Tasks—mediating communication: no evidence.

This result suggests that the current CILS construct provides only minimal evidence of mediation activities and entails the need to broaden this construct with relevant mediation activities, which would lead to revising the format of the test itself in the direction of:

a. plurilingual assessment (from language A to language B), but also translation tasks
b. tasks requiring work within a group
c. tasks involving the description of tables, graphs

An analysis of the results of the Specification phase for mediation strategies showed that, for both the A2 and B1 exams, the mediation scales for which the current exams already provide evidence are the following:

a. in all parts of the exam: "Linking to previous knowledge". In fact, the entire test requires this recall in some way (e.g., To perform the writing task I need to retrieve a set of encyclopedic knowledge that is not necessarily provided by the input itself; at the same time, to understand a written text in the reading comprehension task, I need to do the same thing ...)

b. in all parts of the exam: "Adapting language". Particularly in the speaking task, it is necessary to adapt speech/speech speed

c. in the Reading Comprehension task: "Breaking down complicated information". This is an indispensable operation for the analytical reading of a text (e.g., *bando*).

This indicates that everything has to be made explicit in the actual construct; it is already there but needs to be more fully articulated and transferred to the tasks and the different items.

6 Future directions

The data collected and analysed so far are of fundamental importance for the completion of the alignment process, as they integrate the steps of standardization, standard setting and validation. These steps have been completed, but at present they still require a general revision. On the other hand, the data that will be collected at the end of this process for the A2 and B1 exams will be useful for extending the alignment process with the mediation descriptors to the exams of the remaining CILS levels. The aim is also to investigate whether it is possible to introduce mediation activities and strategies at levels A1 and pre-A1, in particular through the use of para-linguistic communication systems and translanguaging.

In addition, the alignment process appears to be useful for investigating the impact on candidates of assessment in relation to mediation activities and strategies. When candidates are student teachers, this impact should be measurable and of direct relevance to pupils, their families, and their overall school careers. A possible development of this research therefore concerns the engagement of various stakeholders in the educational field, fostering dialogue and debate among researchers and practitioners alike.

7 References

Bagna, Carla, Sabrina Machetti & Anna Maria Scaglioso. 2012. *Il collegamento fra gli esami CILS e il CEFR*. [The linking of the CILS examinations to the CEFR]. In Silvia Cacchiani, Silver Morgan & Marc Seth Silver (eds.), *Standardized language testing: Contemporary issues and applications*. Special issue of RILA. *Rassegna Italiana di Linguistica Applicata* 2012(1), 87-101. Rome: Bulzoni Editore.

Barni, Monica, Sabrina Machetti & Anna Maria Scaglioso. 2010. Linking the CILS examinations to the CEFR: The A1 speaking test. In Waldemar Martyniuk (ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*, 159-176. Cambridge: Cambridge University Press.

British Council, UKALTA, EALTA & ALTE. 2022. *Aligning language education with the CEFR: A handbook*. https://www.britishcouncil.org/exam/english/aptis/research/publications/cefr-handbook (accessed 27 January 2026).

Centro CILS. 2019. *Linee guida CILS*. https://cils.unistrasi.it/1/6/12/Le_Linee_Guida_CILS.htm (accessed 27 January 2026).

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. 2009. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual.* Strasbourg Strasbourg: Council of Europe.

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

Machetti, Sabrina. 2022. Le certificazioni di lingua italiana nei nuovi panorami linguistici globali. Spunti per un'analisi quantitativa. [Italian language certifications in the new global linguistic landscapes: Insights for a quantitative analysis]. *Studi Italiani di Linguistica Teorica e Applicata* LI(2). 452-476.

Machetti, Sabrina & Raymond Siebetcheu. 2017. *Che cos'è la mediazione linguistico-culturale*. Bologna: Il Mulino.

Machetti, Sabrina & Massimo Vedovelli (eds.). 2024. *Manuale della certificazione dell'italiano L2*. [Handbook of L2 Italian language certifications]. Rome: Carocci.

Masillo, Paola & Sabrina Machetti. 2023. Situazioni comunicative asimmetriche: Il test di produzione orale negli esami di certificazione della competenza in L2. [Asymmetrical communicative situations: the test of oral production in the L2 competence exam certificates]. In Valeria Caruso & Marta Maffia (eds.), *Vecchie e nuove forme di comunicazione diseguale: Canali, strutture e modelli. Studi AItLA 17*. Milan: Officina 21.A.

North, Brian & Enrica Piccardo. 2016. Developing illustrative descriptors of aspects of mediation for the Common European Framework of Reference (CEFR): A Council of Europe project. *Language Teaching* 49(3). 1-5.

Peri, Giulia. 2025. Standard per la mediazione linguistica. *Verso un approccio multimodale alle pratiche di valutazione linguistica* [Standards for linguistic mediation. Towards a multimodal approach to language assessment practices]. *Italiano LinguaDue* 17(2), 527-541. https://doi.org/10.54103/2037-3597/30425.

Peri, Giulia. 2025. *Topical knowledge in speaking performances: A scenario-based language assessment for L2 Italian*. Berlin: Peter Lang.

Purpura, James E. 2021. A rationale for using a scenario-based assessment to measure competency-based, situated second and foreign language proficiency. In Monica Masperi, Cristiana Cervini & Yves Bardière (eds.), *Évaluation des acquisitions langagières: Du formatif au certificatif. MediAzioni* 32. A54-A96.

Purpura, James E. & Carolyn E. Turner. 2018. Using learning-oriented assessment in test development (invited workshop). *Language Testing Research Colloquium*. Auckland, New Zealand.

Saville, Nick & Hanan Khalifa. 2016. The impact of language assessment. In Dina Tsagari & Jayanti Banerjee (eds.), *Handbook of Second Language Assessment*, 77-94. Berlin & Boston: De Gruyter Mouton.

Turner, Carolyn E. & James E. Purpura. 2016. Learning-oriented assessment in second and foreign language classrooms. In Dina Tsagari & Jayanti Banerjee (eds.), *Handbook of second language assessment*, 255-272. Boston, MA: De Gruyter.

Vedovelli, Massimo. 2010. *Guida all'italiano per stranieri: Dal Quadro comune europeo per le lingue alla sfida salutare* [A Guide to Italian for Foreigners: From the Common European Framework of Reference for Languages to the "Sifda salutare"]. Rome: Carocci.

8 Biographies

Sabrina Machetti is associate professor of educational linguistics at the University for Foreigners of Siena. She is the director of the CILS (Certification of Italian as a Foreign Language) Centre. Her research interests are focused on testing and assessing Italian as a second and foreign language, language learning and teaching, and language policy.

Giulia Peri earned her PhD in applied linguistics from the University for Foreigners of Siena. Her research interests include teaching and learning Italian as a second and foreign language, language assessment and technology. She is currently a research fellow at the CILS Centre.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Rethinking modern language education in the Netherlands: The CEFR as a compass for national targets[1]

Daniela Fasoglio, Netherlands Institute for Curriculum Development (SLO), Netherlands

https://doi.org/10.37546/JALTSIG.CEFR8-9

The Netherlands has been engaged in an ongoing debate about the need for curriculum reform. In 2021, the Dutch Ministry of Education commissioned the Netherlands Institute for Curriculum Development (SLO) to review national learning objectives, focusing on three key educational domains: qualification, socialization, and subjectification (Biesta 2020). One of the main challenges for modern language education lies in integrating the Common European Framework of Reference for Languages (CEFR) into the national learning targets while ensuring alignment with broader curriculum principles. A preliminary study (Fasoglio and Tammenga 2021) explored several scenarios based on quality criteria such as equity, horizontal and vertical alignment, and curricular coherence. A subsequent case study involving language teachers applied the methodology outlined in Aligning Language Education with the CEFR (British Council et al. 2022) to identify attainable proficiency levels for upper secondary education. As the process moves beyond the design phase, maintaining curriculum quality remains a key priority and depends on close collaboration among curriculum developers, school leaders, teachers, educational publishers, and test developers. Ensuring alignment between learning goals, pedagogy, and assessment is essential (Biggs and Tang 2011). Assessment, in particular, should promote coherence between the CEFR's vision of language learning and use and the goals of the national curriculum.

Keywords: curriculum reform, CEFR, modern language education, curriculum alignment, assessment

1 Background and context

In the last few years, a lively debate has taken place in the Netherlands among scholars, policy makers, and educators about the coherence, purpose, and relevance of school subjects in secondary education in view of the needs of young people to be equipped for participation in today's society. Some of the current national educational targets defined for all streams of upper secondary education have not changed since 2006 despite major shifts in society and technology. This is also the case for modern languages.

To address this, the Dutch Ministry of Education commissioned the National Institute for Curriculum Development (SLO) in 2021 to lead a comprehensive educational reform. The aim of this reform was to ensure that education equips students not only for examinations and employment, but also for life in modern society shaped by diversity, increased mobility, and challenges such as technological development, globalization, and the enhancement of democracy.

1. Many thanks to my dear colleagues Loes Groen and Stéfanie Leunissen for their valuable feedback on the text of this article.

The Ministry's vision builds on educational theorist Gert Biesta's framework of three interconnected core educational purposes: qualification, socialization and subjectification (Biesta 2021, Biesta 2023). While qualification focuses on knowledge and skills for further education or the job market, socialization helps students find their place in society by engaging with democratic values and with different cultures. Subjectification encourages students to grow as individuals, to think critically, to act ethically, and to relate meaningfully to the world around them.

Biesta's approach signals a shift away from education that is primarily results-driven and is centred on measurable results. Instead, it emphasizes the complexity of teaching and learning and values education as a means to support and facilitate students' development by making learning meaningful. This view is based on principles of equity and inclusion: every student, regardless of background, ability or ambition, should have the opportunity to grow intellectually, socially and personally.

Building on Biesta's ideas, updating the national educational targets is not just about new content; it is about rethinking the very purpose of education itself. In a world marked by cultural diversity and rapid changes, schools should reflect the plural linguistic and cultural identities of their students and ensure that the curriculum supports both academic and professional achievement and social justice.

2 Pre-analysis: identifying the context and needs of modern language education

Modern foreign language curricula were, like other subjects in the Dutch curriculum, to be renewed based on Biesta's framework, mentioned in the previous paragraph. Language teachers face difficulties in secondary education regarding contents of the curriculum and of national exams, restrictions on lesson time, and lack of appeal of language subjects. Indeed, language curricula are in need of a boost.

Our first activity consisted of a contextual analysis to assess the needs for modern foreign language curriculum reform in our country (Fasoglio and Tammenga 2021). We initially reviewed relevant written sources: research findings, literature and other (online) publications. We also explored a few comparable contexts in which curriculum reform was taking place or had just been completed (specifically Finland, Ireland and New Zealand). Afterwards, we consulted a number of subject experts, representatives of the national teachers' association, teacher trainers, test experts and researchers. Consultations first took place via written feedback on draft texts, followed by in-depth interviews in focus groups. In our analysis we mapped out relevant and current developments in educational policy, linguistic research, language educational practice, and society in order to lay a solid foundation for the pillars of our reform, and we described a few promising practices. We concluded our analysis with a few recommendations to face the challenges of a substantial renewal of modern language curricula. Most important was the need for an integrated approach to communicative language skills and the inclusion of new subject-specific content in goals, assessment and implementation of the curriculum. Specifically, we mentioned aspects of digital literacy and cultural awareness; knowledge about and reflection on language; alignment with goals, audience, medium and sociocultural context of communication; creative language use; insights in literary texts and other fictional texts such as movies, graphic novels and narrative games. In addition, we affirmed the need to investigate how aspects of plurilingualism and language awareness should take a prominent place in national objectives. In this regard, the *Companion Volume* to the *Common European Framework of Reference for Languages* (CEFR/CV, Council of Europe [CoE] 2020) should provide a framework for aligning objectives, teaching, and assessment and would therefore need to be assigned a visible, official status in the new national targets. Full details of the contextual analysis can be found in Fasoglio and Tammenga (2021).

3 Setting up the renewal process of modern language national targets

The renewal process covered all nine language subjects that have national attainment targets in Dutch secondary education: Arabic, Chinese, English, French, German, Italian, Russian, Spanish and Turkish. As curriculum experts, we developed new examination programmes in close collaboration with language teachers and teacher trainers. We established a curriculum renewal committee consisting of sixteen modern language teachers working in pre-vocational, senior general and pre-university education, and eight language teacher trainers. As curriculum experts, we were responsible for the substantive steering and for ensuring content quality and consistency. The committee started its work in June 2022 and completed it in June 2024. Over this period, we worked in twelve two-day sessions supplemented by meetings in smaller groups, sometimes involving external experts in specific topics such as intercultural competence, digital literacy, and language awareness. Throughout the process, an advisory group provided feedback on draft texts. The advisory group included representatives from the association of foreign language teachers, faculties of humanities, the modern language teacher training network, civil society organizations, and educational publishers. Additionally, two consultations were held with secondary school students and feedback was gathered from Cito, the Dutch institute for test development. Educational researchers from the SLO Advice & Research Department developed an instrument to systematically monitor the consistency of our work in relation to general principles and quality criteria during the design and development phases.

