ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-97
ETS Research Report No. RR–21-27

Designing Efficient L2 Writing Assessment Tasks for Low-Proficiency Learners of English

Shoko Sasayama

Pablo Garcia Gomez

John M. Norris

December 2021

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL*® *Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org **Web site:** www.ets.org/toefl

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

RESEARCH REPORT

Designing Efficient L2 Writing Assessment Tasks for Low-Proficiency Learners of English

Shoko Sasayama, Pablo Garcia Gomez, & John M. Norris

ETS, Princeton, NJ

This report describes the development of efficient second language (L2) writing assessment tasks designed specifically for low-proficiency learners of English to be included in the *TOEFL® Essentials™* test. Based on the can-do descriptors of the Common European Framework of Reference for Languages for the A1 through B1 levels of proficiency, four task types were identified to be prototypical candidate writing tasks for the target test-taker population (i.e., adolescent and adult low-proficiency English learners). Those four task types included: (a) Describe a Photo, (b) Write a Review, (c) Chat With a Friend, and (d) Write an E-mail. These task types were also considered efficient in the framework of the test in that they had the potential to be accessible to low-proficiency learners and to elicit sufficient spontaneous writing for assessment purposes within a short period of time. In the current study, eight assessment tasks, two for each task type, were developed and piloted with 169 A1–B1 learners of English from Japan and Colombia. The findings revealed that the Describe a Photo and Write an E-mail tasks performed the best in eliciting substantial language use and emphasizing distinct performance attributes, both characteristics needed for efficiently measuring test takers' writing proficiency as well as discriminating among proficiency levels at the lower end of the spectrum. The report concludes by highlighting some observations on L2 writing assessment task design for low-proficiency learners of English.

Keywords Assessment task design; efficient assessment tasks; L2 writing assessment; low-proficiency learners; *TOEFL® Essentials™*

doi:10.1002/ets2.12341

This report describes the development of second language (L2) writing assessment tasks for use with low-proficiency learners of English. The development of these tasks was part of the larger project to design an efficient assessment of English language proficiency, the *TOEFL® Essentials™* test. Following a brief review of relevant literature, we introduce salient details regarding the assessment context and purpose as well as the overall approach to designing L2 writing tasks for low-proficiency learners. We then describe the study methodology, report and discuss results, and offer several concluding observations for the design of tasks intended to elicit meaningful communicative performances from low-proficiency learners.

Study Background and Assessment Context

Assessing Second Language Writing in English Language Proficiency Assessment

Large-scale standardized English language proficiency assessment is intended to provide various decision makers with useful information about test takers' abilities, including their ability to communicate effectively on relevant tasks. When assessment is used for high-stakes decision-making purposes, such as university admissions or immigration, it is imperative that the test be designed such that interpretations about the test takers' abilities can be trusted (Xi & Norris, 2021). Such assessments typically provide an indication of test takers' abilities overall as well as being broken down by the four language skills of listening, reading, writing, and speaking; these skills are associated with identifiable task types likely to be encountered by learners (e.g., Chapelle et al., 2008). Assessing each of the four skills is critical in that it offers score users a window into English learners' relative strengths and weaknesses in using English for different purposes, and it is particularly important when learners present unbalanced profiles of proficiency (e.g., high literacy, low oracy skills).

Assessing English L2 *writing* abilities is crucial, especially in circumstances where learners need to demonstrate a sufficient level of writing ability for certain purposes (e.g., readiness for university study) or where identifying writing

Corresponding author: S. Sasayama, E-mail: sasayama@ets.org

development needs for a particular group of learners is a priority. In general, writing proficiency is multifaceted and requires a variety of knowledge and skills, including knowledge of L2 vocabulary and morphosyntax, as well as foundational skills such as spelling and punctuation, composing (e.g., drafting, revising), and macroskills required to understand and respond to genre-specific expectations (Cumming et al., 2021). Correspondingly, there are a variety of possibilities for assessing L2 writing ability, ranging from assessing discrete components of knowledge that are useful in writing (e.g., testing vocabulary by identifying synonyms, testing syntax by building sentences) to other, more performance-based assessments where the test takers are put into a context and asked to use the language for communication purposes by creating some kind of written text (e.g., write an academic discussion post to express an opinion on a topic). Although more discrete assessments may give us useful information about some aspects of the test taker's writing ability, they clearly cannot show us how well a learner can engage in the writing process holistically. In order to get at a learner's ability to write communicatively, we need to put them in a context and have them write something to a specified audience with a specific reason in mind (Cumming et al., 2021; Mislevy et al., 2002).

To this end, a writing *task* or a communication-oriented writing activity can serve as an optimal vehicle. Through the use of writing tasks, we can elicit language use that allows us to measure L2 learners' ability to use the language for communication purposes and to accomplish a task identified to be relevant to and important for them (Long & Norris, 2000). When designed appropriately, a task can replicate a real-world scenario with a specific audience and reason to communicate, and elicit language use that is expected in the real world. Performance elicited by such a task can tell us a lot about what test takers can actually *do* in the target language when put in a similar real-world context.

In designing L2 writing assessment tasks, then, it is important to first identify on which real-world tasks or task types the test takers will need to be able to demonstrate their abilities, from the assessment users' perspectives. Once the target, real-world tasks are identified, item specifications will need to be developed for a given task type. These item specifications define a general description of the task (i.e., what the task is and what it is supposed to test), prompt attributes (e.g., how performance is elicited, task instructions and specifications about prompts and other input such as visuals, questions to be answered), and response attributes (i.e., what expected responses are, such as to write a description of a picture). In order to get at L2 learners' writing ability, the next step is to identify attributes of task performance to be assessed and specify how the responses will be evaluated. In doing so, for task-based assessment, it is particularly important that the scoring criteria reflect the real-world criteria by which the performance is evaluated and that levels of success are clearly defined. Evaluation of the tasks is also a critical part of the development of task-based assessment. This evaluation process will help us ensure the effectiveness, appropriateness, and efficiency of the assessment tasks. It will also help confirm whether they are accomplishing what they were designed to accomplish, including what exactly gets elicited, to what extent task performance can be rated reliably, to what extent the task can discriminate among learners with different proficiency levels, and so on.

Low-Proficiency Learners and Literacy/Writing Skills

For low-proficiency English learners, what are the target, real-world writing tasks that they need to be able to do and demonstrate their abilities on? Although different learners will have distinct needs depending on their particular societal, vocational, or educational settings and personal goals, a good source for general characteristics of learners at different proficiency levels is the Common European Frame of Reference for Languages (CEFR; Council of Europe, 2001). According to the CEFR, beginning at the lowest levels of "Basic User" (A1–A2), some key dimensions of writing proficiency for A1 learners include (a) writing about themselves and imaginary people in simple phrases and sentences, (b) describing familiar everyday objects in very simple language, and (c) composing a short message to a friend to offer or ask for personal information (Council of Europe, 2020). A2 writers are expected to be able to do quite a bit more, especially those who are higher in proficiency within the A2 level (A2+). Some key can-do descriptors for A2 and A2+ writers include (a) writing a diary entry about activities, people, and places in connected sentences, (b) giving opinions about topics of their interest, (c) composing an e-mail to perform some sort of action (e.g., send an invitation), and (d) exchanging information by an e-mail or short letter to answer questions from their interlocutor (Council of Europe, 2020). In transitioning to "Independent Users," B1 writers are expected to be able to "produce … connected texts on a range of familiar subjects … by linking a series of shorter discrete elements into a linear sequence" (Council of Europe, 2020, p. 66). More specifically, they should be able to not only (a) describe a real or imaginary event, but (b) do so by giving accounts of feelings and reactions experienced (in the form of a personal letter), as well as to (c) write a very short report in a conversational format to provide information or to justify an action. These characterizations of what low-proficiency writers can do provide a useful initial domain of task types from which to select and further develop candidate test tasks.

