Researching Generative AI in Applied Linguistics

Edited by
Carol A. Chapelle,
Gulbahar H. Beckett, and Bethany E. Gray

IOWA STATE UNIVERSITY
Digital Press

Researching Generative AI in Applied Linguistics

Edited by

Carol A. Chapelle, Gulbahar H. Beckett, and Bethany E. Gray

Iowa State University Digital Press

Ames, Iowa

Typeset in Times New Roman.
Book Cover Design: Hannah Litterer

ISBN: 978-1-958291-12-2
DOI: https://doi.org/10.31274/isudp.2025.211

Published by

Iowa State University Digital Press
701 Morrill Rd, Ames, Iowa 50011, United States

www.iastatedigitalpress.com

Iowa State University is located on the ancestral lands and territory of the Baxoje (bah-kho-dzhe), or Ioway Nation. The United States obtained the land from the Meskwaki and Sauk nations in the Treaty of 1842. We wish to recognize our obligations to this land and to the people who took care of it, as well as to the 17,000 Native people who live in Iowa today.

Chapter 10

A Validation Study of AI-Generated Prompts in CILS (Certification of Italian as a Foreign Language) B2 Exams

Giulia Peri, Sabrina Machetti, and Paola Masillo
University for Foreigners of Siena

This study presents an ongoing validation analysis of generative AI (ChatGPT-4; OpenAI 2023) in creating test prompts for the written production component of the CILS (Certification of Italian as a Foreign Language) B2 exam. As AI technologies increasingly influence language assessment and test development, a key challenge is ensuring that AI-generated writing test prompts maintain validity, appropriateness, and alignment with assessment constructs. This issue is particularly relevant for high-stakes language certifications, where the demand for new test items is growing, but the number of trained item writers remains limited. The research aims to evaluate ChatGPT-4's capability to generate writing test prompts reflecting the CILS B2 target domain and to investigate how these prompts are perceived by CILS experts and test-takers. Following an argument-based validity framework (Chapelle & Voss, 2021), the study employs a mixed-methods approach to collect evidence for the domain definition inference.

Introduction

The Certification of Italian as a Foreign Language (*Certificazione di Italiano come Lingua Straniera* - CILS) is an Italian language qualification offered by the CILS Center of the University for Foreigners of Siena (Italy). CILS exams are standardized tests of Italian measuring general language proficiency, are aligned with the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* ('CEFR'; CoE, 2001) and are staged

Cite as: Peri, G., Machetti, S., & Masillo, P. (2025). A validation study of AI-generated prompts in CILS (Certification of Italian as a Foreign Language) B2 exams. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 197-218). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.10

on the six CEFR levels (A1 to C2). The target audience for the CILS exams consists of a general adult population with an average cultural background who study the Italian language for educational, professional, or general cultural purposes, both in Italy and abroad.

CILS exams are recognized by universities, employers, and institutions worldwide. Since their introduction in 1993 and over 30 years of activity, approximately 400,000 test-takers have taken the CILS exams in Italy and 87 foreign countries. Currently, all CILS exams are in paper format. Therefore, they are administered in person at CILS Test Centers and are entirely assessed centrally at the CILS Center in Siena (Machetti, 2022). As stated in the CILS Guidelines (Centro CILS, 2009), CILS exams are text and task based. Thus, they present test takers with tasks that relate to situations representative of real-life ones, with the aim of measuring the ability of test-takers to use the Italian language in various domains and contexts. Each CILS exam consists of five components, measuring the following linguistic and communicative skills: listening comprehension, reading comprehension, use of Italian, written and oral production and interaction. Proficiency in Italian written and oral production and interaction is tested through extended-production tasks, which prompt linguistic performance in authentic contexts (CoE, 2001). Task difficulty varies based on the CEFR indications, with moderating factors such as time constraints and type of support provided (Centro CILS, 2009). Regarding the scoring of extended-production tasks in the written component of CILS exams, trained and standardized raters assess test-takers performances based on rating criteria and elaborated rating scales with descriptors, developed according to the proficiency framework. These criteria include communicative effectiveness, register and stylistic appropriateness, morphosyntactic accuracy, lexical adequacy and richness, as well as spelling and punctuation.

The entire CILS test cycle occurs with the support of basic information technology at every stage (e.g., computers, internet, WordOffice, SPSS). However, as the rapid expansion of technology-mediated language assessment develops, test developers are facing the challenge of producing high-quality items efficiently and cost-effectively, stressing the need for practical solutions without impacting negatively on validity (Rossi, 2023). More specifically, the CILS Centre faces challenges related to the increasing workload and the scarcity of well-trained item writers (Gallina, Machetti, & Masillo, 2019). Moreover, the rapid development of artificial intelligence (AI), its spread in the language testing field (Voss, 2024) and changing learner needs prompts the alignment of the assessment process with contemporary technological and educational needs.

Among the most notable advancements in AI-driven language tools is ChatGPT, an advanced model developed by OpenAI (OpenAI, 2023). Built on the Generative Pre-trained Transformer (GPT) architecture, it excels in conversation, content creation, coding, and tutoring. Pre-trained on vast text data and fine-tuned with human feedback, it generates coherent and context-aware responses during interactions. In Italy, the use of ChatGPT began to spread primarily from 2023 onward, with usage mainly linked to informal contexts. A particularly significant aspect of this situation is that ChatGPT was also at the center of much controversy, especially in 2023 and particularly in Italy, where it was blocked by the Italian Data Protection Authority due to privacy concerns. The blocking of ChatGPT may have contributed, directly or

indirectly, to the development of a more skeptical or negative perception of AI tools among both professionals and the public. On the other hand, the challenge of developing and fine-tuning AI technologies for non-English languages is a critical issue in the global deployment of AI solutions (Robinson et al., 2023).