4 Design framework: capturing the core of language learning and teaching

Together with the curriculum reform committee, we formulated a mission statement for the specific area of modern language learning based on Biesta's framework and our preliminary analysis. Both reinforced our beliefs that being aware of how language works, being aware of the cultural implications of language use, and being able to communicate in multilingual contexts belong to the core competences in modern society. Based on these beliefs, we outlined the relevance of modern language teaching and learning in compulsory education, and how it contributes to qualification, socialization and subjectification. We defined its purpose as equipping students with the knowledge, skills, and attitudes to become *proficient, language-aware and culturally aware communicators* in both digital and non-digital multilingual and multicultural contexts. Education in modern languages fosters self-confidence, autonomy, reflection, and creativity in communication. Through exploring, broadening and deploying their plurilingual repertoire in learning and communicating, students expand their horizons, discover their own talents, preferences and opportunities, and develop a deeper understanding of language and culture. They learn to critically engage with diverse media and sources in other languages, enabling them to communicate effectively and appropriately. Education in modern languages empowers students to continue developing their plurilingual repertoire autonomously, both in and beyond the classroom. This will enhance their opportunities in further education and career pathways. By acquiring knowledge and skills in other languages and cultures, students become aware of their own plurilingual and pluricultural identities and potential and learn to approach differing cultural perspectives with openness.

The mission statement, as mentioned above, is built on three pillars. First of all, language users need language in order to perform all kinds of tasks and to achieve various goals—often collaboratively—in diverse contexts, and thereby to participate in society as active citizens. In doing so, they learn to use their entire language repertoire effectively, and to consciously relate to cultural aspects that can influence communication. This requires a holistic and integrated approach to language teaching (Piccardo and North 2019). Placing language learning and use in a social perspective reflects the rationale behind the CEFR (CoE 2020: 30) which makes the CEFR perfectly suitable as a framework for designing and constructing new national targets.

The second pillar is that language subjects encompass not only language skills but also have subject-specific contents that are linked to language as a social phenomenon and as part of identity (e.g., Michel

2024). Think of knowledge and awareness of language structures, similarities and differences among languages and language varieties; of various effects of language contact; of the emotional, social, and even political dimensions of language use (James and Garrett 1991; van den Broek et al. 2022).

Finally, the third pillar is that language and culture are deeply intertwined (e.g., Byrnes 2010). Linguistic expressions are not just a tool for communication but are also carriers of culture.

In line with the above, we developed a framework for the new national educational targets consisting of three domains: A: Communication, B: Language awareness, and C: Cultural awareness. This construction is visualized in Figure 1

Figure 1. Visualization of the content of the new national educational targets in three domains, and their interconnections (Fasoglio et al. 2025)

This diagram shows three overlapping circles for the domains A (light blue), B (dark blue), and C (orange). Each circle summarizes the main contents of the domain. Their overlap in the middle emphasizes that all three domains are essential for effective communication. This involves considering others' perspectives, being aware of cultural aspects and taking them into account, and being aware of one's own language proficiency, usage and impact. Rich, varied and meaningful experiences with the target language in different contexts are vital for building these competencies.

5 Embedding the CEFR

Modern language attainment targets in Dutch secondary education have been linked to the CEFR since 2007. The link to CEFR levels was not originally based on empirical data about what students actually achieve, but on a comparison between national requirements up to 2007 and the descriptions of the CEFR levels. Between 2013 and 2017, Cito carried out an international standard setting for receptive skills. Cito and SLO (in collaboration with Cambridge Assessment for English speaking skills) researched the CEFR attained level for English, French and German writing and speaking skills for some of the educational strands. A summary of these research results and the link to the research reports can be found on the Dutch CEFR portal (SLO 2024; see webpage https://www.slo.nl/thema/vakspecifieke-thema/mvt/erk/leraar-vo/niveaus-vo/bereikte-niveaus/).

A few fundamental discrepancies still persist. Whereas proficiency targets are related to CEFR levels, those levels are not laid down in law. The national school-leaving reading test, accounting for 50% of the final grade, is not based on the CEFR. School boards are responsible for the quality of the school-based tests accounting for the remaining 50%. Diversity among schools regarding the implementation of the CEFR is huge. Guidelines for school-based exams provide guidance for aligning tests to the CEFR; however, they are not legally binding. Language teachers have long been asking for more clarity to help them gain insight into the proficiency level of their students, their progression, and their learning pathways. The renewal of the national educational targets offered opportunities to address these issues.

Two main questions had to be answered during the renewal process. Firstly, how to embed the CEFR in the national targets in such a way that it removes the discrepancies previously mentioned and provides increased guidance to teachers, students, test developers and educational material developers, within the generally prescribed structure for all national attainment targets? Secondly, which CEFR levels for the different languages, language skills and pathways are challenging and at the same time achievable for the different language subjects and educational pathways?

5.1 Ways of embedding

In order to answer the first question, we conducted a preliminary study (Fasoglio et al. 2022) to determine the best way to formalize the CEFR in national targets for secondary education. First, we assessed the needs of the educational sector with regard to the implementation of the CEFR. Based on the aspects that were put forward, we described various scenarios for embedding the CEFR in national targets. The scenarios corresponded to different variables such as: status of the proficiency levels (either advisory or mandatory); ways of differentiating between educational streams; bandwidth (either A, B, C or A1, A2 etc.); implicit or explicit form of processing level descriptions (either adopt the *Can Do* descriptors verbatim or incorporate level indicators in the targets); language or skill level (either one overall CEFR level for a foreign language or specified per skill); level of specificity (what scales and indicators were suitable). We assessed and weighed the scenarios in relation to three curriculum-related quality criteria: equality of opportunity (all students should have the same chance to develop their potential, regardless of their background or other personal circumstances), internal consistency, and horizontal and vertical alignment. We also looked at quality criteria and conditions such as expected usability and expected effectiveness: which of the scenarios would best enable these criteria to be met? Our analysis led to the recommendation that attainment targets should be formulated based on the CEFR global descriptors and that they should be accompanied by explicit CEFR level indications to be included in the preamble. These should be specified for each individual language, language skill and educational stream. Level indicators within the Can Do descriptors of the CEFR needed to be identified and included in the formulation of the attainment targets (see Table 1 for an example).

Table 1. *Draft attainment target for written production and interaction, English senior general education (CEFR B1 level)*

Attainment target 8
The student writes in English, targeted to purpose, audience, context, and medium. This involves: • exchanging information, experiences, feelings, and thoughts in written informal and formal online and offline interaction about familiar topics; • producing straightforward informative and narrative texts with the use of digital tools where applicable; • using appropriate register and conventions; • adjusting language to sociocultural conventions and the perspective of the communication partner; • using varied language structures and expressions.

5.2 CEFR: what are challenging and attainable levels?

The second question we needed to answer was, which CEFR levels for the different languages, language skills and educational streams are challenging enough and at the same time achievable by the end of upper secondary education? In order to make well-founded statements, the curriculum renewal committee collected available information from previous research on the levels attained by students nationwide at the end of upper secondary education. However, only the CEFR levels achieved for English, German, and French—and only for a limited number of language skills and educational streams—had been previously studied. In 2023, further research was conducted in the form of a case study amongst language teachers to determine which CEFR levels are considered both challenging and achievable for upper secondary education (Groen and Trimbos 2023). The methodology of the case study was drawn from *Aligning Language Education with the CEFR* (British Council et al. 2022). The handbook is based on the CEFR/CV and focuses not only on assessment, but also on policy, curriculum and teaching materials, which made it a good fit for our exploration. The fact that the handbook is designed for people who are engaged with the CEFR in a practical way also made it suitable for the purpose of the case study.

The handbook introduces the steps required to align language curricula to the CEFR: familiarization, specification, standardization, standard setting and validation, and offers tools and materials that you can use to document the various steps. Due to time constraints and in view of the specific aims of the case study, the procedure described in the handbook was slightly adapted with respect to the following aspects:

a. For English, French and German, the focus was on familiarization, specification, standardization and standard setting. The focus for the other languages was on familiarization.

b. For all languages, an extra joint session was added for discussion and alignment of estimated levels stemming from the procedure applied in the previous rounds.

c. For all languages, an extra joint session was added to discuss preconditions for implementation.

d. For English, French, German and Spanish, an extra session was added with language institutes to evaluate the outcomes of the case study (triangulation).

Criteria for teachers to participate in the case study included: being qualified for teaching one of the languages involved in the curriculum renewal; being familiar with the CEFR; teaching in exam classes while participating in the case study; using their own classroom materials during the group sessions. Approximately twenty sessions were organized:

- For English, French and German separately: introductory sessions and whole-day sessions, distinguishing between the different educational streams, plus two additional short online meetings for the pre-vocational stream due to insufficient time on the day of the sessions.
- For Spanish: a homework assignment followed by an online session in which teachers assessed the conclusions reached for French and compared them with their own teaching practice.
- For Arabic, Chinese, Italian, Russian and Turkish: a homework assignment followed by a half-day session.

More information about collected data and findings can be found in Groen and Trimbos (2023).

6 Follow-up

The results of the case study provided useful indications of the CEFR levels that were estimated to be achievable at the end of upper secondary education. This information—though not representative of the entire target group—supplemented the data that the curriculum renewal committee had collected from previous research. During the development process, the committee discussed their findings with experts, teacher trainers and representatives of language teachers; they compared the outcomes of the discussions with their own experiences and placed them in the context of continuous learning pathways starting from lower secondary education. An overview of the CEFR levels established by the curriculum renewal committee per language, language skill and educational stream, as well as a justification per language was published at the end of the development process (Fasoglio et al. 2025; see also Table 2). Meanwhile, the Ministry of Education adopted the recommendation from the exploratory study on the formalization of the CEFR in the new modern language attainment targets (Fasoglio et al. 2022).

Table 2. *Proposed minimum required CEFR levels for English and the end of upper secondary education in the Netherlands*

	Pre-vocational				
	Basic	Middle management	Combined/ theoretical	Senior general	Pre-university
Oral and audio-visual comprehension	A2	B1	B1+	B2	C1
Reading comprehension	A2	A2	B1	B2	C1
Oral production	A1	A2	B1	B2	C1
Oral interaction	A1	A2	B1	B2	B2+
Written production and interaction	A2	A2	B1	B1	B2

The new educational targets and the established CEFR levels still need to be tested in schools prior to being formalized by the Ministry of Education and implemented. Starting in September 2025, a trial phase is being carried out by SLO, during which we engage in discussions at schools with language teachers, students and school leaders. With them, we discuss what information, guidance, examples and conditions are needed to shape the intent and the content of the draft educational targets in teaching and assessment. This will lead to the development of a variety of support materials for schools and their teachers. These materials are meant to inspire and are non-prescriptive. They aim to help schools align their programmes of teaching and assessment to the new educational targets.

7 Looking ahead: opportunities and challenges

The national educational targets for modern foreign languages have been renewed based on the idea that language subjects add value to students' development and are essential for their participation in today's multilingual, multicultural society. This idea has been explained in our mission statement. To make students aware of the added value of mastering multiple languages, new or updated content has been included in the attainment targets. The trial phase is a suitable moment to ask students and teachers whether that added value is recognized, and whether the new draft educational targets have succeeded in making it tangible.

CEFR-based attainment targets for modern languages allow schools room for customization and differentiation. Teachers can stimulate students to develop and capitalize on their entire language repertoire and their plurilingual competences in the light of what is important to them for their personal development and participation in society. The more clearly targets are defined, the easier it is to understand what customization is needed to ensure that all students reach those goals.

However, curriculum reform does not automatically lead to better education. Curriculum quality relies on the teacher's critical involvement and good interplay between curriculum developers, teachers, school leaders, testing developers, educational publishers and students. Curriculum making is a social practice (Priestley et al. 2021). It is a non-linear, dynamic process of interpretation, mediation, negotiation and translation across multiple layers of education: from international frameworks and guidelines (in our case, the CEFR) to national requirements, school policy and classroom activities. Curriculum making is, in other words, a fascinating but complex challenge. Formal curricula cannot just be uncritically adopted by teachers; they need to be translated to the school context in a consistent and context-specific way. Therefore, teachers and school leaders need to develop curriculum agency and an enquiring and reflective attitude. To achieve this, shared sense-making and critical engagement with the aims and values of the formal curriculum are essential initial steps. At the same time, sustainable school curriculum changes need a strong synergy between three closely interrelated areas: curriculum development, teacher professional development and school organization development (Priestley et al. 2021). This can only be achieved if schools shape a discursive environment that promotes professional development and cooperative thinking.