With these expectations in mind, particularly for learners at the A1–A2 levels, how should low-proficiency writers' performances on relevant tasks be evaluated? On the one hand, in understanding writing performances of test takers with low L2 literacy skills, several underlying characteristics or indicators of writing proficiency should be taken into account, including (a) text length, (b) lexical diversity, and (c) syntactic complexity. Research has consistently shown that the length of a written text produced by L2 learners increases as their writing proficiency increases (e.g., Crossley & McNamara, 2010; Ferris, 1994; Grant & Ginther, 2000). This research finding is also in line with expectations seen in the CEFR can-do descriptors for A1- versus A2- versus B1-level learners: As they move up along the proficiency spectrum, learners are expected to produce more and more substantial amounts of writing, in particular by combining a series of sentences (A2) or a series of elements into a logical sequence (B1).

Lexical diversity—a type of lexical richness that refers to the range and variety of words used in a given response—is another aspect of task performance that researchers have looked into as a potential indicator of writing proficiency. According to Tack et al. (2017), lexical features—in particular lexical diversity measures—found in short answers of between 30 and 200 words were the most informative predictors (compared with other measures, including syntactic, discursive, and readability features) of English L2 writing proficiency, and they were able to distinguish among A1, A2, B1, B2, and C levels of proficiency on the CEFR. This finding was also confirmed in a study by Crossley et al. (2011) in which lexical density (*M*) was found to be one of the best predictors of English L2 writing proficiency among a variety of lexical measures.

It has also been a common practice for researchers to examine L2 English writers' task performance in terms of syntax, and in particular syntactic complexity. Syntactic complexity can be characterized as "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003, p. 492). Although some syntactic complexity measures (e.g., clauses per T-unit) demonstrate a curvilinear relationship with L2 proficiency, the mean length of T-unit (MLT) has been consistently shown to indicate linear development as the participants' L2 proficiency increases (e.g., Norris & Ortega, 2009). This trend has been observed at the lower end of the proficiency spectrum as well. Jiang et al. (2019), for example, analyzed beginner- and intermediate-level learners' written narrative responses and found that the MLT, mean length of sentence, and dependent clauses per clause were the best measures of syntactic complexity for predicting writing proficiency. Similarly, Khushik and Huhta (2020) found that the length measures, including MLT, were among the indicators of proficiency that were best able to distinguish among A1, A2, and B1 levels of writing proficiency.

On the other hand, complementary to direct indicators of L2 writing, holistic evaluation of writing performance provides important insights into other dimensions of ability. To what extent learners can combine their linguistic resources in a way that is appropriate to the writing genre, address the topic or content of the task, and complete or accomplish the expectations of the task are all phenomena best suited to examination by human raters (e.g., Cumming et al., 2021). Raters, with the aid of rubrics that provide rich descriptions of the key features of expected performances on different tasks at different levels of accomplishment, play an important role in their ability to perceive the complex interplay of such phenomena and render reliable judgments about holistic proficiency distinctions.

Design of Writing Assessment Tasks Tailored for Low-Proficiency Learners

With this background in mind, we set out to develop and investigate several candidate writing assessment tasks intended to provide insights into low-proficiency L2 learners' writing abilities. In this section, we will first introduce our assessment context and the TOEFL Essentials test. We will then describe in depth the design of the four prototypical writing tasks developed specifically for low-proficiency learners of English as part of the TOEFL Essentials test.

Assessment Context: TOEFL Essentials

TOEFL Essentials is a new English language proficiency test in the TOEFL Family of Assessments at ETS. The test covers the four language skills of listening, reading, writing, and speaking, and it is intended for adolescents and adults who wish to provide evidence of their overall English language proficiency level in academic and daily life contexts. Several unique aspects of this test, implemented in response to the assessment mandate, played an important role in determining the types of writing tasks included in the test (see Papageorgiou et al., 2021). First, the tasks needed to be designed such that a full range of proficiency levels could be assessed, including the lower end of the proficiency spectrum, or the range of A1 to B1 on the CEFR. In order to efficiently assess learners across this broad proficiency range, the TOEFL Essentials test also adopted a multistage adaptive testing (MST) methodology to help ensure the most accurate measure of the test takers' language ability. Test-taker performance on the first part of a test section (i.e., the router) determines the content for the second part of the section, where communication-oriented tasks are introduced, so that the difficulty of the assessment tasks matches the ability level of the test takers. Content in the second stage of the writing section is classified into two difficulty levels, "low" or "medium and high."

Second, a balance needed to be struck between efficiency in testing time and the desire to assess learners' ability to communicate meaningfully in English. Although efficient, discrete assessment items, such as Build a Sentence or Listen and Repeat (see Papageorgiou et al., 2021), are consistently shown by research to be highly predictive indicators of test takers' overall proficiency (Davis & Norris, 2021), those items do not allow us to directly observe the extent to which test takers can actually use the language for communication purposes. As reviewed earlier, communication tasks are best suited to assess what test takers can actually do in the target language, as they provide meaningful opportunities to demonstrate their abilities to communicate in English (see Norris, 2018b). Through the use of MST and a combination of distinct item types, the TOEFL Essentials test combines the strengths of both efficient and task-based assessment approaches to achieve both breadth and depth in the measurement of L2 English proficiency. Additionally, to maximize test-taking efficiency, the communication-oriented tasks were also targeted to be completed in a relatively short period of time.

Designing L2 Writing Assessment Tasks for Low-Proficiency Learners of English

With this assessment context in mind, we embarked on the design of communication-oriented writing assessment tasks specifically tailored for low-proficiency learners of English to be included in the low panel of the writing section of the TOEFL Essentials test.

Key Design Goals

Given the assessment context, the key design goals for writing assessment tasks included the following:

Nature of tasks:

- Tasks should replicate target, real-world tasks to gauge test-taker ability to use the language for communication purposes.
- Tasks should be engaging, age appropriate, and presented with a clear communicative purpose.
- Tasks should allow test takers to demonstrate their writing skills in a relatively short period of time to achieve efficiency.

Target proficiency level:

- Tasks should be designed to be accessible for low-proficiency learners within the A1 to B1 range of proficiency on the CEFR.
- Tasks should require limited reading processing, appropriate for low-proficiency learners, to understand the directions and stimulus.
- Tasks should be designed to elicit sufficient language use to distinguish among different levels of proficiency within the lower end of the proficiency spectrum.

Task development quality expectations:

- The stimulus materials provided in the tasks should include both visual (e.g., images, tables) and textual information.
- Tasks should be easily standardized and replicated using a wide range of themes as well as visual and textual information.
- Tasks should elicit enough multisentence text that can be evaluated both by human raters and automated scoring engines.

Task Design

As noted elsewhere, the "TOEFL Essentials test focuses on predicting overall English ability and discerning the likelihood that learners can accomplish real-life English communication tasks" in academic and general life contexts (Papageorgiou et al., 2021, p. 5). That is, although the test features authentic communication tasks, including writing tasks, these tasks are not sampled intentionally from a specific domain or target-language use analysis. Rather, the tasks are intended to provide prototypical opportunities for test takers to show what they can do in English. The CEFR, then, provides an ideal source for defining both what learners should be able to do at distinct levels of proficiency as well as the kinds of representative tasks that might be used to elicit corresponding performances. In order to identify potential writing tasks that would be appropriate for lower-proficiency learners, we first identified target task types that would enable them to demonstrate their abilities by reviewing the can-do descriptors of the CEFR for the A1, A2, and B1 levels of writing proficiency (Council of Europe, 2020). Recapping from above, A1 writers are expected to be able to describe familiar everyday objects and convey and communicate about personal information in writing, using simple sentences that are not necessarily connected to each other. In contrast, A2 writers are expected to be able to write about activities, people, and places in connected sentences, provide opinions or exchange information about familiar (not just personal) topics, and write an e-mail to perform social functions. B1 writers, then, are expected to be able to handle similar task types but in a more sophisticated way, by incorporating feelings and reactions when communicating about their (imaginary) experiences and justifying an action. At this level of proficiency, writers are also expected to be able to produce more substantial language, connecting discrete elements into a linear, logical sequence. Given these expectations, our goal was to develop tasks that would provide test takers with an opportunity to demonstrate their abilities to (a) describe a picture that depicts familiar scenes; (b) provide information about themselves, familiar objects, or experiences; (c) state an opinion about some familiar services, places, or products; and (d) write an e-mail to perform some social function (e.g., an invitation).