The predominance of English in AI development not only impacts the effectiveness of the tool in other linguistic contexts but also its cultural sensitivity and relevance. For instance, English-speaking countries have explored the topic of AI for education more extensively, with cutting-edge studies and practical applications leading the way. For instance, many are the studies looking at the interrelationships between language assessment and AI (e.g., Van Moere & Downey, 2016; Voss et al., 2023; Xi, 2023) or item development and AI (e.g., Attali et al., 2022; Gierl et al., 2012). In the Italian context, research on AI-driven language learning, teaching and assessment is still emerging (e.g., in the field of learning and teaching, see Cinganotto, Sbardella & Montanucci, 2024). However, even in contexts outside of Italy, there appears to be a lack of published studies specifically focusing on the use of AI to generate writing test prompts for written production tasks in the context of formal testing. In this context, the main challenge for the CILS Center is maintaining a high level of validity in its tests while ensuring greater practicality and operational sustainability, qualities that contribute to the usefulness of the test (Bachman & Palmer, 1996). This calls for a rethinking of the test design and administration methods, ensuring that advancements in AI and technology are leveraged effectively without compromising assessment quality.

In response to these challenges, our study explores the feasibility of using ChatGPT-4 to generate writing test prompts for the CILS exams. To ensure a rigorous validation process, this study adopts an argument-based validity framework (Chapelle, Enright, & Jamieson, 2008; Chapelle & Voss, 2021), focusing on the domain definition inference, as previously investigated by research in the English context (e.g., Jun, 2021; Knoch & Chapelle, 2018; Lee, 2023).

The argument-based validity framework is based on the "construction of an interpretative argument that lays out the grounds, inferences, warrants, and claims, all of which provide the foundation for proposed score interpretation and use" (Chapelle, Enright, & Jamieson, 2008, p. 23). At the foundation of the argument is the domain definition inference that establishes the link "between performances in the target domain to the observations of performance in the test domain" (Chapelle, 2012, p. 22; Chapelle, Enright, & Jamieson, 2008, p. 14). As for the "conceptual tools for developing validity arguments" (Chapelle, 2012, p. 20), the warrant supporting the domain definition inference is that observations of performance on the test are representative of performances in the target domain of language use (Chapelle, Enright, & Jamieson, 2008, p. 17). In turn, the warrant relies on two assumptions: (a) that assessment task representative of the target domain can be identified; and (b) that such assessment tasks can be simulated as test tasks. The backing for these assumptions involves a thorough domain analysis and meticulous task construction, as well as evaluations of the success of these processes' outcomes (Chapelle, Enright, & Jamieson, 2008, pp. 17-21). In the writing assessment, investigation of the domain definition inference reveals whether the writing test prompts elicit test-takers' performances that are representative of their writing performances in the target

domain. This is a fundamental concern for all test prompts, as their validity depends on whether they effectively assess the intended construct. However, when writing test prompts are generated by AI rather than human item writers, new factors must be considered, and they must be evaluated to determine whether they align with the established assessment construct and proficiency level.

Given these considerations, our study seeks to collect empirical evidence for the domain definition inference, examining whether AI-generated writing test prompts are representative of the CILS B2 writing test target domain and can be used as writing test prompts for the tasks in the CILS B2 written exam. The domain definition inference is based on the warrant that observations of performance on AI-generated writing test prompts in the CILS B2 exam are representative of the writing skills required at the CEFR B2 level, to which this CILS exam is aligned. This warrant is grounded in three key assumptions:

1. The CILS B2 construct for written interaction and production is well-defined.
2. AI-generated writing test prompts can be designed to be representative of the CILS B2 writing skills (RQ1).
3. AI-generated writing test prompts are appropriate for eliciting B2-level written interaction and production (RQ2).

To support the first assumption, a detailed analysis was conducted to examine the domain relevant to the written interaction and production required for the CILS B2 exam, drawing on the CILS Guidelines (2009) and CEFR B2 descriptors (CoE, 2001). This process confirmed the already established construct definition for written interaction and production tasks at the CILS B2 level. An overview of the construct definition is provided in Appendix A. Since the construct was already explicitly outlined, it did not require further empirical validation in this study.

For the second assumption, empirical validation was necessary to determine whether AI-generated prompts can be identified as representative of the target domain, the CILS B2 writing skills. To assess this, experts were involved in the evaluation of AI-generated writing test prompts, serving as backing for (RQ1). For the third assumption, the study examined test-taker perceptions of AI-generated prompts in terms of difficulty, clarity, and relevance (RQ2).

Therefore, to achieve the goal of this study, the following research questions were formulated stemming from the assumptions illustrated above:

1. To what extent does ChatGPT-4 generate writing test prompts that are perceived as representative of the CILS B2 target domain, according to expert evaluation?
2. How do test-takers perceive ChatGPT-4-generated writing test prompts compared to those created by human item writers in terms of difficulty, clarity, and relevance to the CILS B2 target domain?

Methods

Research Context and Design

This study was conducted using a mixed-method approach (see Instruments for more details) within the context of the CILS B2, which assesses test-takers 'proficiency in Italian as a S/FL at the CEFR B2 level. The exam tests multiple skills (see Introduction), including written interaction and production, organized in two tasks (see Appendix A). As already mentioned in the Introduction, the study used an argument-based approach referencing Chapelle and Voss (2021), with a specific focus on collecting evidence for the domain definition inference (Chapelle & Voss 2021, p. 35).

Participants

CILS Item Writers

The study involved seven Subject-Matter Experts (SMEs) specializing in Second/Foreign Language (S/FL) Italian Language Testing and Assessment. These experts were experienced CILS item writers who had been working for the CILS Center in various senior roles for 7 to 30 years. Their academic qualifications ranged from Master's degrees to PhDs, and they had extensive experience in language test development, task design, and proficiency assessment, particularly within the CILS certification framework.

Test-Takers

The participant sample consisted of 63 test-takers, aged 16 to18, enrolled in an Italian high-school in Istanbul (Türkiye). A purposive sampling method was adopted, as test-takers were specifically selected based on their upcoming participation in the CILS B2 exam in one of the forthcoming sessions organized by the CILS Test Center in Istanbul. This sampling strategy ensured that participants had a relevant proficiency level (CEFR B1 or B2) and a real test-taking motivation, making them suitable for evaluating the AI-generated and human-created writing test prompts.

Instruments

The study employed a set of instruments designed to address the research questions (RQs) by collecting backing (empirical evidence) to support or challenge the assumptions underlying the warrant that AI-generated writing test prompts are representative of the target domain.

