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

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.

> Describe the last job you had or a job you would like to do. You must write between 50 and 80 words.
>
> Write your answer on the Answer Sheet.

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.

> Write an email to a friend and talk about a trip you would like to take soon.
>
> You must write between 50 and 80 words. Write your answer on the answer sheet.

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.