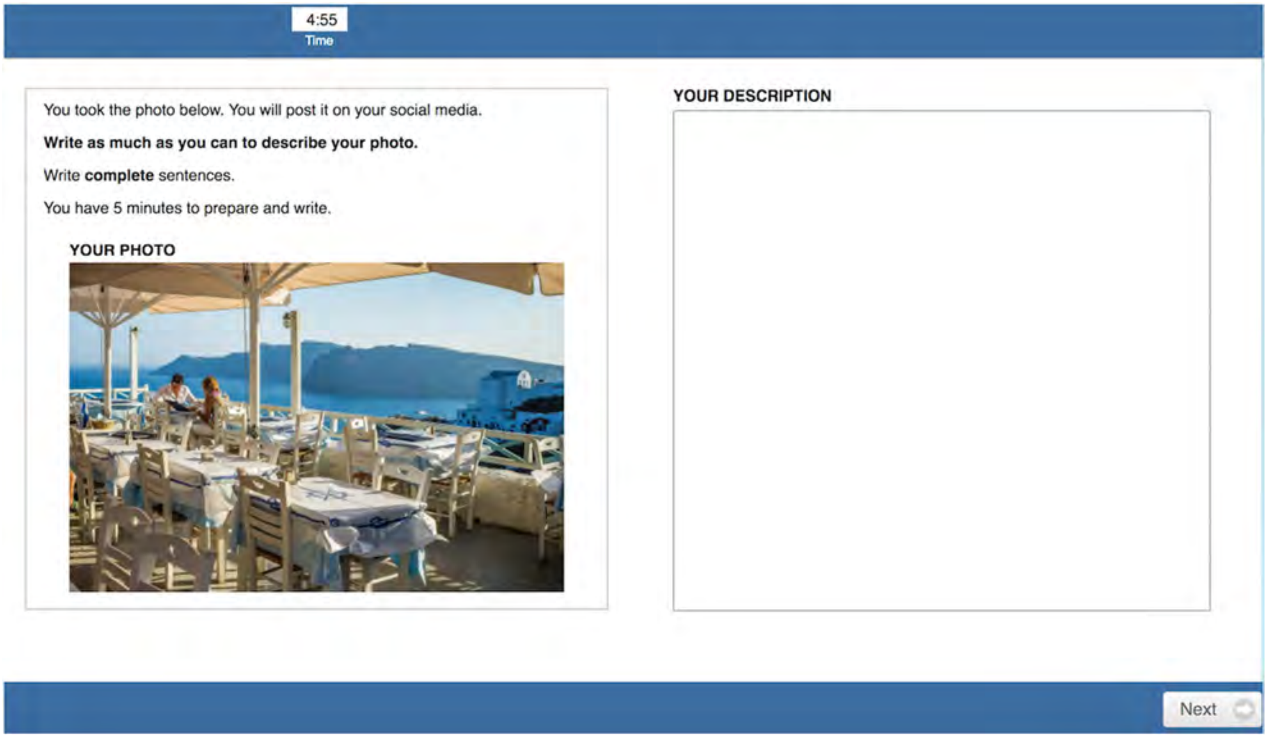
After an initial set of early prototypes, four task types that were thought to elicit the expected language use described above were selected. These task types were Describe a Photo, Write a Review, Chat With a Friend, and Write an E-mail. In what follows, we describe each of the selected task types and their different versions in detail.

Describe a Photo

The Describe a Photo task asks test takers to write a social media post about a photo provided as part of the task input. They are asked to describe the photo to their social media audience, and they are given 5 minutes to read the instructions, prepare, and write their post. In order to investigate the role of choice in this task type, two versions of the same task were developed. As seen in Figure 1, in Form A, test takers are given no choice of a picture and are asked to write about a picture that (in this prototype task) depicts a scene at a restaurant by the ocean where two people are dining—something that was thought to be familiar to many of our target test-taker population (i.e., adolescent and adult low-proficiency learners of English).

In Form B, on the other hand, test takers are given a choice of pictures to describe. The task first presents test takers with three pictures to choose from (see Figure 2). Here, in addition to the picture that depicts two people dining at a restaurant as seen in Form A, two other pictures are presented: (a) a picture of a bookstore and people reading and (b) a picture of a café where two people are working. Those scenes and the related experiences that those pictures likely evoke are also thought to be familiar and personal to the target test takers. Once they choose a picture to describe, test takers then see a screen with a prompt presented with the chosen picture underneath (see Figure 3).

Even though picture description tasks or picture-based narrative tasks are not uncommon in language assessments, our Describe a Photo task is unique in that it is contextualized in a social media environment. This contextualization allows us to provide a real-world scenario for the task, including its intended audience, and make its communicative purpose clear and concrete. The task is deemed appropriate in terms of its difficulty because it is thought to pose a low reading load (not much prompt or input to process). Also, the task clearly targets dimensions of the CEFR at the A1 and A2 levels: namely, describing familiar objects in simple language and composing brief messages about personal and familiar information.



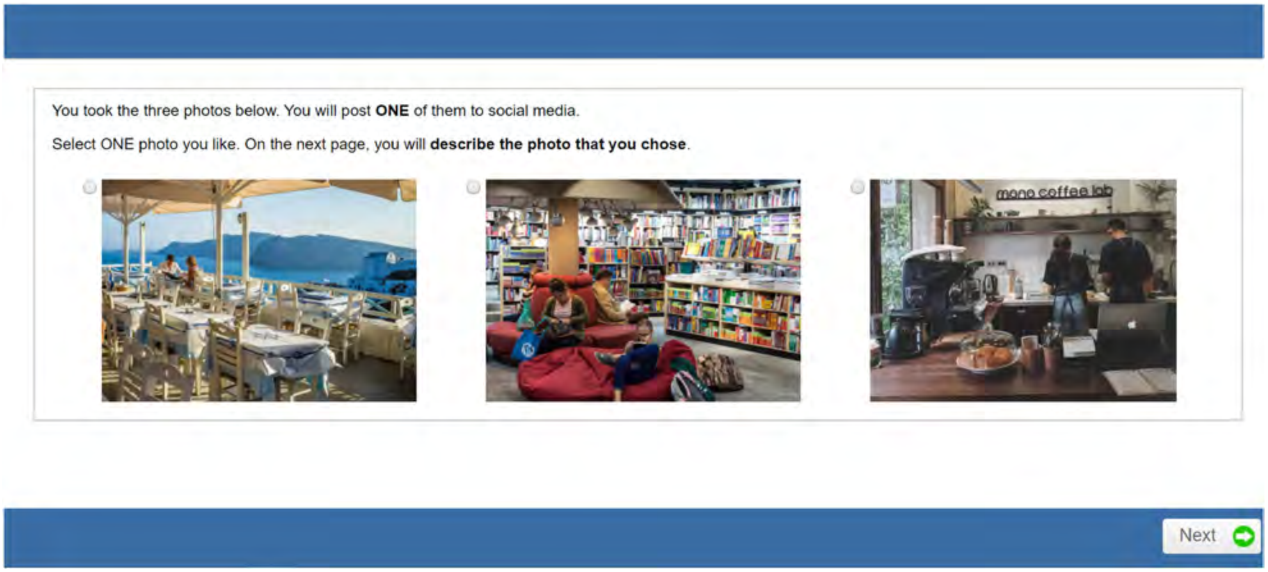
You took the photo below. You will post it on your social media.

Write as much as you can to describe your photo.

Write **complete** sentences.

You have 5 minutes to prepare and write.

Figure 1 Form A of the Describe a Photo task.



You took the three photos below. You will post **ONE** of them to social media.