To address RQ1, we employed the following instruments:

- Two writing prompt models for ChatGPT-4: Predefined writing task prompt models from a previous CILS B2 exam administration (one for Task 1 and one for Task 2) were used as reference for writing test prompt generation.
- AI-generated writing test prompts: Writing test prompts (no. 20) for both Task 1 and Task 2 were generated using ChatGPT-4, following structured input parameters describing CILS B2 human-made writing tasks.
- Expert survey (Survey 1): Survey 1 was designed to gather SMEs' evaluations of AI-generated writing test prompts, focusing on their appropriateness, alignment with the B2

construct, and AI detectability. The survey was administered online and structured into two sections, corresponding to the two writing tasks (Task 1 and Task 2) of the CILS B2 exam. The survey included multiple-selection questions for evaluating the writing test prompts (1-10) and open-ended responses for justifying each selection or give additional comments. Experts were unaware that all prompts had been generated by AI.

To address RQ2, we used the following instruments:

- Human-generated and AI-generated tasks: both human-created writing test prompts (i.e., the writing test prompt models) and two expert-selected AI-generated writing test prompts (one for Task 1 and one for Task 2) were administered to test-takers in a paper-and-pencil format under standardized test conditions. A total of four writing test prompts were administered (two for Task 1 and two for Task 2).

- Test-taker survey (Survey 2): Survey 2 was designed to gather test-taker perceptions after administration regarding the difficulty, clarity and relevance of writing test prompts without revealing whether they were AI- or human-generated. The survey was administered in-person and structured into two sections, corresponding to the two writing tasks (Task 1 and Task 2) in the CILS B2 exam. Each section contained multiple-choice questions (i.e., Likert scale, Yes/No) and open-ended responses, allowing participants to provide both quantitative ratings and qualitative insights on the prompts they reviewed.

Data Collection and Analysis

To collect backing for the domain definition inference, this study adopted a mixed-methods approach, combining descriptive frequency analysis of survey responses with content analysis of participants' evaluations. Data collection was structured to align with the two research questions, ensuring that both SMEs assessments (RQ1) and test-taker perceptions (RQ2) were systematically examined.

RQ1. To What Extent does ChatGPT-4 Generate Writing Test Prompts that are Perceived as Representative for the CILS B2 Domain, according to Expert Evaluation?

To answer this first research question, it was necessary to produce the main materials for the study, namely the writing test prompts. Before generating the writing test prompts with AI, a writing test prompt model from tasks developed by humans for the AI to use as an example was identified (Attali et al., 2022; Bejar, 2002). Two specific writing test prompts, named "Task 1" and "Task 2", were selected from a previous session of the CILS B2 exam (see Appendix B). This choice was guided by post-test statistics, routinely calculated by the CILS Center, which indicated the high quality of these writing test prompts in terms of validity and alignment with the measured construct. The study proceeded with the AI generation of 10 writing test prompts for Task 1 and 10 writing test prompts for Task 2, combining a model-based approach and a rule-based approach methodology (adapted from Kurdi et al., 2020) with ChatGPT-4. The rules provided to the AI through prompts referred to the CEFR B2 descriptors and to the structures identified by the CILS Guidelines for B2-level written interaction and production. Temperature settings were not modified, meaning ChatGPT-4 operated with its default deterministic

configuration to maintain output consistency and reproducibility across multiple prompt generations. Since ChatGPT is known to reflect biases present in its training data (Liu et al., 2024), precautions were taken to prevent the inclusion of inappropriate or sensitive content in the AI-generated writing test prompts. Specifically, the AI was instructed to adhere to the "CILS taboo list" (internal document), which excludes topics related to crimes, war, physical and mental health, etc.

A survey for SMEs (Survey 1) was developed to elicit the SMEs assessment of the AI-generated writing test prompts' adequacy and their detection of any perceived artificial intelligence involvement. Specifically, the survey aimed to gather SMEs input on the appropriateness and authenticity of writing test prompts for assessing writing ability at the B2 level of the CILS exam. SMEs first selected which writing test prompts they considered suitable for Task 1, evaluating aspects such as thematic relevance, difficulty, and alignment with CILS standards. They were asked to explain their selections, providing qualitative insights into the criteria they used. SMEs were then asked to identify which writing test prompts they believed were generated with the support of AI, without knowing that all writing test prompts were AI-generated. They were also invited to explain their reasoning, offering valuable qualitative data on the features or characteristics they associated with AI involvement. These questions were consistently applied to both Task 1 and Task 2 to allow for a comprehensive evaluation. Therefore, SMEs provided quantitative and qualitative feedback to investigate whether they saw AI-generated writing test prompts aligning with the B2 construct (Assumption 2).

RQ2. How do Test-takers Perceive ChatGPT-4-generated Writing Test Prompts Compared to Those Created by Human Item Writers in terms of Difficulty, Clarity, and Relevance to the CILS B2 Target Domain?

Two writing test prompts (one for Task 1 and one for Task 2) were selected from AI-generated outputs by two SMEs (see Appendix B), who conducted a preliminary review to ensure that the writing test prompts did not exhibit substantial divergence from CILS B2 requirements or present issues impeding test takers' performance. This step served as a procedural safeguard prior to piloting, rather than as an intervention to modify the AI-generated content. No substantive changes were made, thereby preserving the integrity of the AI-generated writing test prompts for performance evaluation. These writing test prompts, along with those developed by human item writers, were administered to a sample of 63 test-takers under standardized testing conditions consistent with CILS requirements. Test-takers are allotted 3 hours and 45 minutes for the full CILS B2 exam, which includes all components except for the speaking test. The written production and interaction component typically lasts 1 hour and 10 minutes, covering Task 1 and Task 2. However, in this study, only the written interaction and production section was administered. Since each task included two writing test prompts—one human-generated and one AI-generated—test-takers responded to four writing test prompts in total (two for Task 1 and two for Task 2). To account for the additional writing test prompts, the total time allotted for this session was 2 hours and 20 minutes.

Survey (Survey 2) responses from test-takers were analyzed to evaluate perceptions of

the writing test prompts. The survey explored familiarity with AI, experiences with the tasks, and perceived difficulty. Open-ended responses were further examined to identify themes, including sentiment toward the content and design of the prompts. These findings provided valuable insights into the extent to which the prompts adhered to the test takers' perspectives of expected characteristics of written production tasks in the CILS B2 exam. Therefore, test-takers' responses helped to investigate whether AI-generated prompts are perceived by test takers as appropriate and comparable to human-generated prompts, thereby supporting or challenging Assumption 3.