When it comes to the implementation of the CEFR, the procedures described in the alignment handbook (British Council et al. 2022), possibly adjusted as in Groen and Trimbos (2023), can show their effectiveness in facilitating deep sense-making among language teachers in developing shared curriculum agency. This includes reflection on how to achieve constructive alignment between goals, implementation at school, methodologies and assessment. Constructive alignment is, indeed, an essential condition to make high-quality and sustainable curricula possible, both at the intended and implemented level (Biggs and Tang 2011).

Additionally, further research and trialling is needed for high-stakes assessment that is valid and reliable, both central and school-based, to define suitable assessment constructs and rating systems, as the CEFR itself is not really designed for rating purposes. Intensive synergy between curriculum and test developers is crucial to set out assessment specifications, forms and methods that fit the intended

purpose of the new national targets, and at the same time meet feasibility conditions in their contexts of use such as procedures, resources, time, costs, etc. In line with the principles of the entire educational reform, assessment should synchronously stimulate an approach that prioritizes promoting and improving student learning and learning autonomy. This is another challenging journey, where the renewal of the national attainment targets serves merely as the starting point.

8 References

Biesta, Gert. 2020. Risking ourselves in education: Qualification, socialisation, and subjectification revisited. *Educational Theory* 70(1). 89-104. https://doi-org.proxy2.library.illinois.edu/10.1111/edth.12411.

Biesta, Gert. 2023. Becoming contemporaneous: Intercultural communication pedagogy beyond culture and without ethics. *Pedagogy, Culture & Society* 31(2). 237-251. https://doi.org/10.1080/14681366.2022.2164341.

Biggs, John & Catherine Tang. 2011. *Teaching for quality learning at university: What the student does*, 3rd edn. Columbus, OH: McGraw-Hill Education.

British Council, EALTA, UKALTA & ALTE. 2022. *Aligning language education with the CEFR: A handbook.* https://ealta.eu/documents/resources/CEFR%20alignment%20handbook.pdf (accessed 21 Oct 2025).

Byrnes, Heidi. 2010. Revisiting the role of culture in the foreign language curriculum. *The Modern Language Journal* 94, 315-336.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume*. Strasbourg: Council of Europe. https://rm.coe.int/commoneuropean-framework-of-reference-for-languages-learning-teaching/16809ea0d4 (accessed 21 Oct 2025).

Fasoglio, Daniela, Machteld Moonen & Marjon Tammenga. 2022. *Verkenning voor de formalisatie van het Europees Referentiekader voor Talen in het Nederlandse onderwijs*. [Exploration for the formalization of the European Framework of Reference for Languages in Dutch education]. Amersfoort: SLO.

Fasoglio, Daniela & Marjon Tammenga. 2021. *Startnotitie Moderne Vreemde Talen. Bovenbouw voortgezet onderwijs* [Preliminary policy document modern foreign languages. Upper secondary education]. Amersfoort: SLO. https://www.slo.nl/@20066/startnotitie-moderne-vreemde-talen/ (accessed 21 Oct 2025).

Fasoglio, Daniela, Stéfanie Leunissen & Marjon Tammenga. 2025. *Toelichtingsdocument. conceptexamenprogrammema's moderne vreemde talen voor vmbo, havo en vwo. Versie 2.* [Explanatory paper draft examination programmes modern foreign languages for pre-vocational, general and pre-university education. Version 2.]. Amersfoort: SLO. https://www.slo.nl/thema/meer/actualisatie-kerndoelen-examenprogramma/actualisatie-examenprogramma/examenprogramma-moderne-vreemde-talen/@24933/toelichtingsdocument-mvt-vmbo-havo-vwo/ (accessed 21 Oct 2025).

Groen, Loes & Bas Trimbos. 2023. *Verkenning ERK-niveaus* [Exploration of CEFR levels]. Amersfoort: SLO.

James, Carl & Peter Garrett. 1992. *Language awareness in the classroom*. London: Longman.

Michel, Marije. 2024. *Un strålande futuro needs mehr dan ðeə langues*. Inaugural lecture. University of Groningen, 16 February. https://doi.org/10.21827/65688fb08df31.

Piccardo, Enrica & Brian North. (2019). *The action-oriented approach: A dynamic vision of language education.* Bristol: Multilingual Matters. https://doi.org/10.21832/PICCAR4344.

Priestley, Mark, Stavroula Philippou, Daniel Alvunger & Tiina Soini. 2021. Curriculum making: A conceptual framework. In M. Prestley, S. Philippou, D. Alvunger & T. Soini (eds.), *Curriculum making in Europe: policy and practice within and across diverse contexts.* Emerald.https://www.emerald.com/insight/publication/doi/10.1108/9781838677350

SLO. 2024. Het ERK [The Dutch CEFR portal]. https://www.slo.nl/thema/vakspecifieke-thema/mvt/erk/ (accessed 21 Oct 2025).

van den Broek, Ellen, Helma Oolbekkink-Marchand, Ans Van Kemenade, Pauline Meijer & Sharon Unsworth. 2022. Stimulating language awareness in the foreign language classroom: Exploring EFL teaching practices. *Language Learning Journal* 50(1). 59-73. https://doi.org/10.1080/09571736.2019.1688857.

9 Bibliography

Daniela Fasoglio is senior curriculum developer and modern language specialist at the Netherlands Institute for Curriculum Development (SLO). She is involved in the thorough revision of the language curricula in Dutch secondary education. She has coordinated several CEFR dissemination and implementation activities and has been a member of the CEFR expert group, and later the language policy expert group at the Council of Europe. Her expertise includes curriculum analysis, design and evaluation, as well as development and implementation processes. She is engaged in the integration of such issues as culture, mediation and plurilingualism in language education.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

The CEFR in Japan: A tale of two approaches in English and Japanese language teaching

Masashi Negishi,Tokyo University of Foreign Studies. Japan
Yukio Tono,Tokyo University of Foreign Studies, Japan

https://doi.org/10.37546/JALTSIG.CEFR8-10

This article explores the varying degrees of implementation and influence of the Common European Framework of Reference for Languages (CEFR) in Japan, focusing on a comparative analysis of its introduction to English language teaching (ELT) and Japanese language teaching (JLT). Drawing on a framework that operationalizes key CEFR concepts—the action-oriented approach, the role of social agents and proficiency levels—this study analyses curriculum documents, textbooks and assessment tools. The findings reveal a marked difference in the adoption strategies: a cautious, laissez-faire approach in ELT, and a more top-down, mandated approach in JLT. The ELT context demonstrates inconsistent alignment, with some progressive teachers and materials developers filling gaps left by national curricula, while the Japanese language education context shows a strong, albeit not yet widespread, alignment in accredited institutions. The article concludes by discussing the inherent "power" of the CEFR, not as a prescriptive standard, but as a framework that can drive reform, highlighting the need for targeted training and support to achieve broader, more uniform impact.

Keywords: CEFR, CEFR-J, English language teaching, Japanese language teaching, the action-oriented approach (AoA), social agents, proficiency levels, the power of the CEFR

1 Introduction

The *Common European Framework of Reference for Languages* (Council of Europe [CoE] 2001) has emerged as a globally significant tool for language education, providing a common basis for the description of language curricula, syllabuses, textbooks and assessments. The CEFR's core principles — the AoA, the concept of the language learner as a social agent (SA), and the illustrative proficiency scale — are believed to have had a significant influence on approaches to teaching and assessment in different parts of the world. Despite the CEFR's original intent not to be prescriptive, its adoption by national education systems often involves complex adjustments to existing educational traditions and institutional structures.

This article examines the dynamics of CEFR implementation in Japan by comparing two distinct contexts: English language teaching (ELT) in secondary schools and Japanese language education for foreign learners. Both contexts have been influenced by the CEFR, but their pathways to adoption and their resulting degrees of alignment differ significantly. The analysis is based on our observations, a survey of textbooks and tests, and an examination of official documents from the Ministry of Education, Culture, Sports, Science and Technology (MEXT). By contrasting these two cases, this study aims to shed light on the conditions under which the CEFR can effectively drive educational reform and the challenges that arise when its principles are introduced into different institutional environments.

2 Conceptual framework and operationalization

The CEFR views language users and learners as 'social agents' who accomplish tasks in specific contexts and fields of action. The AoA is a key concept of the framework, moving beyond decontextualized language exercises to focus on meaningful communication. As Piccardo and North emphasize, "the real-life task, where social agents are engaged, is the core of the AoA, since it provides the unifying frame within which all actions make sense and serve a purpose" (2019: 40-41).Furthermore, in this paradigm, assessment is based on "what the social agent can do in a real situation" (397). It should be noted, however, that while classroom tasks designed for learning purposes cannot really be considered as authentic and real-life in the sense that Piccardo and North mean, such tasks can be seen as having their own intrinsic pedagogical authenticity and they can serve as useful proxies for real-life tasks in the world beyond the classroom.

For the purpose of this analysis, we have operationalized the key concepts of the CEFR, i.e., task authenticity, social agents and the levels, to evaluate the alignment of educational materials and assessments. Task authenticity is used to measure the implementation of the AoA. Tasks are judged to have high authenticity if they are meaningful, contextualized communicative activities, as opposed to decontextualized drills. The concept of social agents is operationalized by examining the presence of a specific, identifiable context in test and textbook instructions. For example, a task is considered to involve a social agent if the instruction places the learner in a given situation (e.g., "You are a customer in a restaurant …"). This practice, which has roots in communicative language testing, is crucial for assessing how well the CEFR's social agent concept is being integrated. Finally, the levels of the CEFR are evaluated by determining whether a program or assessment explicitly specifies CEFR proficiency levels.

This framework allows for an analysis of the degree of alignment, recognizing that CEFR integration is not a binary state but a continuous spectrum. This continuous model is essential for capturing the nuanced implementation observed in Japan's language educational landscape. Textbooks serve as a means of delivery, while classroom assessments and entrance examinations represent the assessment element of the education system.

3. The case of English language education in Japan

3.1 The CEFR-J project: the early years (2004-2012)

To contextualize the revision of English language education in Japan, we must first introduce the CEFR-J project (Negishi et al. 2013). This initiative comprised several government-funded research projects that investigated how to construct a language framework like the CEFR and how to properly adapt it to the Japanese context. Prior to the CEFR's publication, there was virtually no comprehensive understanding of Japanese learners' English proficiency distribution—only fragmented data from local examinations such as the EIKEN Tests existed. When the CEFR was published in 2001, it initially attracted attention from only a small group of specialists, some of whom did not fully grasp the meaning of CEFR levels and even claimed that we should aim at C level.

In 2004, we launched the KAKEN project, collecting information on international proficiency benchmarks for English while researching domestic English proficiency guidelines across primary, secondary and tertiary education levels. During our search for international standards we encountered the recently published CEFR. Its growing European influence and impressive scope, size and depth led us to abandon our plans to create original proficiency guidelines. Instead, we decided to explore the CEFR's potential as a "descriptive tool" to analyse the situations, needs, and goals of English language teaching in Japan, and to determine how it should be adapted for this purpose.

In 2008, a newly funded KAKEN project called the *CEFR-J project* brought together 15 researchers from various fields of applied linguistics. To better understand the CEFR's design and construction process, we created our own version of CEFR-aligned calibrated descriptors. Since English language learners in

Japan were predominantly at lower proficiency levels, we focused our branching on lower CEFR levels, as recommended in the CEFR (CoE 2001: 32). We developed the CEFR-J *Can Do* descriptors based on the following principles:

1. Add Pre-A1
2. Divide A1 into three sublevels: A1.1, A1.2, A1.3
3. Divide A2, B1, and B2 into two sublevels: A2.1, A2.2; B1.1, B1.2; B2.1, B2.2
4. No change for C1 or C2
5. Adapt *Can Do* descriptors to the Japanese context

We carefully developed and calibrated the descriptors in accordance with the CoE guidelines. Project members were assigned to five skill areas—spoken interaction (SI), spoken production (SP), listening (L), reading (R), and writing (W—for descriptor development. The format of the CEFR-J descriptors followed that of the self-assessment grid (CoE 2001: 26-27). While we were aware of the four modes of communication (reception, production, interaction and mediation) distinguished in the CEFR, we adopted the terms used in the self-assessment grid—understanding, speaking, and writing—which were more familiar to stakeholders at that time. In 2009, we invited Dr Tony Green from the University of Bedfordshire to conduct a workshop on descriptor development. Following his recommendations, we analysed our original descriptors by breaking them down into three components: (1) task [action], (2) condition and (3) criteria [or text for receptive skills]. Tables 1 and 2 show examples of these. We consulted the European Language Portfolio's descriptor list (Lenz and Schneider 2004) and revised our descriptors to include all three components, ensuring they were comparable across difficulty levels.