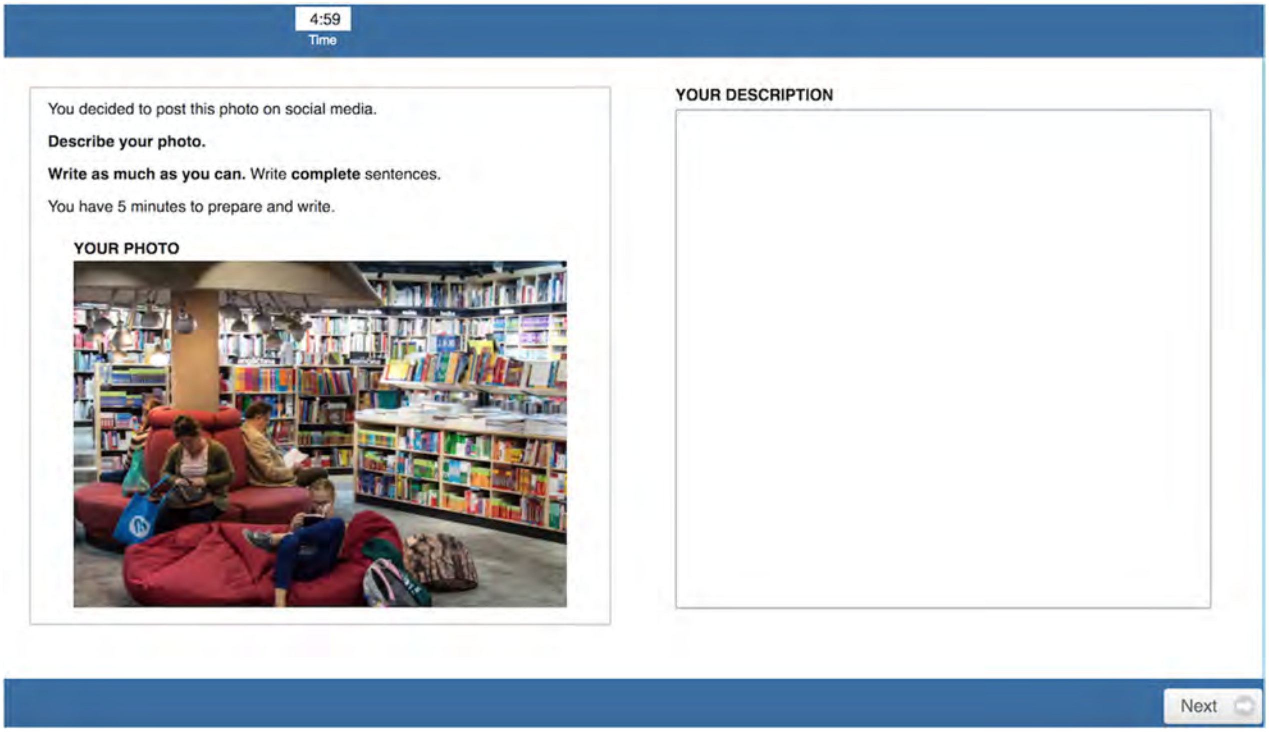
Select ONE photo you like. On the next page, you will **describe the photo that you chose**.

○ ○ ○

Figure 2 Form B of the Describe a Photo task: Screen 1.

Write a Review

The Write a Review task asks test takers to respond to a friend's question about a service or a product posted on social media. Information about the service or product in question, together with a relevant image, is provided as part of the task input, and test takers need to use all information presented in order to successfully answer the question. Two versions of the Write a Review task were created to look into whether the subject of a review has any effect on performance. Form A asks test takers to answer a question about a café, Form B asks them to answer a question about a tablet, and both forms instruct them to provide an opinion about what they think of the service or product, using the information given



You decided to post this photo on social media.

Describe your photo.

Write as much as you can. Write **complete** sentences.

You have 5 minutes to prepare and write.

Figure 3 Form B of the Describe a Photo task: Screen 2.

(see Figures 4 and 5). Participants are given 5 minutes to read the instructions, plan, and write a response to this Write a Review task.

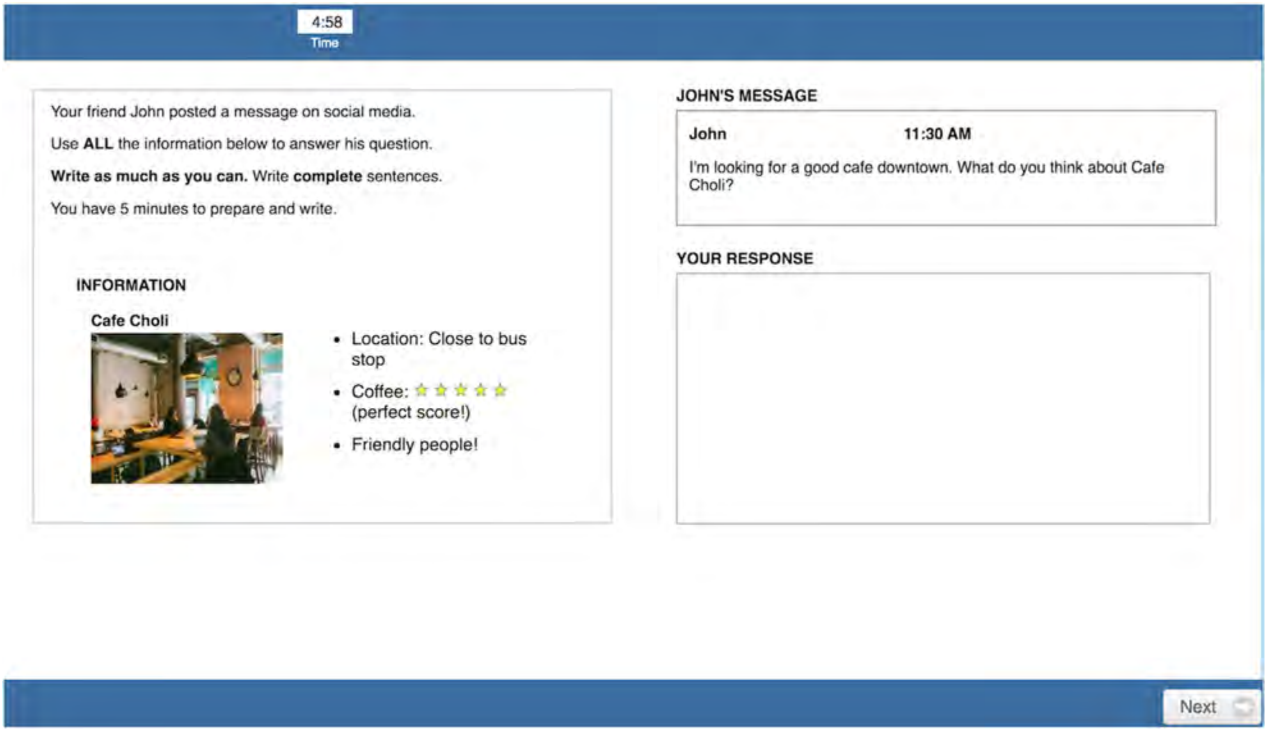
For this task type, not having to come up with content for a response, or *what* to say, eases the overall cognitive demand and helps test takers focus more attentional resources on their language use, or *how* to say it (Sasayama, 2015). Given the target test-taker population (i.e., low-proficiency learners of English), it was deemed important to provide a list of things in the input that they can write about, in the form of information about the service or product in question. Here, though, caution was exercised in providing only minimal linguistic scaffolding: The task includes telegraphic language, rather than complete sentences, that test takers are expected to integrate and manipulate in their responses to successfully accomplish the task.

It is also worth highlighting that this task is embedded in a social media setting, just like the Describe a Photo task, and it provides a real-world communication goal and audience. Unlike the Describe a Photo task, however, the Write a Review task is designed for the range of A2 to B1 learners and includes dimensions typical of those expected at the A2–B1 levels of the CEFR. In particular, connected text is expected in this task to share information and state an opinion about familiar topics.

Chat With a Friend

The Chat With a Friend task asks test takers to respond to a friend's questions about a social media post. In this task, test takers need to respond to questions from a friend in a conversational format about familiar places and activities. Test takers are presented with a picture that they had posted on their social media account with a sentence describing the picture. They also see a brief comment and a series of three simple questions about the post from their friend, which serves as a type of scaffolding to help test takers write about the given topic. Two versions of the Chat With a Friend task were created to explore whether the topic of a chat has any effect on performance. The original post was about a trip near the ocean in Form A (see Figure 6) and about a new bike in Form B (see Figure 7). In both forms, test takers are given 5 minutes to read the instructions, prepare, and write their response to the questions.

Like the other two tasks, this task also makes use of the social media setting to replicate communicative, information-exchange interactions typical of those expected at the lower end of the proficiency spectrum. This task targets the A1



Your friend John posted a message on social media.

Use **ALL** the information below to answer his question.