This study focuses solely on expert and test-taker perceptions, without analyzing test-taker performances. Future research could extend this work by examining test-taker responses to AI-generated versus human-crafted prompts, to further assess whether AI-generated prompts not only align with CILS task specifications but also successfully elicit B2-level writing performances.

Results

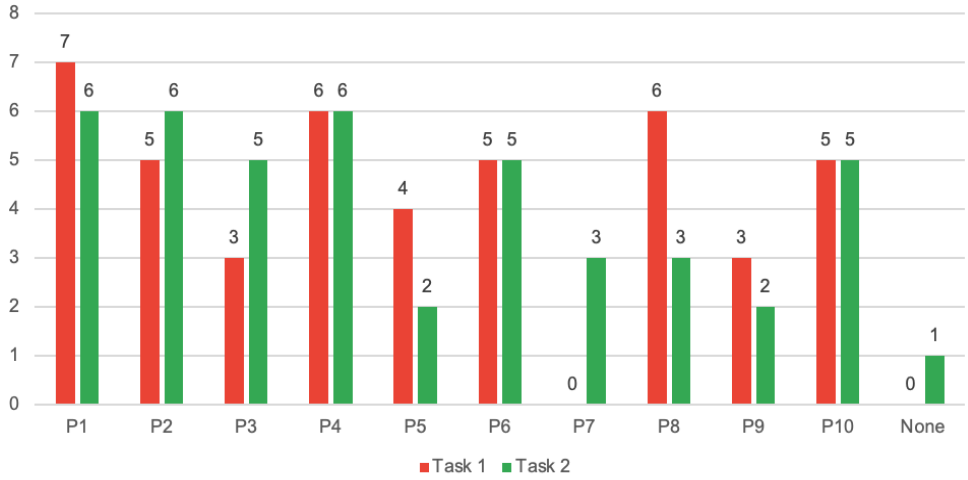
For RQ1, the results of Survey 1 were analyzed. Regarding the first question of Survey 1, the SMEs' responses are summarized in Figure 10.1, which also includes the content of the question in the caption. Figure 10.1 refers to the responses related to both the writing test prompts generated for Task 1 (no. 10) and those generated for Task 2 (no. 10).

For Task 1, Prompt 1 (P1) received unanimous approval from all SMEs, along with Prompts P4, P5, P8, and P10, which were also frequently identified as suitable. In contrast, P7 was not selected by any SME. For Task 2, P1, P2, P4, and P10 received the highest endorsements. However, there was greater variability for the remaining prompts. Prompt P5 and P9 were least endorsed, with only 2 experts selecting them. Notably, one SME deemed none of the writing test prompts appropriate for Task 2. Overall, across the 20 AI-generated writing test prompts, there was considerable variation in the acceptability among experts.

Regarding the open-ended question addressed to SMEs about the reasons behind the adequacy of each writing test prompt, their responses focused on the following aspects. For the prompts generated for Task 1, the SMEs considered the selected writing test prompts more or less relevant based on their suitability for assessing written production and their alignment with the B2 level descriptors of the CEFR (e.g., SME.6 "The writing test prompt is appropriately structured for a B2 level because the candidate is required to reflect on a topic, describe a personal experience in detail by explaining its advantages and disadvantages, and provide arguments."). Additionally, the SMEs reflected on the appropriateness of the content, the use of similar writing test prompts in previous CILS exam sessions with which they were familiar, and the use of specific verb forms (e.g., SME.5 "The test-takers is required to produce a text on a topic that may fall within their field of interest, using arguments and examples to support their ideas [B2]. The verb 'talk about' [ita. *parla*] is more suitable for a speaking task, so I would suggest changing it accordingly.").

Figure 10.1

Survey 1, Question 1: Number of SMEs Selecting 20 AI-generated Prompts as Suitable for Task 1 (Red) and Task 2 (Green) in the CILS B2 written production component (N=7)



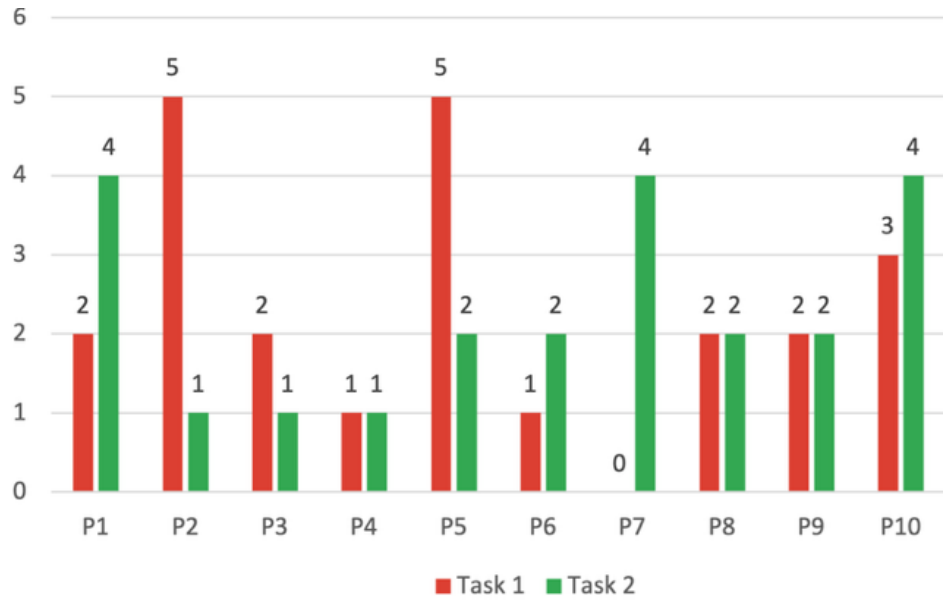
Note. P = prompt; None = no prompt is appropriate.

For the writing test prompts generated for Task 2, the SMEs primarily commented on their adequacy by considering morphosyntactic elements, the characteristics of the prompt formulation, and their relevance for written production and the complexity of the themes addressed (e.g., SME.4 "The topic should be related to the test-taker's field of interest at the B2 level"). Furthermore, for Task 2, the SMEs found it particularly useful to reflect on the interest and engagement generated by the prompts and mentioned the appropriateness of certain specific terms used in the prompts (e.g. SME.5 "I would remove the gerund—as it is not a structure typically expected at B2—and replace the term 'dean' [ita. *decano*] with a more general one.").