Table 1. *An example of a "broken-down" CEFR-J spoken interaction Can Do descriptor*

A1.3 Spoken Interaction	Performance	Criteria (Quality)	Condition
I can ask and answer simple questions about very familiar topics (e.g., hobbies, sports, club activities), provided that people speak slowly and clearly with some repetition and rephrasing.	I can ask and answer … questions about … topics (e.g., hobbies, sports, club activities)	simple very familiar	provided that people speak slowly and clearly with some repetition and rephrasing.

Table 2. *An example of a "broken-down" CEFR-J Listening Can Do descriptor*

B2.1 Listening	Task	Text	Condition
I can follow extended speech and complex lines of argument provided the topic is reasonably familiar.	I can follow	extended speech and complex lines of argument	provided the topic is reasonably familiar.

In 2010, we completed the alpha version of the 120 CEFR-J descriptors covering Pre-A1 to B2.2 across five language activities (L, R, SI, SP, W). This approach was necessary because our main goal was to understand the construction process of the CEFR-like framework, and creating hundreds of Can Do descriptors similar to the ELP was not feasible. In summer 2010, we conducted an extensive sorting

exercise with 241 primary and secondary school teachers. The results were encouraging, with an overall Spearman rank-order correlation of ρ = .928. Based on these findings, we made minor revisions to the alpha version and prepared a beta version for a large-scale Can Do questionnaire survey in 2011.

The 2011 survey included 5,468 participants (1,685 lower secondary, 2,538 upper secondary, and 1,245 university students). Unlike the original CEFR descriptors, which were calibrated against teachers' perceptions of learner proficiency, we delivered our survey directly to students. This approach was necessary because Japanese class sizes are much larger than those in Europe, making it unlikely that teachers could accurately judge what individual students can do. After conducting IRT analysis, we revised the descriptors and released version 1 of the CEFR-J in March 2012.

3.2 The CEFR-J project: the RLD and implementation (2012-2024)

The release of the CEFR-J raised awareness among English language teaching communities in Japan about the CEFR's growing influence globally. For example, the EIKEN Tests, Japan's most widely used English assessment, began aligning its test grades with CEFR levels. In 2013, MEXT published a booklet introducing Can Do descriptors and encouraged secondary schools to create their own statements. This initiative achieved limited success, however, as MEXT did not direct schools to use the CEFR as a reference point, leaving many schools struggling to define appropriate descriptor content across proficiency levels. Nevertheless, the CEFR-J succeeded in making the CEFR more accessible to Japanese educators.

In the decade following Version 1's release, the CEFR-J project concentrated on two areas: developing Reference Level Descriptions (RLDs) for the CEFR-J and supporting various stakeholders—MEXT, local education boards, schools and publishers—in using the CEFR-J according to their specific needs.

3.2.1 CEFR-J RLD Project

Reference Level Descriptions (RLDs) involve developing resources for a specific language to make the CEFR concrete by detailing the lexis, grammar, and functions needed at each level (CoE 2005). As part of the CEFR-J project, we explored how to develop RLDs effectively using the following data-intensive methods. We assembled a new KAKEN research team with specialists in corpus linguistics, machine learning and language assessment. Between 2012 and 2020, we developed textbook and learner corpora classified by CEFR levels and evaluated machine learning methods to identify key linguistic features for level classification (Tono 2013). This work produced several open resources: the CEFR-J Wordlist, Grammar Profile, Text Profile and Error Profile. Full details are available on the CEFR-J website (https://www.cefr-j.org/download_eng.html).

Another significant initiative was the CEFR-J Can Do Test—a set of performance assessments measuring users' ability to complete tasks described in specific descriptors. We created test samples for 100 CEFR-J descriptors, with sample versions publicly available on the CEFR-J website (https://www.cefr-j.org/download_eng.html#cefrj_testasks).

3.2.2 Encouraging the use of the CEFR-J for ELT in Japan

Between 2018 and 2020, MEXT formed a working group to revise the national curriculum (Course of Study), with one of the present authors (Yukio Tono) serving as a committee member. Tono informed committee members about the CEFR's international influence and the CEFR-J project's work. MEXT subsequently incorporated several CEFR principles into the Courses of Study. The 2020 curriculum for secondary schools divides speaking into spoken production and spoken interaction, mirroring the original CEFR categories. Teaching objectives are now framed as Can Do statements, and vocabulary requirements reference the CEFR-J word list. Despite these CEFR-influenced changes, MEXT avoided explicitly stating that the Course of Study was aligned with CEFR levels, deeming such policy declarations

inappropriate for an official legal document. Consequently, those unfamiliar with the CEFR might not recognize the CEFR's influence on the revised curriculum.

In our final KAKEN project (2020-2024), we collaborated with two local schools—a lower-secondary school (LSS) in Saitama City and an upper-secondary school (USS) in Kyoto—to integrate CEFR perspectives into their programmes. For the Saitama project, we collected students' written production data over two years. Using tasks partially based on the CEFR-J Can Do Test, we administered the same writing assignments throughout this period to track students' progress from A1 to A2. In the Kyoto project, we analysed one lesson from the English textbook in detail. We partnered with a Japanese English teacher who invited us to participate in lesson planning, task development and assessment using CEFR-J resources. We video-recorded and transcribed all classes for this lesson to create the CEFR-J Classroom Observation Corpus. In this corpus, we annotated all teacher and student utterances for classroom discourse functions and language related to Can Do objectives. This observation data directly connects to the final Performance Test results, highlighting the relationship between input, interaction and output. The results were reported at the International Symposium of the CEFR-J 2025, where one of the invited speakers, Barry O'Sullivan, commented that the CEFR-J project was "one of the most successful implementation models based on a series of empirical studies and practices" (O'Sullivan 2025).

3.3 Impact of the CEFR: the case of English language teaching in Japan

The fact that the Courses of Study did not explicitly state CEFR levels suggests they are loosely, rather than strongly, aligned with the CEFR. It should be noted, however, that MEXT has decided to set attainment targets using the CEFR for their national surveys in the Fourth Basic Plan for the Promotion of Education: A1 or above for lower secondary school graduation and A2 or above for upper secondary school graduation (https://www.mext.go.jp/content/20240228-soseisk02-100000597_09.pdf).

The implementation of the CEFR principles, i.e., AoA, social agent and levels, in ELT in Japan can be examined by looking at two of the three components of the Comprehensive Learning System: delivery and assessment (British Council et al. 2022).[1]

3.3.1 Delivery: textbooks and local curricula

At the LSS level, an analysis of textbooks from a city with which our CEFR-J project group worked closely reveals a complex picture. The authorized textbooks, which follow the loosely aligned Courses of Study, show weak alignment with the CEFR's core principles such as AoA and social agent. They often contain decontextualized grammar exercises and lack authentic communicative tasks. However, many schools create their own supplementary, locally produced textbooks, which, in this specific city, show a much higher degree of alignment with the CEFR. This is mainly because the teachers involved hold beliefs that are consistent with the CEFR's emphasis on social agents and the AoA. These local materials effectively fill the gap left by the authorized textbooks, demonstrating a bottom-up, teacher-driven effort to align pedagogy with CEFR principles. However, it should be noted that this may not be the case for all the local boards of education.

Similarly, at the USS level, authorized textbooks generally exhibit weak alignment. The primary focus remains on grammar exercises and decontextualized practice, which contrasts sharply with the communicative, task-based approach advocated by the CEFR. There are some newly edited English textbooks with strong alignment with the CEFR or the CEFR-J, but their use so far seems to be limited to a small number of schools, and it will take time to familiarize teachers with CEFR-based textbooks with a strong emphasis on AoA.

1. The Comprehensive Learning System (CLS) is a conceptual model that argues for the success of language education through the close, systematic alignment of all core system components—specifically, curriculum, delivery (teaching), and assessment—rather than treating them as independent elements (British Council et al. 2022).

3.3.2. Assessment: entrance examinations and classroom tests

The assessment landscape for ELT in Japan is highly varied, ranging from classroom tests to high-stakes entrance examinations. This variety leads to a highly uneven degree of CEFR alignment.

For LSS students, two types of high-stakes entrance examinations are crucial for admission to USS: public and private. The public USS entrance examinations show some degree of CEFR alignment, with certain sections featuring task-based and contextualized items. This partial alignment reflects a gradual shift in public sector testing. However, private USS entrance examinations often show no such features, relying instead on traditional, decontextualized grammar and vocabulary items. This disparity creates a schism in the assessment landscape, as many students take both types of tests.

For USS students, the situation is even more complex. The Japanese university entrance examination system includes the Common Test for University Admissions (CTUA), individual university entrance examinations, and commercial tests. While classroom assessments and individual university entrance examinations often lack CEFR alignment, both the CTUA and commercial tests (e.g., EIKEN Tests, IELTS) claim to be aligned with the CEFR.[2] These tests feature a high proportion of task-based and contextualized items, effectively functioning as assessments of what social agents can do in real-life situations. This creates a disconnect: USS students are taught using weakly aligned materials, but they are assessed for university admission using strongly aligned tests. This forces students to adapt to a different approach to language use for high-stakes examinations, suggesting a positive backwash effect where the CEFR-aligned tests are driving changes in student learning strategies, if not in the curriculum itself.

This approach can be characterized as a laissez-faire strategy, whereby the provision of information regarding the CEFR and the CEFR-J, in addition to the tools and tests developed, is facilitated through our annual symposium. Subsequently, the decision regarding adoption or adaptation of these materials is entrusted to educators, publishers and test providers. This approach has resulted in varying degrees of alignment with the CEFR, and the impact is uneven.

4 The case of Japanese language education

In contrast to the gradual, cautious adoption in ELT, the implementation of the CEFR in Japanese language teaching (JLT) for foreign learners has been a much more direct, top-down approach. This was primarily driven by external pressure from institutions and students in Europe and beyond who required a standardized way to measure and report their Japanese language proficiency. They needed to demonstrate their language level for academic credit or for a visa and often required CEFR-aligned certificates.

In response to this demand, MEXT took a decisive step by publishing the *Frame of Reference for Japanese Language Education* (FRJ) (Subdivision on the Japanese Language of the Council for Cultural Affairs, 2021). This was followed by a regulatory move: MEXT established an accreditation system for Japanese language schools. To be accredited, these institutions must demonstrate a clear alignment with the FRJ in their curriculum design, teaching content and assessment. This mandatory requirement has created a powerful incentive for schools to adopt the CEFR.

An analysis of accredited Japanese language schools reveals a very high degree of alignment with the CEFR. The level of admission and the target level of achievement at the end of the programme are both defined according to the CEFR. The curriculum, delivery and assessment in these institutions are

2. References on the alignment of EIKEN with the CEFR: https://www.eiken.or.jp/eiken/en/grades/#:~:text=In%20a%20move%20to%20test,C1.
References on the alignment of IELTS with the CEFR: https://ielts.org/organisations/ielts-for-organisations/compare-ielts/ielts-and-the-cefr#:~:text=To%20help%20test%20users%20understand,and%20other%20Cambridge%20English%20Qualifications.

intentionally designed around the AoA and the concept of social agency. This structured, mandated approach has been highly effective in achieving a high degree of CEFR alignment within the accredited sector.

It is conceivable that a high level of alignment may have been achieved in accredited Japanese language schools. The impact has been significant, but not yet pervasive; the programmes of non-accredited schools do not align with the CEFR. It appears that these non-accredited schools harbour a desire to be accredited, yet lack the necessary understanding of how to achieve this. Consequently, the implementation of training programmes that are designed to facilitate familiarity and alignment with the CEFR is imperative.

5 Discussion: the power of the CEFR

The comparison between ELT and JLT in Japan is interesting because it illustrates two different pathways for CEFR implementation. In ELT, a laissez-faire approach has led to a varied, uneven degree of alignment. Delivery in schools often lags behind, while assessment, especially in high-stakes university admissions, is moving toward CEFR principles, such as AoA and SA as well as CEFR levels. This creates a tension that is driving change from the top, with tests leading the way. The CEFR, in this context, acts as a guiding framework that educators and test providers can voluntarily adopt, with adoption often being a response to a market demand for internationally recognized proficiency standards.

In JLT, the CEFR's adoption was top-down and mandated. This approach led to strong alignment in accredited institutions but has not yet reached the entire sector. The CEFR, in this case, functions as a powerful regulatory tool, with MEXT acting as a coordinating body, monitoring its use through the accreditation process. This contrasts with the CEFR's original intent, which explicitly states, "We have NOT set out to tell practitioners what to do" (CoE 2001: xi). Nevertheless, as the CEFR provides a comprehensive framework, it possesses an inherent power to influence and structure educational systems, particularly when adopted by an authoritative body.