Write as much as you can. Write **complete** sentences.

You have 5 minutes to prepare and write.

INFORMATION

Cafe Choli

- Location: Close to bus stop
- Coffee: ☆☆☆☆☆ (perfect score!)
- Friendly people!

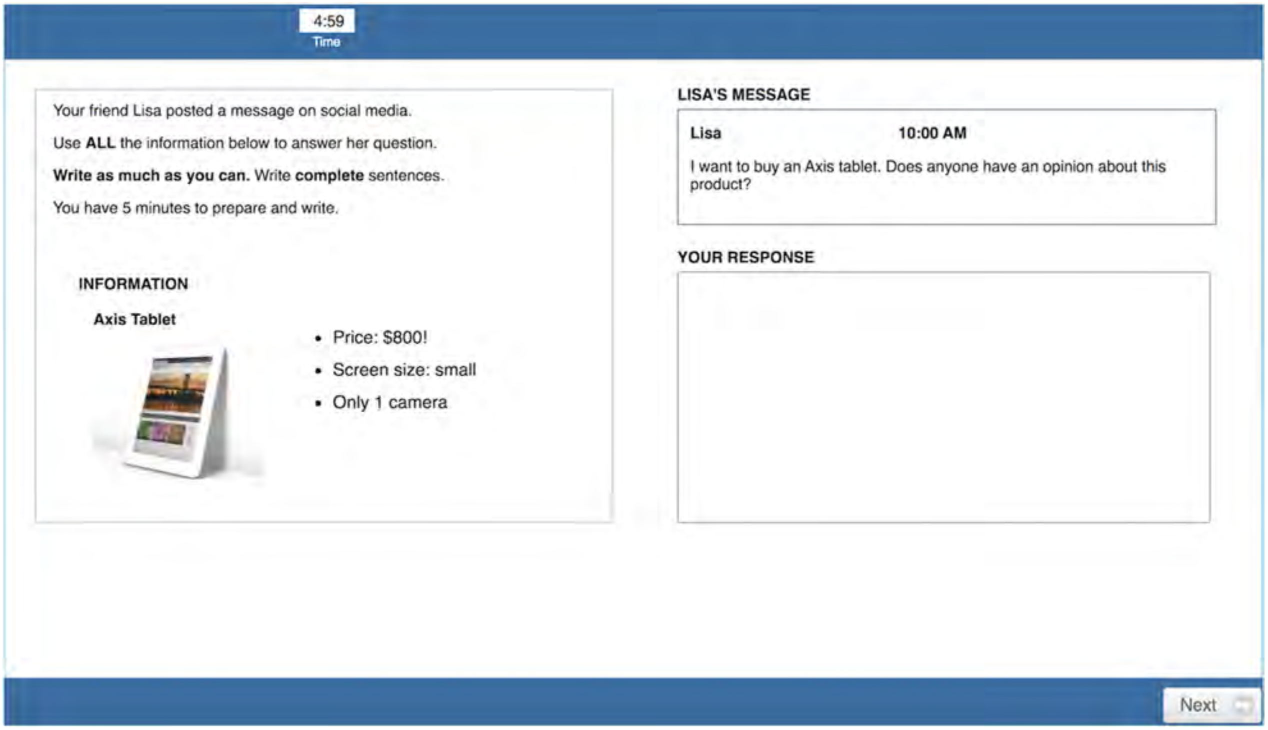
JOHN'S MESSAGE

John **11:30 AM**

I'm looking for a good cafe downtown. What do you think about Cafe Choli?

YOUR RESPONSE

Figure 4 Form A of the Write a Review task.



Your friend Lisa posted a message on social media.

Use **ALL** the information below to answer her question.

Write as much as you can. Write **complete** sentences.

You have 5 minutes to prepare and write.

INFORMATION

Axis Tablet

- Price: $800!
- Screen size: small
- Only 1 camera

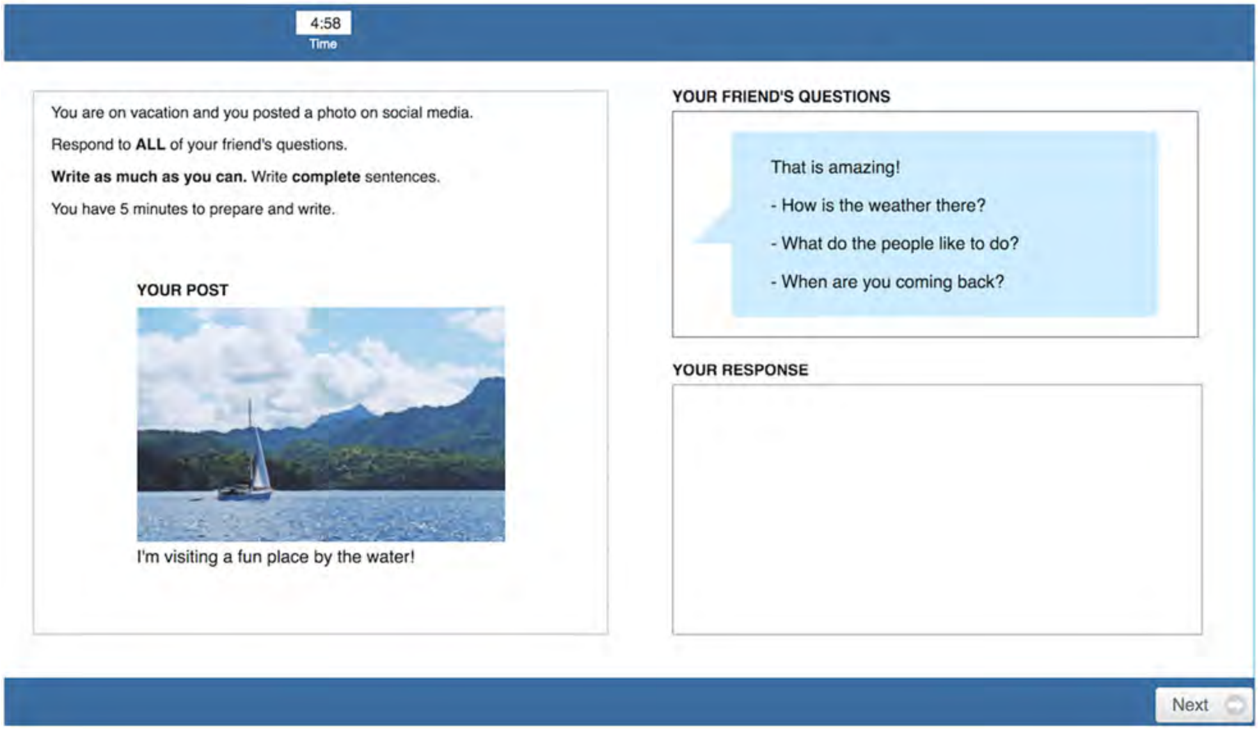
LISA'S MESSAGE

Lisa **10:00 AM**

I want to buy an Axis tablet. Does anyone have an opinion about this product?

YOUR RESPONSE

Figure 5 Form B of the Write a Review task.



You are on vacation and you posted a photo on social media.

Respond to **ALL** of your friend's questions.

Write as much as you can. Write **complete** sentences.

You have 5 minutes to prepare and write.

YOUR POST

I'm visiting a fun place by the water!