Regarding the second survey question on the use of AI for generating the writing test prompts, Figure 10.2 presents the collected data. As with the previous case, Figure 10.2 includes responses related to both the writing test prompts generated for Task 1 (no. 10) and those generated for Task 2 (no. 10).

Figure 10.2

Survey 1, Question 2: Number of SMEs Identifying 20 Writing Test Prompts as AI-generated (N=7)



Note. P = prompt

For Task 1, prompts P2 and P5 were most frequently suspected of being AI-generated, with 5 indications each. Some SMEs also identified P1, P3, P9, and P10 as potentially AI-generated, while P7 received no such indications. Notably, 2 SMEs indicated they believed none of the writing test prompts were AI-generated. For Task 2, prompts P1, P7, and P10 received the highest number of AI-generation suspicions, with 4 indications each. As with Task 1, 2 SMEs believed that none of the prompts were AI-generated.

The distribution of responses and accompanying comments suggest that there is no unanimous perception among experts about which prompts were AI-generated (e.g., SME.3 "There are no indicators that can distinguish whether the inputs were formulated by a human item writer or by artificial intelligence."). Also, this lack of agreement suggests the possibility that the AI-generated writing test prompts were not easily distinguishable from those created by humans. It is also plausible that their diversity was sufficient to approximate the variety typically expected in human-generated writing test prompts.

For RQ2, and after a human preliminary review process (see Data Collection and Analysis), the two best generated prompts—one for Task 1 and one for Task 2—were selected to be administered alongside the human-generated writing test prompts. Prompts 1 and 2 were the human-generated ones, while Prompts 3 and 4 were the AI-generated ones. However, this information was never revealed to test-takers. After the administration, test-takers responded to Survey 2. The following figures present the data collected.

The first question of the second survey investigated the perceived difficulty of the writing test prompts according to test-takers. The test takers' responses to this question are summarized in Figure 10.3 below.

Figure 10.3

Survey 2, Question 1: Histograms Showing Test Takers' Judgments of the Difficulty of Four Prompts (Prompt 1 and 2 Written by Humans; Prompts 3 and 4 AI-Generated) (N=63)

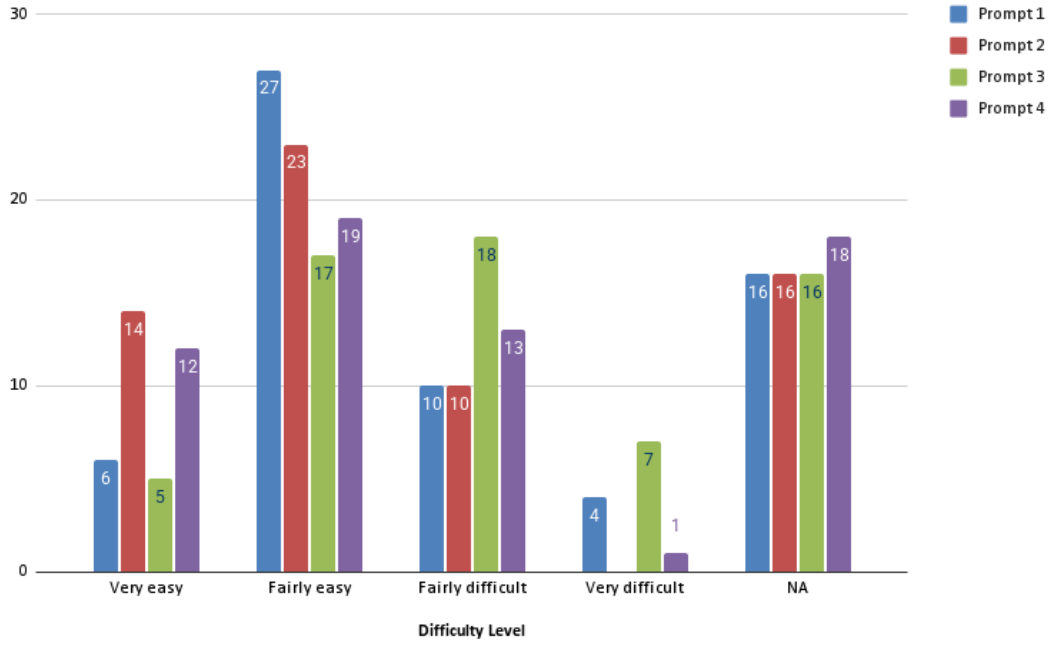


Figure 10.3 shows that "Fairly easy" has the highest number of responses in most cases, particularly in Prompts 1 and 2. The "Fairly difficult" rating is consistent but not dominant across prompts, with a spike in Prompt 3 (18 responses). In fact, there is a noticeable rise in both the "Fairly difficult" and "Very difficult" responses in Prompt 3, which indicate that this writing test prompt was perceived as more challenging compared to others. In contrast, Prompt 2 has higher "Very easy" and "Fairly easy" ratings, suggesting it was the least challenging overall.

A comparison between the human-generated prompts (1&2) and AI-generated ones (3&4) reveals differences in perceived difficulty, suggesting that AI-generated writing test prompts were generally perceived as more difficult. As shown in Figure 10.3, human-generated writing test prompts were more frequently rated as easier. For Prompt 1, 33 (70.2 % of valid responses) test takers rated it as "Very easy" or "Fairly easy", and for Prompt 2, 37 (78.7% of valid responses) test takers gave it these rating, making it the easiest overall. In contrast, the AI-generated writing test prompts, particularly Prompt 3, were perceived as more difficult. As mentioned before, Prompt 3 was judged as "Fairly difficult" or "Very difficult" by 25 (53.2 % of valid responses) test takers.

To better interpret test takers' perceptions of the writing test prompts, the second and third questions of Survey 2 investigated the familiarity of test takers with AI. Familiarity with AI

was considered a contextual factor to take into account, as individuals with greater exposure to AI might have different perceptions or expectation regarding the writing test prompts' characteristics. Furthermore, because the study investigated test takers perceptions of potential AI involvement in writing test prompt creation (Figure 10.6), understanding their baseline familiarity with AI was essential to interpret their judgements.

The data related to AI familiarity is presented in Figures 10.4 (Test-takers' knowledge of AI) and 10.5 (Test-takers' usage of AI) below.