The key point to note is that the CEFR's power is dependent on the approach taken. A laissez-faire approach can result in instances of innovation and alignment, driven by individual educators and market forces, but this can be a slow and uneven process. A top-down, mandated approach can achieve rapid and high-level alignment within a specific sector, but it requires significant institutional support and training to ensure widespread adoption. In both cases, the CEFR acts not as a static standard but as a catalyst for educational reform.

6 Conclusion

The CEFR's journey in Japan demonstrates its adaptability and influence as a framework for language education. Both ELT and JLT have engaged with its principles, but through different mechanisms. The ELT experience highlights the power of bottom-up initiatives and high-stakes testing to drive change, even in the absence of a strict mandate. The Japanese language education experience shows that a top-down, regulatory approach can achieve rapid and strong alignment.

Ultimately, the effectiveness of the CEFR in any national context depends on the synergy between curriculum design, delivery and assessment. If a system is not fully aligned with the CEFR, educators must actively work to fill the gap themselves, supplementing textbooks and adapting assessments to align with the framework's principles. This requires significant effort but is essential for creating a Comprehensive Learning System. The CEFR provides the conceptual tools for this work, and the experiences of Japan offer valuable lessons on how to deploy them, whether through cautious encouragement or decisive mandates.

7 References

British Council, EALTA, UKALTA & ALTE. 202. *Aligning language education with the CEFR: A handbook*. http://www.ealta.eu.org/documents/resources/CEFR%20alignment%20handbook.pdf (accessed 27 January 2026).

CEFR-J Version 1.0. 2012. https://www.cefr-j.org/index.html (accessed 27 January 2026).

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. 2005. *Reference level descriptions for national and regional languages (RLD): Draft guide for the production of RLD*. Version 2. Council of Europe.

Lenz, Peter & Günther Schneider. 2004. A bank of descriptors for self-assessment in European Language Portfolios. Strasbourg: Council of Europe. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168045b15f (accessed 27 January 2026).

Negishi, Masashi, Tomoko Takada & Yukio Tono. 2013. A progress report on the development of the CEFR-J. In Evelina D Galaczi & Cyril J Weir (eds.), *Exploring language frameworks. Proceedings of the ALTE Krakow conference, July 2011*, 135-163. *Studies in language testing* 36. Cambridge: Cambridge University Press.

O'Sullivan, Barry. 2025. Reflections on the local and global significance of the CEFR-J. Invited talk at the CEFR-J 2025 International Symposium, Kyoto University, 26 March.

Subdivision on the Japanese Language of the Council for Cultural Affairs. 2021. https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/hokoku/pdf/93736901_01.pdf (in Japanese) (accessed 27 January 2026).

Tono, Yukio. 2013. Criterial feature extraction using parallel corpora and machine learning. In Ana Díaz-Negrillo, Nicolas Ballier & Paul Thompson (eds.), *Automatic treatment and analysis of learner corpus data*, 169-204. Amsterdam: John Benjamins.

8 Biographies

Masashi Negishi is a professor emeritus at Tokyo University of Foreign Studies, Japan. He earned his PhD from the University of Reading, UK. He has played a leading role in numerous English Language Teaching (ELT) research projects in Japan. His contributions include heading the CEFR-J Project and participating in the development of English proficiency tests and national education surveys. His current research interests are focused on the application of the *Common European Framework of Reference for Languages* (CEFR) to language teaching in Japan and the development of CEFR-J based tests.

Yukio Tono is a professor of corpus linguistics at Tokyo University of Foreign Studies and serves as Director of the World Language Center at TUFS. He received his PhD from the University of Lancaster, UK. His research focuses on corpus applications in foreign language teaching and learning, L2 vocabulary acquisition, pedagogical lexicography and resource development for the CEFR-J. He serves as principal investigator of the CEFR-J project together with Masashi Negishi.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Implementation and impact of the CEFR in Costa Rica's foreign language education system

Ana C. González-Ramírez, University of Costa Rica, Costa Rica
Walter Araya-Garita, University of Costa Rica, Costa Rica

https://doi.org/10.37546/JALTSIG.CEFR8-11

This article explores the implementation of the Common European Framework of Reference for Languages (CEFR; Council of Europe [CoE] 2001) and its impact on Costa Rica's syllabus reform of the foreign language education system. Since its official adoption by MEP (Ministry of Public Education of Costa Rica) in 2016, the CEFR has guided curriculum reforms, shifting from traditional content-based teaching—focused on grammar and vocabulary—to a student-centered, communicative approach. Findings highlight improvements in students' proficiency levels and teacher professional training, though challenges remain, such as unequal resource distribution and lesson-time constraints. The article underscores the need for systematic teacher training, ongoing policy adjustments, and implementation in higher education. Recommendations include increasing lesson time and data-driven resource allocation to enhance language education reach, particularly in underserved regions.

Keywords: CEFR, Costa Rica, English language learning, bilingual education, language testing

1 Introduction

Before 2016, language education in Costa Rica—particularly in English—was shaped mostly by traditional pedagogies. Despite national coverage, Costa Rica lacked a unified framework to assess language proficiency systematically. As a result, significant disparities developed in instructional quality and assessment criteria across educational regions.

To address these challenges, the Ministry of Public Education (MEP) adopted the descriptors and guidelines included in the 2001 edition of CEFR in 2016. The use of its comprehensive, internationally recognized scale to describe language proficiency from A1 (beginner) to C2 (mastery) (CoE 2001) aimed to modernize Costa Rica's language education system by aligning curricula, assessment, and instructional practices with global needs. This alignment required teachers to adapt classes and create or modify class materials to achieve the specific communicative outcomes of each proficiency level. Instruction had to reflect the increasing complexity of linguistic content and real-world tasks expected at each stage. Teachers were encouraged to move beyond traditional grammar-based instruction and integrate performance-based activities that foster authentic language use. This shift demanded ongoing reflection, adaptation, and professional development to ensure consistency between teaching, assessment, and the CEFR descriptors.

The implementation of the CEFR in Costa Rica has unfolded through a gradual and structured process that reflects the country's long-term commitment to sustainable educational reform. Beginning with the institutional integration of CEFR principles between 2016 and 2018, efforts focused on engaging key stakeholders, designing CEFR-based lesson plans, and promoting communicative, task-oriented assessment practices. During the following phase (2019-2020), focus shifted to consolidating these reforms through quality assurance, supervision of use, and teacher development initiatives. Since 2021, the process has entered a third stage, one of refinement and full implementation, characterized by the widespread adoption of CEFR-aligned practices across all educational levels and the integration of proficiency-based assessment and certification systems.

This article traces the process of CEFR implementation in Costa Rica and analyses its impact on curriculum design, pedagogy, teacher training and assessment. Drawing on institutional documentation, national test results and recent evaluations, the authors critically examine both the successes and ongoing challenges of this reform process. They conclude by offering policy recommendations and future directions.

2 Pre-CEFR language education practices

Costa Rica's engagement with English language education dates to 1825, when the Grammar-Translation Method dominated instruction, reinforcing the notion of language as written content to be memorized and recited (Córdoba-Cubillo et al. 2005). Toward the end of that century, some schools began to adopt the Direct Method, which encouraged oral interaction and more naturalistic learning (Rojas-Díaz 2021). However, because many instructors were foreign teachers or English-speaking Costa Ricans who lacked a pedagogical background, its implementation was uneven and often ineffective. In the 20th century, efforts to professionalize English teaching gradually advanced, particularly with the establishment of teacher-training programmes at the University of Costa Rica (UCR) in the 1950s (Marín-Arroyo 2013). Although this milestone marked a significant step forward, the Audiolingual Method promoted at the time still fell short of developing the communicative competence later emphasized by international frameworks. By the 1990s, Costa Rica began to experiment with Communicative Language Teaching; however, institutional inertia, unequal resource distribution (Córdoba-Cubillo et al. 2005), and washback to traditional national evaluations limited its widespread adoption.

Assessment practices mirrored the instructional emphasis on form over function. The former national assessment tool that MEP would design to determine students' proficiency, called *Bachillerato Exam*, focused on reading comprehension but lacked guiding proficiency scales or communicative benchmarks (Marín-Arroyo 2013). Because teachers focused on preparing their students for the test, other formative activities were neglected: students were not provided with much listening or speaking practice in class, memorization of vocabulary and grammatical structures was encouraged, and English classes were given in Spanish to "ease" students' learning process.

Efforts to depart from these text-based practices started to take place in the early 2000s. Early initiatives of the Office of Foreign Languages centered on improving listening and speaking abilities (MEP 2021a). However, the lack of a national language framework—such as the CEFR—meant that teaching goals, classroom and assessment practices, and student proficiency expectations varied significantly across schools and regions.

A 2015 diagnostic report revealed that teachers found the English syllabus difficult to interpret and implement, overly subjective, and lacking alignment with the CEFR (MEP 2015). Its content was considered outdated, of limited relevance to learners, and mostly ineffective as it did not adequately address the communicative competencies required in contemporary contexts. More importantly, proficiency outcomes remained below expectations despite twelve years of instruction largely due to the absence of clear exit benchmarks and misalignment between curriculum, assessment, and classroom practices.

In sum, the pre-CEFR era showed a system that was fragmented and in need of alignment, paving the way for the CEFR's adoption as a unifying framework.

3 Adoption and familiarization with CEFR (2016–2018)

3.1 Educational system

In 2016, MEP spearheaded a national curriculum overhaul that resulted in the official adoption of the CEFR as the guiding benchmark framework for its foreign language education programmes. Although no further updates have been made to the English programmes since then, meaning that the CEFR *Companion Volume* (CEFR/CV; CoE 2020) has not yet been incorporated into national policy or official curricular frameworks, the use of CEFR as the guide to setting exit proficiency levels brought about many learning opportunities and changes.

The new 2017 English syllabus provided the rationale for the adoption of CEFR's 'Can Do' descriptors as the programmes' learning objectives. According to MEP, the CEFR was the right choice as it "provides a common basis for the development of language syllabi, curriculum guidelines, textbooks, and assessment, describes what language learners do at different levels of proficiency within domains and scenarios, using self-affirmative language, [and] provides a common terminology that can be adapted for all languages and educational contexts" (MEP 2017: 27). MEP now aimed to emphasize the social nature of communication, and CEFR descriptors were to guide such a transformation.

The traditional text-based, teacher-centered system had to change toward a student-centered, task-based learning one that mirrored CEFR's competence work more closely. The previously emphasized Communicative Approach progressively shifted to the now-favored Action-oriented Approach (AoA). Official MEP documents detailed the new roles of teachers and learners, provided examples of teaching learning strategies, as well as sample mediating strategies and activities, and all within the framework of CEFR.

The implementation of the new curriculum was to be made progressively. The long-term goal was that, by starting in 2017 with seventh graders, results could be seen by 2021. Figure 1 shows the assignment of CEFR bands to each level in primary and secondary education in Costa Rica (MEP 2016a). The high school system would graduate students operating at an exit level of B1.2 (ready for B2 proficiency instruction). As can be observed in Table 1, 'new' bands were created (A1.1 and A1.2, for example) and assigned to each grade. It should be noted that grades 7-9 receive three weekly lessons of 40 minutes each, while grades 10 and 11 receive 5 weekly lessons of 40 minutes each. CEFR proficiency levels were redistributed to meet the greatest number of band descriptors across grades based on the number of instructional hours assigned. In bilingual high schools, for example, English instruction in grades 10 and 11 has on average five weekly lessons of 40 minutes for 41 weeks, a total of 136 hours. As Table 1 shows, each level has its own CEFR benchmarks, which have also been distributed accordingly.

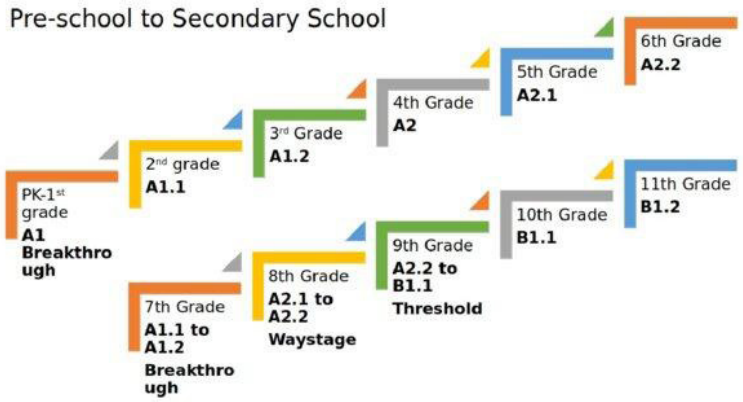

Figure 1. Distribution of CEFR proficiency bands as exit profiles across grades in primary and secondary education in Costa Rica

Table 1. *Proposed plan of distribution of CEFR Benchmarks according to the number of instructional hours of English and expected proficiency level to achieve in regular public high schools*

Grade level	Yearly Teaching Hours	CEFR Guidelines / Benchmarks	Exiting proficiency level
7	82	A1 90-100	A1-A2.1
8	82	A2 180-200	A2.2
9	82		B1.1
10	136	B1 350-400	B1.1
11	136		B1.2

Source: National English advisor, M. Granados, personal communication, 27 October 2025.