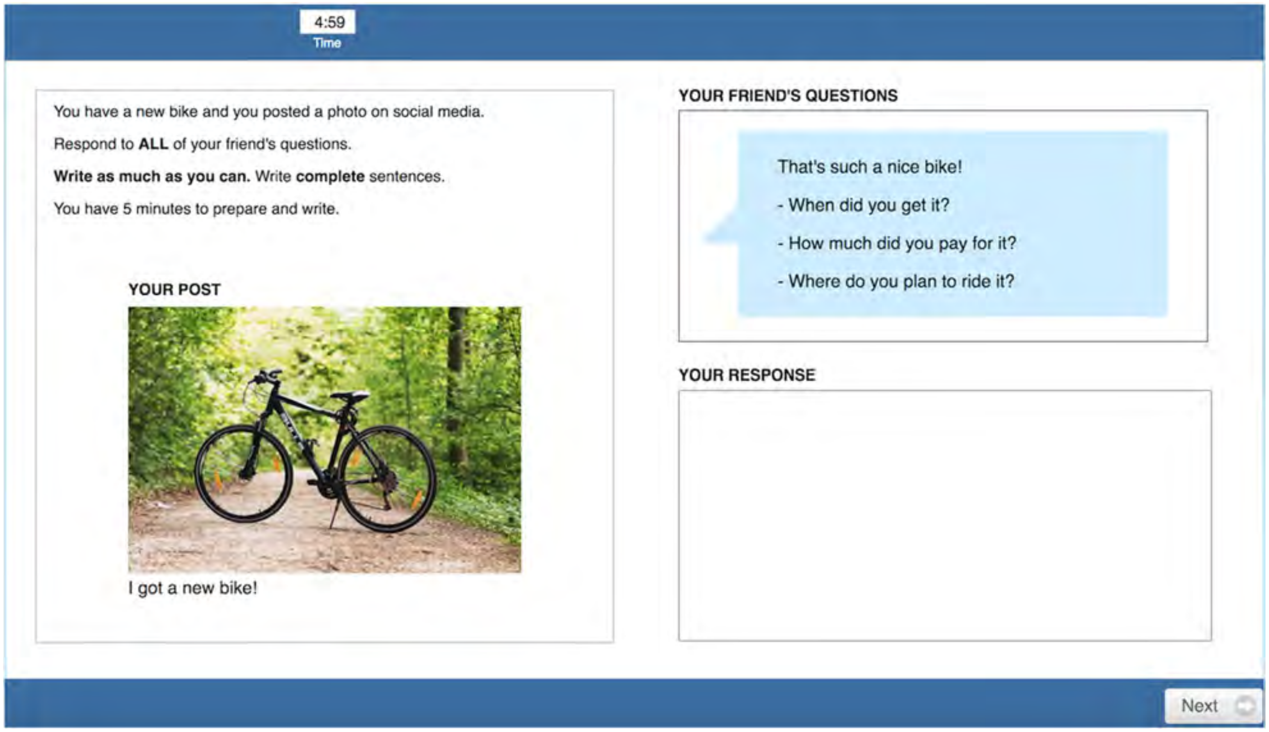
YOUR FRIEND'S QUESTIONS

That is amazing!

- How is the weather there?
- What do the people like to do?
- When are you coming back?

YOUR RESPONSE

Figure 6 Form A of the Chat With a Friend task.



You have a new bike and you posted a photo on social media.

Respond to **ALL** of your friend's questions.

Write as much as you can. Write **complete** sentences.

You have 5 minutes to prepare and write.

YOUR POST

I got a new bike!

YOUR FRIEND'S QUESTIONS

That's such a nice bike!

- When did you get it?
- How much did you pay for it?
- Where do you plan to ride it?

YOUR RESPONSE

Figure 7 Form B of the Chat With a Friend task.

and A2 levels of proficiency on the CEFR, in particular by requiring test takers to write a brief message about familiar, everyday topics.

Write an E-mail

The Write an E-mail task asks test takers to write an e-mail message for a specific communicative purpose, such as extending an invitation. As seen in Figures 8 and 9, in this task, test takers are presented with a scenario where the audience/interlocutor and the purpose of communication are specified. Unlike the other three task types, the task prompt provides a greater degree of scaffolding or guidance about how to respond to the task. In order to investigate whether task requirements given in the form of scaffolding have any effect on performance, two versions were created. Form A asks test takers to (a) choose one movie out of the two presented as part of the task input, (b) invite a friend, Ralf, to the movie, and (c) give him some information about the movie (i.e., name and when it starts). Form B, on the other hand, challenges test takers a bit further and asks them to (a) invite a friend, Daniela, to a movie, (b) give her information about the two movies, and (c) state which movie they prefer and justify their opinion. In both forms, the information on the two movies is presented in a table format to ease construct-irrelevant cognitive demands, like having to make up details of the movies (Sasayama, 2015).

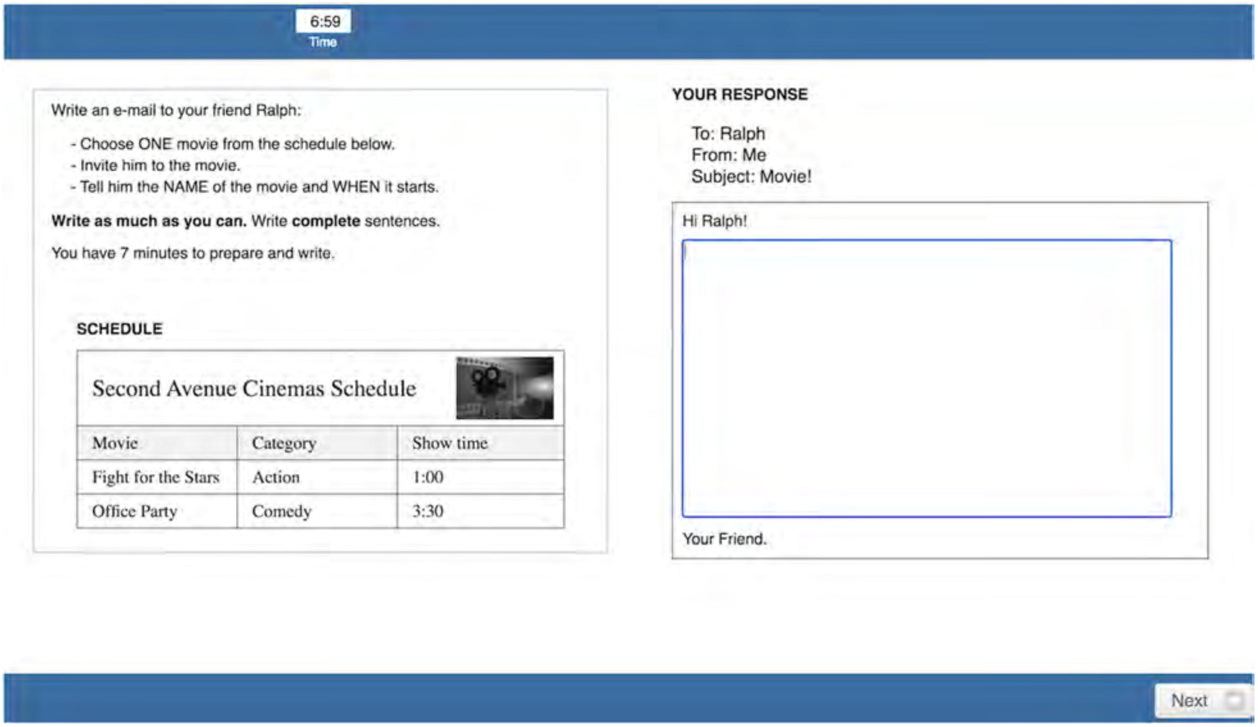
This task includes dimensions typical of those expected at the A2–B1 levels on the CEFR. Test takers are expected to produce a coherent text, connecting discrete elements into a logical sequence to extend an invitation, share information, and in the case of Form B, state and support a preference. Given the more elaborated task prompt to be processed and more extended responses expected, test takers are given 7 minutes to read the instructions, plan, and respond to the Write an E-mail task.

Henceforth, the Describe a Photo tasks will be referred to as Task 1, the Write a Review tasks as Task 2, the Chat With a Friend tasks as Task 3, and the Write an E-mail tasks as Task 4. Table 1 summarizes the characteristics of the four task types.

Development of a Scoring Guide

As the next step in the assessment development process, a scoring guide was designed to gauge the extent to which low-proficiency English learners are able to accomplish the four task types described above. A decision was made to have a single scoring guide for all four task types to enable comparisons across the different tasks in terms of overall task performance and aspects of language delivery.[1] This decision also helped to rule out a priori assumptions about differences among the tasks by placing them on an equal footing when it came to scoring performances. The development of the scoring guide was an iterative process. We started with a version that specified three basic levels of success (not successful, somewhat successful, successful). After reviewing 40 sample responses to each task type, we felt that it was necessary to be able to make more fine-tuned judgments about the test takers' performances, which eventually led to a 6-point scoring guide. It is also noteworthy that the scoring guide was developed with the eventual likelihood of automated scoring in mind, and it therefore entailed linguistic aspects that have been shown by research to be indicators of L2 writing proficiency, including the text length, lexical and grammatical range, and syntactic complexity (see the review above) as well as accuracy.