Figure 10.4

Survey 2, Question 2: Test-takers' Knowledge of AI (N=63)

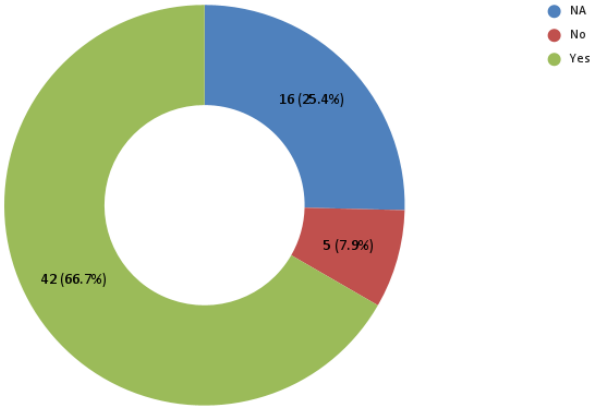


Figure 10.5

Survey 2, Question 3: Test-takers' Usage of AI (N=63)

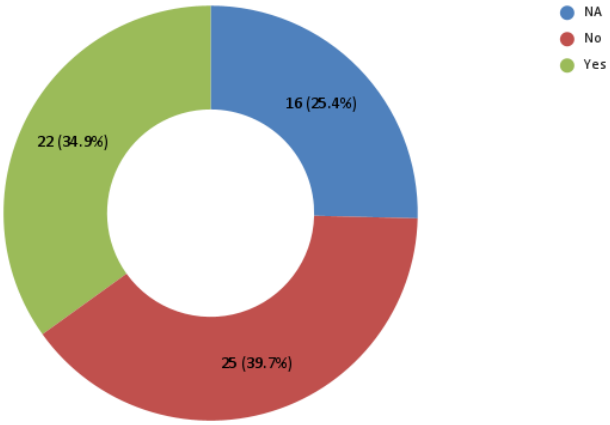


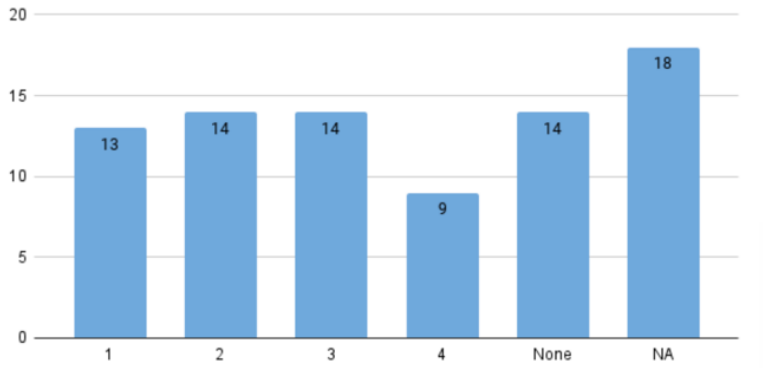
Figure 10.4 reveals that the majority of test-takers (66.2%) were familiar with the concept of AI. However, as shown in Figure 10.5, a comparable proportion of test-takers reported never having used AI (39.7%) and having used it in the past (34.9%).

For the fourth question of the second survey, test-takers were asked to give their opinion on the potential support of AI in creating the prompts they responded to. The data collected is

presented in Figure 10.6 below.

Figure 10.6

Survey 2, Question 4: Test-takers' Perception of AI Involvement in Writing Test Prompts (N=63)



Note. NA = the number of test-takers who did not respond.

As can be seen in Figure 6, Prompts 2 and 3 received the highest number of responses (14 each), indicating that test-takers were more likely to attribute these writing test prompts to AI (while only Prompt 3 was AI-generated) compared to Prompts 1 and 4 (while only Prompt 1 was human-generated). Interestingly, an equal number of test-takers (no. 14) believed that none of the prompts were created using artificial intelligence, reflecting a degree of skepticism or uncertainty regarding AI's role in content generation. Additionally, 18 test-takers did not engage with the question (NA), the highest count among all response categories. This may suggest that the question was unclear to some test-takers or that they lacked confidence in their ability to discern AI-generated content.

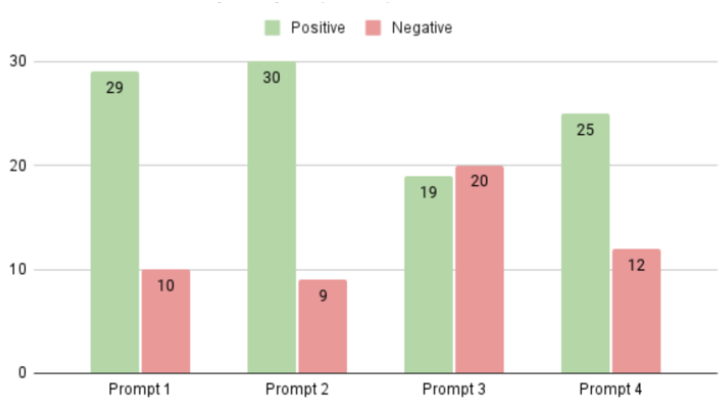
The final question of the second survey invited test-takers to reflect on their experiences with the writing test prompts in greater detail. Test-takers were asked to provide a brief description of their feelings while responding to each prompt. This included sharing whether they enjoyed the task, any difficulties they encountered, and whether they found the topic appropriate for their age group. The open-ended nature of this question aimed to capture nuanced insights into their engagement with the writing test prompts and their overall perceptions of the task. Test-takers' responses were grouped based on sentiment towards the content of the writing test prompt (Figure 10.7). Positive responses where test-takers expressed enjoyment, ease, or appropriateness of the topics were coded as positive (e.g., TT[2].3 "It was easy to write because I had already done something similar in my language course."). Negative responses, where test-takers mentioned difficulties, discomfort, or felt the topic was inappropriate for their age or context were coded as negative (e.g., TT.6 "The topic […] it's not very suitable for my age since we don't have much experience."). In two instances the test-takers' response was not clearly identifiable as positive or negative, as their language proficiency was probably too low to explain

[2] TT = Test taker

their point of view clearly (e.g., TT.9 “If I don’t know the language I want to use, I will have difficulty writing the email.”). These responses were not categorized.