3.2 Teachers' professional profile

Adoption and familiarization efforts with the CEFR also required teachers to further professionalize their language proficiency, pedagogy and assessment practices. Therefore, early in the process, the MEP launched a series of nationwide workshops, conferences and seminars to help improve teacher competencies.

Some teachers welcomed the initiatives as intellectually stimulating and professionally enriching. Others, however, voiced concerns about unrealistic expectations—particularly the demand to deliver CEFR-based outcomes in contexts where infrastructure, resources and digital access remain limited. Such concerns were also founded on the new societal expectations on teachers' professional identity: with the new syllabus their role is considered pivotal in achieving national bilingualism goals, and they are thus held accountable through standardized testing of students (e.g., the PDL-MEP).

In addition, teachers were now required to have a B2-C1 English level and a sound understanding of the theoretical foundations of the new English programmes (Programa Estado de la Nación 2017). To address these needs, the results obtained from the CEFR-aligned proficiency tests given to teachers in 2015 were used "to design and implement training courses to improve [their] communicative language ability and teaching practices" (MEP 2017: 9). For example, the programme *¡Actualizándonos!* is a continuous professional development initiative born in 2016 that encompasses training in both linguistic proficiency and pedagogical mediation (MEP 2016b).

CEFR-focused training opportunities were also provided. The first training workshops helped teachers learn about the CEFR, understand its objectives, and comprehend that its adoption would transform their English teaching practice from structural grammar-based instruction to meaningful, communicative, and task-based learning that was more in accordance with current global needs and profiles. The second wave of training initiatives helped teachers explore how AoA and CEFR could be incorporated into their lesson plans, classroom activities and assessments to effectively sequence lessons across a continuum of proficiency. Therefore, national workshops as well as certification and in-service training programmes further familiarized educators with key concepts such as 'Can Do' statements, learning progression scales, and the AoA principles, as outlined in the CEFR/CV (CoE 2020). Finally, a third type of training aimed to teach instructors how to use CEFR-aligned assessment elements such as integrated skill tasks, contextualized rubrics, and performance-based judgments, which reinforced the CEFR's holistic vision of language use. Assessment practices evolved toward criterion-referenced, transparent and descriptive scales that promote coherence between classroom activities and evaluation practices through the use of performance indicators that were specifically distributed and even expanded for the Costa Rican context.

Governmental authorities shared support materials that would help teachers plan their lessons with the new AoA and CEFR-based learning outcomes in mind. For example, the *Resource Kit for Seventh Grade LEBS and Bilingual Groups* provided teachers with additional ideas and resources to strengthen

and practice students' oral and written production. "The intention of this kit is to make the syllabus can-dos accessible through a set of tasks for each competence" (MEP 2018a: 4). Thus, this material blended AoA principles and CEFR band descriptors and poured this understanding into lesson plan templates, sample lesson plans, and classwork activities.

3.3 Assessment practices

The introduction of new national standards and CEFR-aligned learning outcomes also guided assessment practices across public schools (Alianza para el Bilingüismo 2018). Official documents highlighted the need for objective oral assessments grounded in CEFR descriptors (MEP 2018c). Likewise, initiatives such as the INCO programme—conversational English workshops for secondary students—were guided by methodological documents that aligned learning indicators with CEFR levels for curriculum and assessment design (MEP 2018).

The adoption of the CEFR encouraged the construction of assessment processes that incorporate, as Bachman (2009) notes, deep expertise across linguistics, pedagogy, and psychometrics. Such assessments should also prove effective through the collection of validity evidence (Bachman and Palmer 1996).

Consequently, a revamp of the assessment processes under this light became a priority towards the collection of "empirically based information to both close learning gaps and generate evidence of learning successes" (MEP 2017: 10).

3.4 Prueba de Dominio Lingüístico (PDL-MEP)

One key development in response to the new approach was the English Diagnostic Test (PDL-MEP), Costa Rica's large-scale English proficiency test. This test measures students' communicative abilities and evaluates the effectiveness of the educational system in achieving CEFR-based goals, ensuring institutional accountability for language outcomes. Its design aligned internationally recognized descriptors with national curriculum goals and local delivery conditions (Quesada-Pacheco and Araya-Garita in press; Dimova et al. 2020) by adapting to technological inequalities across regions (Quesada-Pacheco and Araya-Garita in press), engaging stakeholders (Quesada-Pacheco et al. 2023), and addressing the limitations inherited from the discontinued Bachillerato Exam.

The implementation of the PDL-MEP results from a systematic, multi-stakeholder engagement process between the UCR programme PELEx (Programa de Evaluación en Lengua Extranjera) and the MEP. The active participation of specific actors—national and regional advisors, school directors, IT staff and pedagogical coordinators—ensures clarity, transparency and effective communication. Engagement begins with inter-institutional planning meetings and continues through the joint design of sampling frameworks, material distribution, mock testing and technical training. This sustained collaboration culminates in the delivery of official testing, reporting and certification, exemplifying an inclusive, coordinated and accountability-driven approach to large-scale language assessment.

The implementation of the PDL-MEP not only enabled the comparability and standardization of data but also positioned CEFR-aligned testing as a pivotal mechanism for driving systemic educational improvements through evidence-based decision-making. Decisions about school quality, equity and curricular reform must be data-informed and pay rigorous attention to fairness, validity and transparency (AERA et al. 2014; ILTA 2024), which resonates with McNamara's (2000) view of language assessment as an activity that entails ethical and social consequences.

In response, the PDL-MEP adopts O'Sullivan's (2021) Comprehensive Learning System, which promotes the integration of curriculum, instruction and assessment. In this model, CEFR descriptors serve not merely as scoring references but as the foundation for teaching objectives, classroom practices and feedback processes, including new materials, assessment tools, training workshops and CEFR literacy

initiatives. The PDL-MEP applies this approach by delivering results linked to CEFR bands, offering diagnostic feedback and promoting curricular alignment.

4 Consolidation and evaluation of CEFR use (2019-2020)

Foreign language instruction based on the CEFR standard scale for proficiency has encouraged teachers to create student-centered environments that foster autonomy and active learning (Quesada-Pacheco and Araya-Garita in press). Clearer learning goals that allow monitoring of student progress and tailor instruction to individual needs, if necessary, now guide lesson planning. Classroom activities encourage more realistic use of language in the classroom through authentic tasks. As a result, learner autonomy and engagement have become paramount, in alignment with CEFR's learner-centered philosophy.

Assessment practices have also evolved in accordance with CEFR standards. The use of meaningful, contextualized language tasks that integrate all skills, address communicative competences and employ performance-based rubrics has resulted in more useful measures of student proficiency in the personal and academic domains. This composite of features results in assessment tools with enhanced construct validity that can guide instructional decisions and support learning (Quesada-Pacheco et al. 2023).

Empirical evidence confirms the positive impact of CEFR implementation on English language learning in Costa Rica. Figure 2 compares the percentage distribution of English listening proficiency levels (CEFR bands) among Costa Rican secondary school students as assessed by two tests. The 2008 test only assessed listening; it was created entirely by MEP, and according to it (MEP 2021a), it followed the CEFR's framework of language assessment. The 2019 test was the PDL-MEP, which was created by the UCR and was aligned to the CEFR. As observed, in 2008 students in the A1 band (64.7%) were predominant, meaning most students were basic language users with limited communicative ability; whereas a much smaller number reached A2 (18.7%), B1 (8.0%), B2 (5.5%), and C1 (2.3%). In contrast, the 2019 results reveal a clear upward shift in proficiency compared to those from 2008, with more students reaching A2 and B1 levels, although the majority remained below the independent user threshold (B2).

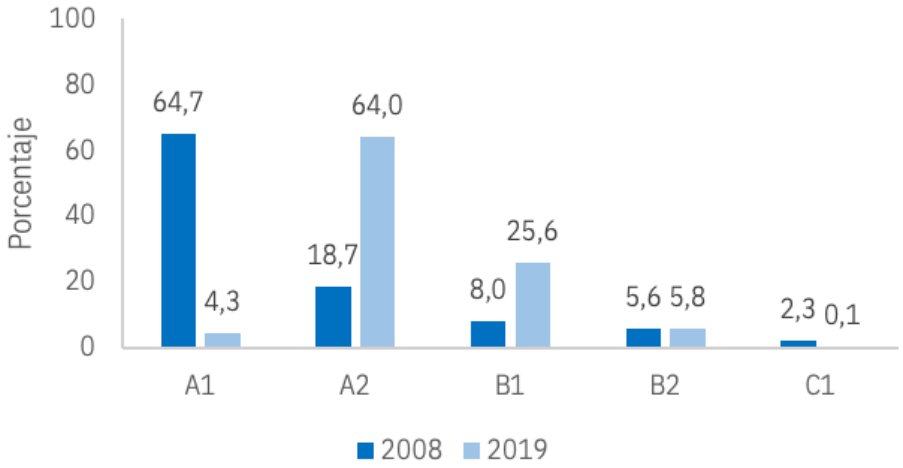

Figure 2. Percentage distributions of the results obtained by high school students in the English Proficiency Tests of listening, 2008 vs. 2019. Reprinted from the *Educational Policy for Language Promotion* (MEP 2021a)

The contrast between the 2019 and 2021 general performance results also confirms progress. As Figure 3 shows, in 2021 although the largest group of students was still in the A2 band, there were more students who moved to more proficient bands—B2 and C1—showing an improvement from the 2019 results. The number of students in A2 dropped sharply to 57%. There was also a notable increase in B2

(12%) and a slight rise in C1 (2%). Data from the national proficiency test administered in 2019 and again in 2021, after the adoption of CEFR, reveals significant improvements in student performance. Based on official information obtained from MEP's 2008 listening test, the proportion of students attaining A2 and B1 proficiency has increased notably since then, indicating that systemic reforms grounded in CEFR principles are yielding measurable progress.

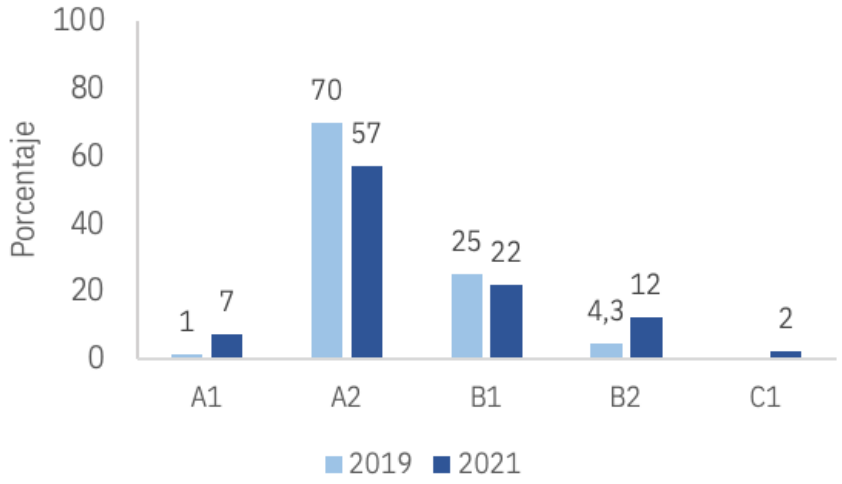

Figure 3. Percentage distributions of the results obtained by high school students in the English Proficiency Tests for four skills, 2019 vs 2021.[1] Reprinted with permission Quesada-Pacheco et al. (2023)

5 Refinement and full implementation (2021-present)

The adoption of the CEFR has acted as a catalyst for a profound transformation of English language education in the national public system. Classroom practices shifted toward a more communicative, learner-centered model rooted in authentic language use (Quesada-Pacheco et al. 2023) that is constantly and consistently improving. The creation of bilingual classrooms[2] and the full adoption of CEFR-aligned curricula have accelerated reform. Regional workshops such as *English on the Spot* (MEP 2024) and *English for Conversation* (MEP 2018b) now provide secondary students with additional practice to improve their oral skills.

Furthermore, English lesson coverage in preschool and primary education has been increased (MIDEPLAN 2018). Now, the entire country's foreign language education system in English is aligned to CEFR.