Figure 10.7

Survey 2, Question 5: Sentiment Analysis of Test-takers Responses (N=39)



As shown in Figure 7, the analysis of sentiment toward the four writing test prompts reveals variations in test-taker perceptions. Prompt 2 received the most positive feedback, with 30 test-takers expressing favorable views and only 9 negative responses, suggesting it was the most well-received and effective writing test prompt. Similarly, Prompt 1 was positively regarded by the majority, with 29 positive responses and 10 negative ones, indicating strong overall approval. Prompt 4 also elicited predominantly positive feedback (25 positive responses), though it had slightly higher negative sentiment (12 responses) compared to Prompts 1 and 2. In contrast, Prompt 3 stood out as the most polarizing, with almost an equal split between positive (no. 19) and negative (no. 20) sentiments. This division may indicate challenges with its content, complexity, or alignment with expectations.

Discussion

The findings of this study provide empirical evidence on ChatGPT-4’s capability to generate writing test prompts that align with the CILS B2 exam, offering backing for the domain definition inference (Chapelle & Voss, 2021). Considering the judgment of SMEs, the study found an overlap between writing test prompts identified as AI-generated and those considered adequate for the exam, supporting Assumption 2 that AI-generated content can meet expert standards for B2-level writing prompts. However, expert feedback suggests areas for improvement, particularly by refining language complexity and ensuring alignment with the exam’s format, which could further refine the quality of AI-generated writing test prompts. This finding emphasizes the need for human oversight in AI-generated test materials.

Test-takers provided backing for Assumption 3 by expressing their judgment on the writing test prompts. The majority of test takers rated the writing test prompts as “Fairly easy”, with only one writing test prompt perceived as quite challenging. Although the comparison between the human-generated writing test prompts and AI-generated writing test prompts

revealed that the AI-generated ones were generally perceived as more difficult, this perception does not necessarily indicate inappropriateness, particularly when accompanied by positive sentiment (see below).

Interestingly, while 50% of the writing test prompts were commonly believed to be created by AI, 14 test-takers believed that none of the writing test prompts were created using AI, suggesting their uncertainty about AI's role in content generation. Despite this, test-takers had mainly a positive sentiment towards the writing test prompts.

These findings contribute to the validation of the domain definition inference, supporting the assumption that AI-generated prompts can be designed to align with the characteristics of the CILS B2 written interaction and production construct (RQ1) and are perceived as appropriate for eliciting B2-level writing skills (RQ2). However, full validation of the warrant requires further evidence, as this study does not yet evaluate test-taker written performances on AI-generated writing test prompts, as previously noted. The next stage of this research will focus on the evaluation inference (Chapelle & Voss, 2021), analyzing whether the written responses produced on AI-generated writing test prompts demonstrate the expected linguistic and strategic competencies aligned with CEFR B2 descriptors.

Conclusion

This study examined ChatGPT-4's capability to generate writing test prompts for the CILS B2 exam, providing backing for the domain definition inference. The findings suggest that AI-generated prompts can serve as viable alternatives to human-created ones, particularly in terms of clarity, thematic relevance, and perceived difficulty. However, refinements are needed in language complexity and adherence to CILS format, underscoring the continued need for human oversight in AI-assisted test development.

To further validate these findings, the next phase of this research will analyze test-taker written performances on AI- and human-generated prompts, offering additional backing for the domain definition inference while also contributing to the evaluation inference (Chapelle & Voss, 2021). This step will assess whether AI-generated prompts elicit responses that align with CEFR B2 descriptors.

Thus, this study contributes both practically and theoretically to AI-assisted language assessment. Practically, it explores how AI can enhance test development efficiency by reducing the time and effort required for writing test prompts. Theoretically, it advances understanding of the validity and appropriateness of AI-generated prompts in high-stakes language testing, particularly within S/FL Italian. These insights serve as a foundation for future research on AI's role in language test design, reinforcing the importance of validation frameworks in AI-driven assessment.

References

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*, Article 903077. https://doi.org/10.3389/frai.2022.903077

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–217). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410602145-10

Centro CILS. (2009). *Linee guida CILS*. Guerra Edizioni. https://cils.unistrasi.it/public/articoli/12/linee_guida_cils_pdf(2).pdf

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, *29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Chapelle, C., & Voss, E. (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press. https://doi.org/10.1017/9781108669849

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. Routledge.

Cinganotto, L., Sbardella, T., & Montanucci, G. (2024). Dagli algoritmi alle competenze linguistiche: Il ruolo dell'intelligenza artificiale nell'educazione linguistica online. *The Journal of Language and Teaching Technology, 6*, 1-12. https://italian.rutgers.edu/images/PDFs/6.Cinganotto.Sbardella.Montanucci_JLTT_2024.pdf

Council of Europe (CoE). (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/1680459f97

Gallina, F., Machetti, S., & Masillo, P. (2019). Il valutatore delle competenze linguistico-comunicative: Riflessioni e proposte per una figura professionale. In C. Bagna & V. Carbonara (Eds.), *Le lingue dei centri linguistici nelle sfide europee e internazionali: Formazione e mercato del lavoro* (pp. 67–83). Edizioni ETS.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Jun, H. (2021). Justifying the interpretation and use of an ESL writing final examination. In C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 235–263). Cambridge University Press. https://doi.org/10.1017/9781108669849.014

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35(4)*, 477–499. https://doi.org/10.1177/0265532217716735

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, A. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Lee, E. (2023). L2 students' views on writing tools: Investigating domain definition within an argument-based validation framework. *English Teaching*, *78*(1), 125–144. https://doi.org/10.15858/engtea.78.1.202303.125

Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, *211*, Article 104977. https://doi.org/10.1016/j.compedu.2023.104977

Machetti, S. (2022). Le certificazioni di lingua italiana nei panorami linguistici globali: Spunti per un'analisi quantitativa. *Studi Italiani di Linguistica Teorica e Applicata, 51*(2), 452–476. https://perma.cc/8VH2-9KWY

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://openai.com

Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for high- (but not low-) resource languages. In P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), *Proceedings of the Eighth Conference on Machine Translation* (pp. 392–418). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.40

Rossi, O. (2023, May). *Using AI for test item generation: Opportunities and challenges* [Webinar]. EALTA Webinar Series. https://perma.cc/33N4-8QST

Van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341–358). De Gruyter Mouton. https://doi.org/10.1515/9781614513827-023

Voss, E. (2024). Artificial intelligence in language assessment. In A. Kunnan (Ed.), *The concise companion to language assessment* (2nd ed., pp. 112–125). Wiley-Blackwell.