Teacher professional development has become a strategic priority, with tenure now contingent on achieving advanced proficiency (C1) and with sustained training opportunities supported by inter-institutional alliances. National advisory committees have devoted much of their efforts toward the creation of additional resources that refer to the CEFR/CV to support teachers' classroom practices (MEP 2023) and language assessment (MEP 2021b).

The definition of exit profiles for secondary education—B2 for bilingual and technical schools, B1 for other modalities—has further standardized learning outcomes and learning achievement. Systematic, CEFR-based assessment mechanisms now inform policy decisions, ensuring that diagnostic data directly shape targeted training and curriculum design, in accordance with McNamara (2000).

1. These findings are based on students who finished high school under the previous English programme, in operation until 2019, before the first graduating class of the new English programme, implemented in 2017, finished in 2021.
2. Bilingual classrooms are in high schools that provide 14 English lessons per week: 5 listening/speaking, 5 reading/writing, and 4 literature. There are eight such schools in the country.

Institutional cooperation, notably between MEP and PELEx, has strengthened assessment systems and addressed logistical and pedagogical challenges. Such collaboration resulted in a tool that effectively aligned curriculum, teaching, learning and assessment. Although the PDL's administration was not nationwide in the last two years, the current national English proficiency test (created and administered since 2024 by MEP) also responds to the new CEFR-based approach (MEP 2017).

6 Policy evolution and future directions

Costa Rica's implementation of the CEFR remains an evolving interpretive process. As positive results have begun to appear, attention is shifting toward consolidating the policy strategies they stem from and replicating them. The next phase of the reform emphasizes more effective responsiveness to classroom realities, data trends, and long-term national goals for bilingualism and educational equity.

One key future direction involves systematic curriculum revisions informed by both teacher feedback and student learning data. The insights gained by practitioners will play a critical role in ensuring that curriculum materials are pedagogically sound, contextually relevant and technologically appropriate. In fact, the MEP is now innovating with online placement tests, interactive self-assessment tools and task-based learning activities mapped to CEFR levels to enhance CEFR-aligned teaching and assessment (Araya-Garita and González-Ramírez 2024).

The vast amount of data that can be obtained through these digital means should continue to be used to strengthen curriculum implementation and classroom practice by welcoming CEFR experts' advice, holding training workshops by academia, and providing close advisory sessions. Integrating classroom-based evidence with national performance data allows policymakers to better align instructional content with realistic proficiency outcomes (Araya-Garita and González-Ramírez 2024).

Costa Rica's long-term bilingualism goals have driven the adaptation of the CEFR across educational levels, with ongoing efforts to extend CEFR-based instruction into higher education and vocational training to ensure continuity in language development. Concurrently, policies to enhance teacher qualifications, redistribute resources, and address local instructional needs reflect a commitment to data-driven policymaking aimed at raising standards and promoting equitable progress among all learners.

7 Challenges and limitations

While teacher training represents one of the most notable achievements of implementing a CEFR-based curriculum, the underlying professional development model has its limitations. Teachers in rural areas, non-traditional learning centres and underfunded schools often lack access to ongoing, context-sensitive professional development and may be ill-equipped to fully adopt communicative methodologies or interpret CEFR-aligned assessment results (Araya-Garita 2021). Many of these educators reported difficulty in translating CEFR theory into classroom practice as they believe they lack mentorship, enough reflection time, or clear curricular adaptation guidelines and models. Teachers have also highlighted the disconnect between the programme's theoretical demands and the realities of their instructional settings (Programa Estado de la Nación 2023).

Many public schools still face challenges such as large class sizes, limited instructional time, and wide proficiency differences among students, hindering teachers' capacity to assess language proficiency effectively. As a result, assessment often remains focused on the four skills rather than the four modes of language use outlined by the CEFR. This results in a continued overemphasis on grammar, vocabulary, and reading comprehension, while speaking, listening, and writing are often neglected (Quesada-Pacheco et al. 2023; Fallas-Godínez and Araya-Garita 2024). According to Araya-Garita and González-Ramírez (2024), the unequal allocation of instructional time across different school types stands as one of the most pressing concerns.

These disparities are further compounded by rising expectations linked to national bilingualism goals and high-stakes assessments such as the PDL-MEP test, placing additional stress on educators, many of whom are required to meet CEFR-based outcomes in conditions far from ideal (Elizondo-González and Araya-Garita 2026).

Structural challenges also hinder equitable implementation of CEFR-based instruction. Limited instructional time, especially at the primary level, reduces the opportunity for students to build strong linguistic foundations early on. Persistent differences in access to quality instruction between urban and rural areas, as well as in boys' and girls' speaking performance, threaten to widen achievement gaps.

Female students consistently underperformed in oral interaction tasks, which suggests gendered patterns of engagement or confidence in communicative settings within the classroom. Similarly, students in rural and marginalized areas lag behind those in urban centres, largely due to unequal access to qualified teachers, digital tools, and adequate resources. These inequities raise critical concerns regarding the consequential validity of the CEFR-aligned reform (Messick 1995).

8 Conclusions and recommendations

The adoption of the CEFR has significantly transformed Costa Rica's national language education system, aligning it more closely with international standards and best practices (Quesada-Pacheco et al. 2023). One of the most important outcomes has been a fundamental pedagogical shift toward student-centered environments that emphasize meaningful communication about familiar topics in more authentic scenarios. Curriculum, instruction and assessment are now more coherently aligned across grade levels, enabling teachers to design their classes with clearer objectives in mind and monitor learner progress using common reference points.

To consolidate the gains obtained so far and ensure sustained progress, several recommendations are provided. First, English instruction time should be increased, particularly in early grades, to build foundational skills and support long-term language acquisition. Second, targeted investment must prioritize rural and under-resourced schools to reduce inequities in teacher preparation and access to learning materials. Third, future teacher training programmes should emphasize context-responsive models that promote classroom-based inquiry and encourage peer collaboration. Fourth, gender and regional disparities in performance should be systematically monitored and addressed through data-informed interventions. Lastly, national CEFR-aligned proficiency targets should be clearly articulated with structured learning pathways that enable more students to reach B2 or higher levels (Araya-Garita and González-Ramírez 2024).

A shift toward more sustained, inclusive, and context-responsive teacher development is also needed to ensure the long-term success of the CEFR in Costa Rica. Teachers require long-term accompaniment to fully understand, internalize and adapt CEFR descriptors to their local realities. Online CEFR resource hubs can also provide language instructors with more equitable access to context-tailored teaching materials, sample lesson plans created by actual teachers, and assessment tools. Peer-coaching and teacher learning communities can facilitate co-construction of CEFR-aligned practices. Collaborative spaces would allow teachers to share strategies for adapting CEFR descriptors to real classroom conditions and to contextualize CEFR progression scales based on student performance (Araya-Garita 2021).

Ultimately, teacher professional development is the cornerstone of effective CEFR implementation, enabling educators to make informed, autonomous, and equitable pedagogical decisions. Since the CEFR is a flexible framework rather than a prescriptive syllabus, its successful adoption depends on teachers' ability to interpret and adapt its descriptors to local contexts—designing learning pathways, identifying learner needs, and assessing language use in authentic and fair ways.

To ensure fair and meaningful learning outcomes for all students, Costa Rica must continue to prioritize integrated approaches that combine curriculum design, pedagogical practice, and assessment—anchored in the principles of the CEFR (Quesada-Pacheco et al. 2023).

Strengthening digital infrastructure and access to it can help promote more equitable access to standardized assessment and mitigate the digital divide, particularly in remote regions. Expanding teacher training in language assessment and digital literacy is vital to help educators interpret test data and align instruction effectively (Araya-Garita 2021).

In Costa Rica, the adoption of CEFR has resulted in major advancements in the light of the objectives it was meant to help achieve: to modernize instruction, reduce regional disparities, and enable comparability of outcomes across educational institutions (Alianza para gel Bilingüismo 2018; Quesada-Pacheco et al. 2023). With continued political will, stakeholder engagement, and inclusive policy design, Costa Rica can continue to build a more robust, equitable, and internationally competitive language education system.

9 References

Alianza para el Bilingüismo. 2018. *Estrategia Nacional de Bilingüismo 2018-2022.* San José: Ministerio de Educación Pública de Costa Rica.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. 2014. *Standards for educational and psychological testing.* New York: American Educational Research Association.

Araya-Garita, Walter. 2021. Dominio lingüístico en inglés en estudiantes de secundaria para el año 2019 en Costa Rica. *Revista de Lenguas Modernas* 34. 1-21.

Araya-Garita, Walter & Ana González-Ramírez. 2024. Impact and implementation of the CEFR framework in Costa Rica's language education system. Poster presented at the International Conference "Responding to the CEFR Alignment Handbook", Blanquerna - Universitat Ramon Llull, Barcelona, 18-19 October.

Bachman, Lyle. 2009. *Fundamental considerations in language testing*, 2nd edn. Oxford: Oxford University Press.

Bachman, Lyle & Adrian Palmer. 1996. *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Córdoba-Cubillo, Patricia, Rossina Coto-Keith & Marlene Ramírez-Salas. 2005. La enseñanza del inglés en Costa Rica y la destreza auditiva en el aula desde una perspectiva histórica. *Actualidades Investigativas en Educación* 5(2). https://doi.org/10.15517/aie.v5i2.9153.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. https://www.coe.int/en/web/common-european-framework-reference-languages (accessed 8 April 2024).

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

Dimova, Slobodanka, Køhler Simonsen & Jane Kling (eds.). 2020. *Local language testing: Design, implementation, and development.* Berlin: Springer.

Elizondo-González, Fabian & Walter Araya. 2026. The Geography of Language Learning: How Region and School Type Shape English Proficiency in Costa Rica. *Revista Electrónica de Investigación Educativa.* https://doi.org/10.15517/pj482p23.

Fallas-Godínez, Alejandro & Walter Araya-Garita. 2024. Computer or paper-based delivery mode: An analysis for testing English reading strategies. *Revista de Lenguas Modernas* 40. https://doi.org/10.15517/r1m8dy28.

International Language Testing Association. 2024. *Code of ethics and best practices.* https://www.iltaonline.com (accessed 15 March 2024).

Marín-Arroyo, Edwin. 2013. *La enseñanza del inglés en Costa Rica en el siglo XIX: Una respuesta al modelo económico.* Cartago: Instituto Tecnológico de Costa Rica.

McNamara, Tim. 2000. *Language testing.* Oxford: Oxford University Press.

Messick, Samuel. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50(9). 741-749. https://doi.org/10.1037/0003-066X.50.9.741.

Ministerio de Educación Pública. 2015. *Informe de diagnóstico: Programas de estudio de inglés para educación preescolar, primaria y secundaria.* San José: Ministerio de Educación Pública.

Ministerio de Educación Pública. 2016a. *Programa de Estudio de Inglés: Primer Ciclo de la Educación General Básica.* San José: Ministerio de Educación Pública.

Ministerio de Educación Pública. 2016b. *Plan Nacional de Formación Permanente: ¡Actualizándonos!* San José: Instituto de Desarrollo Profesional Uladislao Gámez Solano. https://idp.mep.go.cr/sites/all/files/idp_mep_go_cr/publicaciones/actualizandonos_version_final_3.pdf (accessed 20 October 2025).

Ministerio de Educación Pública. 2017. *Programa de Estudio de Inglés Plan de Estudios Liceo Bilingüe Secciones Bilingües Español-Inglés Tercer Ciclo.* San José: Ministerio de Educación Pública.

Ministerio de Educación Pública. 2018a. *Resource Kit for Seventh Grade LEBS and Bilingual Groups.* https://recursos.mep.go.cr/lebs_y_sebi/data/documents/resourcekit.pdf (accessed 20 October 2025).

Ministerio de Educación Pública. 2018b. *DTCEd_344_02_2018_INCO: Documento INCO.* https://recursos.mep.go.cr/2023/sitios-ingles/setimo/sitio_7/inco/dtced_344_02_2018_inco.pdf (accessed 15 October 2025).

Ministerio de Educación Pública. 2018c. *La Prueba Oral para medir la comprensión y producción oral en lenguas extranjeras.* https://drea.mep.go.cr/sites/default/files/publicaciones-anexos-2022/PRUEBA%20ORAL%2014%20nov%202017_0.pdf (accessed 20 October 2025).

Ministerio de Educación Pública. 2021a. *Política Educativa de Promoción de Idiomas: Hacia una Costa Rica bilingüe.* https://ddc.mep.go.cr/sites/all/files/ddc_mep_go_cr/archivos/politica_educativa_para_la_promocion_de_idiomas.pdf (accessed 20 March 2024).

Ministerio de Educación Pública. 2021b. *Guidelines on How to Create Indicators of Learning for the Suggested Pedagogical Mediation of the English Curriculum.* https://recursos.mep.go.cr/lebs_y_sebi/data/documents/bilingue/7%20Orientacionesbilingue%20v.f.pdf (accessed 10 October 2025).

Ministerio de Educación Pública. 2023. *Orientaciones: La interacción y la producción oral en continuo.* https://recursos.mep.go.cr/lebs_y_sebi/data/ORIENTACIONES.pdf (accessed 24 October 2025).

Ministerio de Educación Pública. 2024. *Lineamientos Taller English on the Spot (EOS).* https://drea.mep.go.cr/sites/default/files/publicaciones-anexos-2025/Lineamientos%20Taller%20English%20on%20the%20Spot%20%20EOS%20VF%2025-11-2024.pdf (accessed 24 October 2025).

Ministerio de Planificación Nacional y Política Económica [MIDEPLAN]. 2018. *Plan Nacional de Desarrollo y de Inversión Pública del Bicentenario 2019-2022*. San José: Gobierno de Costa Rica.

O'Sullivan, Barry. 2021. The comprehensive learning system: Designing and implementing coherent language education. In Barry O'Sullivan & Dina Tsagari (eds.), *The Routledge handbook of second language assessment*, 11-23. London: Routledge.

Programa Estado de la Nación. 2017. *Sexto informe Estado de la educación*. San José: Consejo Nacional de Rectores. https://estadonacion.or.cr/?informes=sexto-informe-estado-de-la-educacion-2017 (accessed 29 October 2025).

Programa Estado de la Nación. 2023. *Noveno informe Estado de la educación*. San José: Consejo Nacional de Rectores. https://estadonacion.or.cr/?informes=informe-estado-de-la-educacion-2023 (accessed 29 October 2025).

Quesada-Pacheco, Allen, Walter Araya-Garita & José Fallas-Godínez. 2023. La enseñanza y aprendizaje del inglés en la secundaria pública costarricense del siglo XXI: Innovaciones, brechas y desafíos. *CONARE Repository*. https://repositorio.conare.ac.cr/handle/20.500.12337/8517 (accessed 20 March 2025).

Quesada-Pacheco, Allen & Walter Araya-Garita. In press. Large-scale online English language testing in Costa Rica: The pioneering PELEx experience. Berlin: Springer.

Rojas-Díaz, Katherine. 2021, April 20. Costa Rica fortalece su ruta hacia el bilingüismo con Política Educativa de Promoción de Idiomas. Ministerio de Educación Pública (MEP).

10 Biographies

Ana C. González-Ramirez, MA in Teaching English as a Foreign Language, is a faculty member at the University of Costa Rica, where she has taught for over sixteen years Her research focuses on language assessment and the localization of testing practices to enhance the quality and fairness of English education in Costa Rica. As an active member of the PELEx program, she has contributed to studies on test constructs, regional score variation, and CEFR alignment. Her professional interests also include ESP course and curriculum design, teacher development, and educational outreach.

Walter Araya-Garita is a full professor and researcher at the University of Costa Rica (UCR) specializing in language testing and assessment, with over 25 years of experience. He is the founder of the Language Assessment and Training Program (PELEx) at UCR and has played a key role in advancing language evaluation initiatives nationally and regionally. Walter has served as secretary of the Latin American Association for Language Testing and Assessment (LAALTA) for two years. He holds an MA in Teaching English as a Foreign Language and an MSc in Administration from the University of Costa Rica, as well as a specialization in Educational Planning from the National Institute of Educational Planning and Administration (NIEPA) in India.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Some concluding reflections

David Little, Trinity College Dublin, Ireland
Neus Figueras, University of Barcelona, Spain
Lynda Taylor, University of Bedfordshire, Great Britain

https://doi.org/10.37546/JALTSIG.CEFR8-12

From the beginning, the editorial group that prepared *Aligning language education with the CEFR: A handbook* (British Council et al. 2022) anticipated that the experience of use would quickly reveal a need for revision and further elaboration. That was one of the chief motivations for organizing the 2024 Barcelona conference. As we explained in our introduction, the articles in this special issue began life as presentations at that conference, and some of them report on alignment projects that made use of the Handbook. On the basis of the feedback, both explicit and implied, provided by our authors, it seems to us that the following issues need to be addressed when planning CEFR alignment projects and preparing a revised edition of the Handbook.

Online vs. in-person alignment procedures. All the alignment projects reported in this special issue were conducted at least partially online, whereas the Handbook is mostly focused on in-person procedures. In most cases, reasons of cost and convenience are likely to tip the balance in favour of working online, for which a revised edition of the Handbook should offer more detailed guidance, taking account of key differences between the two modalities (Kanistra 2025; Kollias 2023).

Four modes of language use vs. four skills. Even when they have been newly designed using the CEFR/CV's mediation scales, the tests reported in these articles are described in terms of the four skills of traditional second language pedagogy rather than the CEFR's four modes of language use (to which only Carolyn Westbrook and Aidan Holland refer). Given that the CEFR was first published a quarter of a century ago, this is both strange and shocking. Clearly, alignment with the CEFR can never be entirely valid if this mismatch is maintained—something to which a revised edition of the Handbook should give prominence.

The CEFR's action-oriented approach. The 2001 CEFR describes its approach to the description of language use as "action-oriented". The term coheres with the concept of language user/learners as "social agents … who have tasks (not exclusively language-related) to accomplish" (Council of Europe [CoE] 2001: 9), and its pedagogical implications are confirmed by the text box at the beginning of Chapter 2: "Language use, embracing language learning …" (CoE 2001: 9). The 2001 CEFR nevertheless declines to make pedagogical recommendations: "it is not the function of the Framework to promote one particular language teaching methodology, but instead to present options" (Council of Europe 2001: 142). By contrast, the 2020 CEFR/CV advocates an action-oriented approach to language teaching and learning (see also Piccardo and North 2019), though without explaining this significant change of stance

or the relation the new stance assumes between language teaching/learning and language use. Based on the concept of a Comprehensive Learning System (O'Sullivan 2020), the Alignment Handbook is concerned with all dimensions of language learning—curriculum, materials and teaching/learning as well as assessment. Clearly, a revised edition will need to explore what an action-oriented approach to teaching and learning entails, and how precisely it is related to an action-oriented description of language use.

Making greater use of the CEFR's discursive content. To date, regardless of its specific focus, alignment with the CEFR has tended to focus on levels, scales and descriptors; rather little attention has been paid to the discursive content of either the CEFR or the CEFR/CV. The article by Diego Cortés Velásquez and Elena Nuzzo provides a welcome exception to this general rule: the project on which they report started from the discussion of tasks in Chapter 7 of the 2001 CEFR. When revising the Alignment Handbook, the new editors could usefully refer to key conceptual discussions in the CEFR and CEFR/CV, including the discursive characterizations of the proficiency levels in Chapter 3 of the 2001 CEFR (CoE 2001: 33-37) and Appendix 1 of the 2020 CEFR/CV (Council of Europe 2020: 173-175).

The role of AI technologies. In the five years since work began on the Alignment Handbook, the potential of AI systems to support and perhaps abbreviate alignment processes has expanded significantly. Clearly, this must be addressed in a revised edition.

At the end of 2025, and following the publication of a brief supplement to the Handbook in which some of these issues are addressed, the editorial group responsible for the Alignment Handbook—Neus Figueras, David Little, Barry O'Sullivan, Nick Saville and Lynda Taylor—passed the baton to a new group—Dave Allan, Armin Berger, Nathaniel Owen, Graeme Seed and Carolyn Westbrook. These colleagues and the institutions they represent—the British Council, EALTA, UKALTA and ALTE—have agreed to continue the work that started in 2018. We conclude, however, by pointing out that the ultimate responsibility for honest and successful use of the CEFR and the Handbook, and for any consequent improvements in language education, lies with everyone involved in language education.

References

Kanistra, Paraskevi. 2025. *Evaluating the Item Descriptor (ID) Matching method in a face-to-face and a synchronous virtual environment*. Peter Lang.

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

Piccardo, Enrica & Brian North. 2019. *The action-oriented approach: A dynamic vision of language education*. Multilingual Matters.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Title: **CEFR Journal—Research and Practice**

Type: Online journal
URL: https://cefrjapan.net/journal
Contact: journal@cefrjapan.net

Edited by: Japan Association for Language Teaching (JALT)
CEFR & Language Portfolio SIG
Fergus O'Dwyer (editor)
Dmitri Leontjev (editor)
Elif Kantarcıoğlu (editor)
Maria Gabriela Schmidt (coordinator, editor)
Morten Hunke (liaison officer, editor)
Alexander Imig (treasurer, website editor)

ISSN: 2434-849X
DOI https://doi.org/10.37546/JALTSIG.CEFR

Publication Statement of the CEFR Journal

All work published in the CEFR journal is original work, with contributions from the named authors.

The work published in the CEFR journal is open access and licensed under a Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

Authors of articles are allowed to retain publishing rights and hold the copyright without restrictions. The CEFR Journal actively checks for plagiarism using Turnitin. The CEFR Journal does not charge article processing charges or submission charges.

Any issues regarding complaints and appeals, conflicts of interest / competing interests; ethical oversight; post-publication discussions and corrections should be directed to journal@cefrjapan.net

The administration costs of the CEFR Journal are covered by the JALT CEFR & Language Portfolio SIG; the CEFR Journal has no paid advertisements or other revenue sources.

In the event that the CEFR Journal is no longer published, electronic backup and preservation of access to the journal content shall be made available via CLOCKSS.

Submission (Call for Papers)

This journal attempts to fall somewhere in between an inaccessible academic journal (long waiting times, fairly strict guidelines/criteria) and a newsletter (practical in nature but lacking in theoretical support/foundation), linking research of a practical nature with relevant research related to foreign language education, the CEFR, other language frameworks, and the European Language Portfolio. While the CEFR was introduced by the Council of Europe and intended for use, first and foremost, within Europe, the influence of the CEFR now needs to be attested in many places beyond European borders. It has become a global framework, impacting a variety of aspects of language learning, teaching, and assessment across countries and continents beyond the context for which it was originally created. As such, there is a pressing need to create a quality forum for sharing research, experiences, and lessons learned from applying the CEFR in different contexts. This journal provides such a forum where people involved or interested in processes of applying the CEFR can share and learn from one another.

We are continuously seeking contributions related to foreign language education, the CEFR, other language frameworks, and the European Language Portfolio. We are particularly interested in specific contextual adaptations.

Currently, we have a new Call for abstracts out. Due to current necessities and demand, we are looking to give your experiences with online, remote, and e-learning in conjunction with the CEFR, the CEFR/CV, or portfolio work the spotlight it deserves. In these months many practitioners are accruing valuable best and potentially also worst practice experience. We would like to offer a forum to share such valuable insights in future volumes. Until 30 August 2026 we are looking for abstracts at: journal@cefrjapan.net.

Please contact the editors with any queries and submit to: journal@cefrjapan.net

Guidelines

Submission:	31 August 2026
Contributions:	Articles (research), reports (best practice), news (work in progress), research notes, book reviews
Language(s):	English (British, American, international) preferred, but not mandatory. Other languages by request, with an extended abstract in English.
Review type:	Peer review, double blind

Reviewer guidelines:

We ask all reviewers to make every reasonable effort to adhere to the following ethical guidelines for **CEFR Journal—Research and Practice** articles they have agreed to review:

1. Reviewers must give an unbiased consideration to each manuscript submitted for consideration for publication, and should judge each on its merits. Since, we employ a double-blind review, the text you have been provided with ought to have no reference to race, religion, nationality, sex, gender, seniority, or institutional affiliation of the author(s). Please, notify us immediately where any such information is still detectable in the anonymised text you received.
2. Reviewers should declare any potential conflict of interest prior to agreeing to review a manuscript, including any relationship with the author that may potentially bias their review.
3. Reviewers are strongly advised to keep the peer review process confidential; information or correspondence about a manuscript should not be shared with anyone outside the peer review process.
4. Reviewers should provide a constructive, comprehensive, evidenced, and appropriately substantial peer review report. For your convenience, we are providing you with a 'reviewing matrix' you may choose to use at your own discretion. We would also like to kindly ask you to provide us in the journal editorial team with a final overall assessment of the text's publication potential—please, see bottom of this document.
5. Reviewers must avoid making statements in their report, which might be construed as impugning any person's reputation.
6. Reviewers should make all reasonable effort to submit their report and recommendation in a timely manner, informing the editor if this is not possible.
7. Reviewers should call to the journal editor's attention any significant similarity between the manuscript under consideration and any published paper or submitted manuscripts of which they are aware.

Author instructions:

Adapted version of deGruyter Mouton guidelines for Language Learning in Higher Education (CercleS) and style sheet.

CEFR Journal—Research and Practice

Japan Association for Language Teaching (JALT)
CEFR & Language Portfolio SIG (CEFR & LP SIG)

ISSN 2434-849X