Voss, E., Cushing, S. T., Ockey, G. J., & Yan, X. (2023). The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and

practice. *Language Assessment Quarterly, 20*(4–5), 520–532. https://doi.org/10.1080/15434303.2023.2288256

Xi, X. (2023). Advancing language assessment with AI and ML–Leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly, 20*(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488

Appendix A

Table A1

CILS B2 Exam Level Description (CILS Guidelines, 2009, p. 29)

Competence profile	It is the level that certifies full autonomy in communicative competence in Italian as a foreign language. The test-taker is able to communicate effectively during a stay in Italy for study purposes and to manage interactions with the Italian language and culture, even for work-related reasons. Interaction with native speakers develops without excessive effort or tension. A test-taker at this level of competence can understand the main ideas of even complex texts that relate both to everyday life and to more abstract concepts. Oral and written production is communicatively effective, even if it contains some errors. This is the minimum level of competence required for access to the Italian university system, for completing a study cycle within a short-term mobility program for students, as well as for teachers and researchers. It is also necessary for benefiting from scholarships awarded by the Italian state and for undertaking an internship as part of a diploma course or within companies.
Syllabus of Morphosyntactic Structures	In addition to the structures covered in the previous levels, test-takers are required to understand and manage the following structures of the Italian language: Allocutive pronouns Indefinite pronouns and adjectives Combined pronouns Pronominal particles Conjugation of the active and reflexive forms of regular and irregular verbs, as well as modal verbs, in the following moods and tenses: • Present indicative • Present perfect indicative (passato prossimo) • Imperfect indicative • Past absolute indicative (passato remoto) • Pluperfect indicative (trapassato prossimo) • Simple and anterior future indicative • Present conditional • Past conditional • Present and imperfect subjunctive • Present and past infinitive • Imperative Passive form (only recognition) Impersonal verbs Most common adverbs of judgment and doubt. Simple sentence structure: Volitional clauses using the subjunctive, indicative, and infinitive.

	Complex sentence structure: Coordinated disjunctive, conclusive, and correlative clauses. Complex sentence structure: Subordinate clauses including subject, final, comparative, real-conditional, explicit concessive, explicit consecutive, and implicit temporal clauses.
Pragmatics and Language Use	The test-taker is able to express themselves confidently, clearly, and politely in either a formal or informal register, depending on the situation and the person involved. They can appropriately participate in discussions, effectively managing turns of speech. The test-taker is familiar with the social and communicative norms and rules typical of various everyday situations.
Vocabulary	The test-taker is able to navigate texts containing words from the Basic Vocabulary of the Italian language, as well as common vocabulary, up to a maximum of 7%. In oral and written production, they can use words from the core lexicon and also part of the high-frequency vocabulary.

Table A2

CILS DUE-B2 Written Production Component (CILS Guidelines, 2009, p. 31)

Profile	The test-taker is able to produce written texts using simple structures while conveying information clearly and effectively from a communicative perspective on familiar topics or subjects of personal interests (see also CEFR B2 writing production and interaction descriptors).
No. of tasks	2 Task 1: Descriptive/Narrative text Task 2: Structured text (email/letter)
Task types	Description of familiar people or places, storytelling involving given characters, diary-style narration of a trip, an event, or a particular episode; informal letter to relatives or friends to express gratitude, request information, or share experiences while highlighting key points.
Text length	Task 1: 100 to 120 words. Task 2: 50 to 80 words.
Test duration	1 hour and 10 minutes

Appendix B

Table B1

Administered Writing Test Tasks

Type of generation	Task 1	Task 2
Human	(Ita.) L'amicizia è una componente importante nella vita di tutti. Che cosa ne pensi? Parla delle tue esperienze positive e negative in proposito. (Eng.) Friendship is an important aspect of everyone's life. What do you think? Discuss your positive and negative experiences on the topic.	(Ita.) Sei indeciso se fare una vacanza-studio all'estero. Scrivi un'email a un tuo professore dove: - spieghi dove, quando e che tipo di esperienza vorresti fare; - chiedi la sua opinione ed eventuali suggerimenti. (Eng.) You are unsure about whether to take a study vacation abroad. Write an email to one of your professors where you: - Explain where, when, and what kind of experience you would like to have; - Ask for their opinion and any possible suggestions.
AI	(Ita.) Rifletti sull'importanza dello sport e del gioco di squadra nello sviluppo personale. Hai esperienze personali che evidenziano questi aspetti? (Eng.) Reflect on the importance of sports and teamwork in personal development. Do you have any personal experiences that highlight these aspects?	(Ita.) Vuoi migliorare le tue competenze linguistiche e decidi di iscriverti ad un corso di livello avanzato. Scrivi un'email ad una scuola di lingue e chiedi informazioni sul corso, sui requisiti di ammissione e sulla data di inizio. (Eng.) You want to improve your language skills and decide to enroll in an advanced-level course. Write an email to a language school asking for information about the course, admission requirements, and the start date.

About the Authors

Sabrina Machetti is the Director of the CILS (Certification of Italian as a Foreign Language) Centre, Associate Professor of Educational Linguistics. Her main research interests are focused on language testing and assessment, learning and teaching Italian as a foreign/second language, and language policies.

Giulia Peri earned her Ph.D. in Applied Linguistics at the University for Foreigners of Siena. Her research interests focus on S/FL Italian teaching and learning, language testing and assessment and technology. She is a Research Fellow at the CILS Centre (Certification of Italian as a Foreign Language) and the Project Coordinator for the L2 Italian scenario-based assessment project at the University for Foreigners of Siena.

Paola Masillo is Ph.D. in "Linguistics and Teaching Italian as second language", University for Foreigners of Siena. Her main research focuses on learning Italian as a foreign/second language, language assessment and language policies. She is currently working in the technical-scientific and data processing area at the CILS Centre of the University for Foreigners of Siena.

Author Note

All authors contributed equally to the conceptualization of the research. Giulia Peri was responsible for writing the Methods and Results sections. Sabrina Machetti wrote the Introduction. Paola Masillo wrote the Conclusions. The authors declare no conflicts of interest.