The resulting 6-point scoring guide was designed to evaluate the quality of the responses holistically across different dimensions, both in terms of overall success in completing the tasks and consistency in language use (see Appendix A). A Score 5 response is characterized as a fully successful response with a great deal of elaboration and consistent facility in language use in terms of a range of vocabulary and grammar as well as accuracy in grammar and mechanics. A Score 4 response is a generally successful response where a good amount of elaboration can be observed, even though language use may lack consistency. A Score 3 response is characterized as a partially successful response, and Score 3 can be awarded in two different ways. It is awarded either to (a) a response that is well elaborated but displays issues in language use (e.g., attempt to make complex sentences can be observed but the response lacks grammatical accuracy) or (b) a response that lacks elaboration but shows good command of the language in terms of a range of vocabulary and grammar as well as accuracy. Responses at Scores 2 and 1 are specified as unsuccessful responses where limitations are obvious, both in terms of elaboration and language use, with a Score 2 response expected to



Write an e-mail to your friend Ralph:

- Choose ONE movie from the schedule below.
- Invite him to the movie.
- Tell him the NAME of the movie and WHEN it starts.

Write as much as you can. Write **complete** sentences.

You have 7 minutes to prepare and write.

SCHEDULE

Second Avenue Cinemas Schedule

Movie	Category	Show time
Fight for the Stars	Action	1:00
Office Party	Comedy	3:30

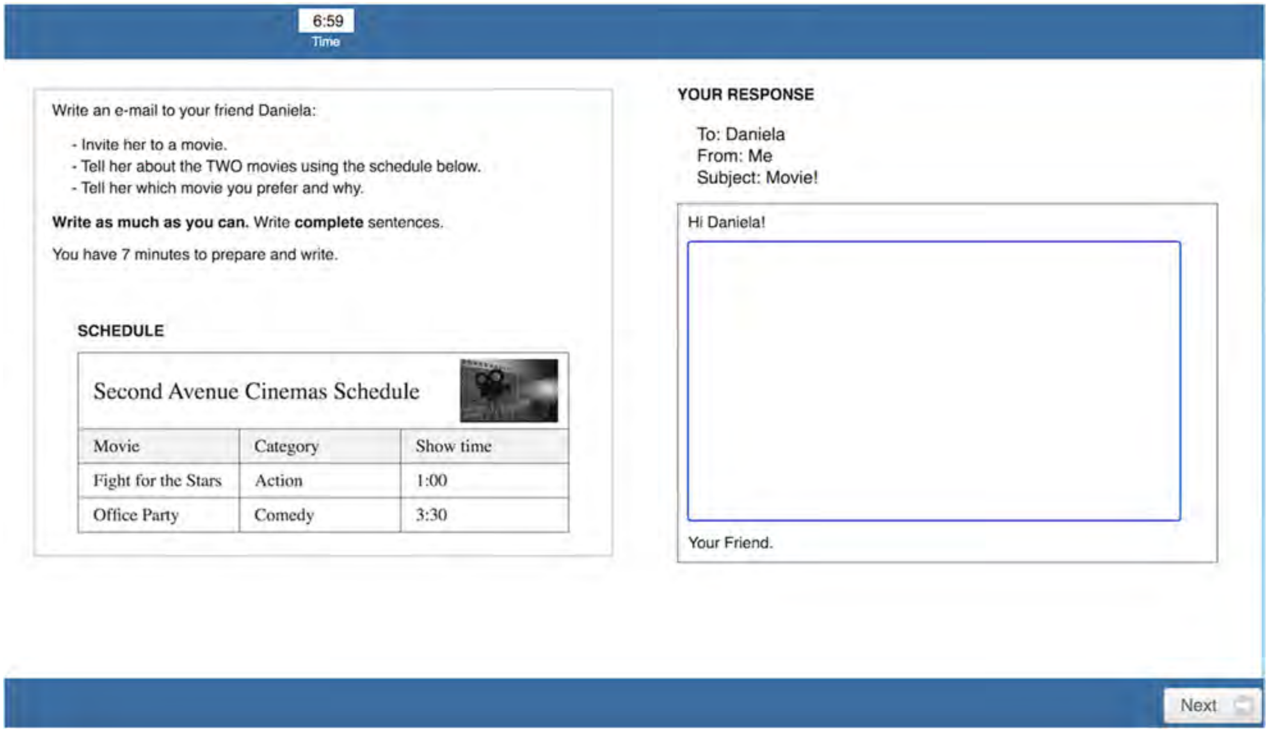
YOUR RESPONSE

To: Ralph
From: Me
Subject: Movie!

Hi Ralph!

Your Friend.

Figure 8 Form A of the Write an E-mail task.



Write an e-mail to your friend Daniela:

- Invite her to a movie.
- Tell her about the TWO movies using the schedule below.
- Tell her which movie you prefer and why.

Write as much as you can. Write **complete** sentences.

You have 7 minutes to prepare and write.

SCHEDULE

Second Avenue Cinemas Schedule

Movie	Category	Show time
Fight for the Stars	Action	1:00
Office Party	Comedy	3:30

YOUR RESPONSE

To: Daniela
From: Me
Subject: Movie!

Hi Daniela!

Your Friend.

Figure 9 Form B of the Write an E-mail task.

Table 1 Summary of the Four Task Types Developed

	Task 1	Task 2	Task 3	Task 4
Task type	Describe a Photo	Write a Review	Chat With a Friend	Write an E-mail
Target proficiency levels	A1–A2	A2–B1	A1–A2	A2–B1
Task context	Social media post	Social media post	Social media post	E-mail
Scaffolding (how to respond)	Little guidance	Some guidance	Some guidance	Substantial guidance
Scaffolding (what to write about)	Pictorial only	Pictorial and textual	Pictorial and textual	Textual only[2]
Response time (min)	5	5	5	7

show some attempt to produce text beyond simple sentences. Last, Score 0 is reserved for a response that is insufficient to assess test takers' ability to deal with a given task (e.g., blank, copied from the prompt, written in a non-English language).

The Current Study

As a critical step in the test development process, we conducted an initial pilot test of the new writing tasks—specifically designed to target lower-proficiency learners of English—to address several fundamental questions about their effectiveness and efficiency. The primary purpose of this pilot study was to finalize the selection of writing tasks for inclusion in the TOEFL Essentials test. In order to do so, we posed the following five research questions:

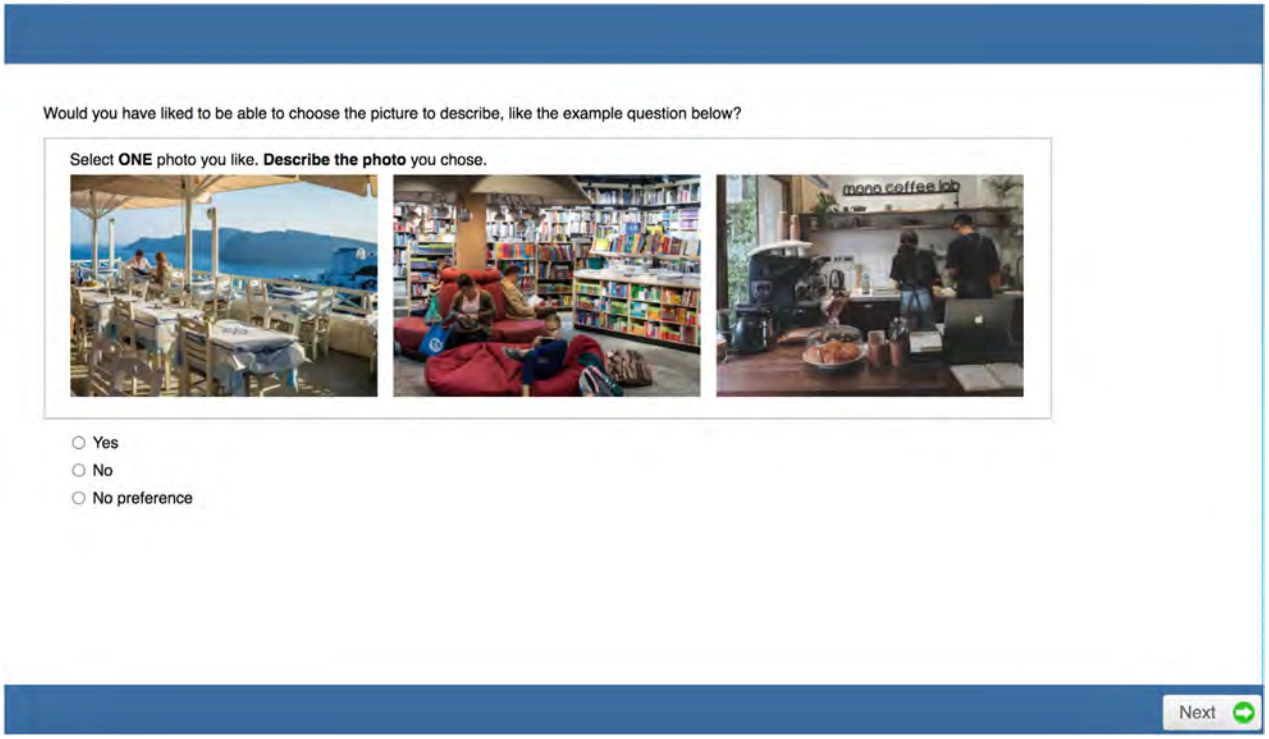
1. To what extent were the test takers able to attempt a response to the four writing task types designed for lower-proficiency learners of English?
2. To what extent were the tasks rated reliably, and was there any difference in rater reliability between the four task types?
3. Which among the four task types demonstrated sufficient levels of discrimination among learners at different low-proficiency levels?
4. To what extent did performances on the four task types differ in terms of (a) holistic scores, (b) amount of writing elicited (fluency), (c) lexical diversity, and (d) syntactic complexity?
5. Which of the two versions of the same task type functioned better in terms of eliciting the test takers' writing abilities?

Method

This section describes the methodology for the current study, including (a) participants, (b) materials, (c) procedures, and (d) data analysis.

Participants

A total of 169 learners of English, who were identified by their English instructors to be at the A1, A2, or B1 level of proficiency on the CEFR, participated in the study. Sixty-seven participated from Japan, and of those, 21 were A1 learners and 46 were A2. The remaining 102 participating students were from Colombia; 17 of them were A1 learners, 50 were A2, and 35 were B1. Thus, in total, 38 A1 learners, 96 A2 learners, and 35 B1 learners participated in the study. The majority of the participants (165 out of 169) were undergraduate university students at the time of the study, and 143 out of 169 fell into the age range of 18–22. One hundred were female, 66 were male, and the remaining three chose either the "other" category or the "prefer not to respond" option about their gender. About one third of the participants (57), mainly from Colombia, had studied English for 2 years or less, and another third (63), mainly from Japan, had studied English for 6–8 years. Twenty participants had spent 3–5 years studying English, and the remaining 29 had learned English for 9 years or more.



Would you have liked to be able to choose the picture to describe, like the example question below?

Select **ONE** photo you like. **Describe the photo** you chose.

○ Yes

○ No

○ No preference

Figure 10 Perception question about choice displayed after Task 1 (Form A: No choice condition).

Materials

In addition to the four writing task types introduced above, a handful of other materials were used in the current study, including (a) a C-test, (b) a perception question, and (c) a background questionnaire.

C-Test

A C-test (Norris, 2018a) was administered to all participants as a quick measure of their L2 proficiency. This C-test adopted the standard design of deleting the second half of every second word in a coherent, paragraph-length text. It consisted of one passage about the role of dancing for ancient humans, determined in previous internal test development research to be a relatively easy passage, and had 20 blanks in total. Participants were given 7 minutes to complete this one-passage C-test. Cronbach's alpha reliability estimate was found to be $\alpha = .84$ with the current population of the test takers, suggesting that it would prove consistent in distinguishing even among participants with low English proficiency.

Perception Question

The study also included one perception question after the Describe a Photo task (Task 1), which asked about the participants' preference related to choice. In Form A, the participants were not offered the possibility of choosing which picture to write a description about. After Task 1, those participants were asked if they *would have liked* to have been able to choose the picture to describe (see Figure 10).

In Form B, the participants did have a choice of the pictures, and those participants were asked if they liked having that choice (i.e., "Did you like having the choice of pictures to describe?").

Background Questionnaire

Participants were asked to fill out a background questionnaire at the beginning of the session to provide their demographic information, including questions on gender, age, level of education, and duration of English language learning.

Procedures

Participants engaged in the study online at home, using their own computers. They were randomly assigned to Form A or Form B and were given a unique ID and a password associated with each form. As a result, 85 participants were assigned to Form A and 84 to Form B. The two groups were found to be equivalent in terms of their overall proficiency levels, with the participants in Form A scoring 10.48 ($SD = 4.81$) on average on the C-test and the participants in Form B scoring 10.43 ($SD = 4.01$). Participants were instructed to go to the study URL and log into the study session using the given ID and password. Once they logged in, they were asked to choose their first language (L1) so they could view and complete the consent form and the background questionnaire in their L1. All participants then encountered the four writing tasks, one version of each task type, in the following order: (a) Describe a Photo (Task 1), (b) Write a Review (Task 2), (c) Chat With a Friend (Task 3), and (d) Write an E-mail (Task 4). Between Task 3 and Task 4, all participants engaged in two short reading comprehension tasks (which will not be discussed in this report). After Task 1, they answered one question to provide their opinion about being able to choose which picture to describe. At the end of the session, after Task 4, participants were presented with the C-test. Participants were instructed to complete all research tasks in one sitting, and the entire session took approximately half an hour for most of them, with the average time spent across all participants being 36 minutes.

Data Scoring, Coding, and Analysis

In this section, we discuss (a) how the responses to the writing tasks were scored, (b) how the linguistic indices were coded and calculated, (c) how the C-test was scored, (d) how the perception question data were tallied, and (e) how all data were analyzed statistically.

Writing Tasks: Holistic Scores

The participants' writing responses were double scored by three raters. The raters participated in two training sessions led by the first two authors. In the training sessions, we reviewed the scoring guide, looked at several sample responses for each task, and conducted practice rating to make sure that the raters were all calibrated. During the practice rating, for each response, each rater provided a score independently, discussed rationales for the score, and then the group resolved any issues or discrepancies in the use of the scoring guide. After the training, the raters each scored 450 responses, so that each response was rated by two raters. The average score between the two raters was used as the final holistic score for each participant. Note that all responses, including the ones written in a non-English language or blank responses, were scored by the raters. Following the scoring guide, those non-English or blank responses were given a score of zero.

Writing Tasks: Linguistic Indices

The participants' writing responses were also analyzed in terms of several linguistic indices, including fluency (total number of words produced), lexical diversity (total number of word types, ratio of original words), and syntactic complexity (mean length of T-unit, or MLT). First, the responses were checked to ensure that they were appropriate for the linguistic analyses. Some participants wrote their responses in the L1, and 16 responses (made by five participants) written in a non-English language were excluded from the analyses. The remaining responses were reviewed to check whether they consisted of arbitrary keystrokes or whether they were simply copied from the prompt. No additional responses were eliminated from the analyses for these reasons.

Fluency

Fluency, or the amount of writing, was measured by counting the total number of words produced by each participant for each task. Contractions, such as *I'm* and *don't*, were counted as two words.

Lexical Diversity

Lexical diversity was analyzed from two perspectives. First, as a simple way to gauge low-proficiency learners' ability to produce a variety of vocabulary, lexical diversity in this study was measured by counting the sheer number of word

Table 2 Total Number of Words Included in the Task Prompt/Input in Each Version of the Task Types

	Task 1		Task 2		Task 3		Task 4	
	Form A	Form B	Form A	Form B	Form A	Form B	Form A	Form B
# of words included in task prompt/input	40	35	73	71	75	78	80	81

types (i.e., unique, nonrepeated words) included in each response (Treffers-Daller et al., 2018). Following Yu (2010) in counting the total number of types, words that belong to the same lemma or word family (e.g., *go, goes, went, gone*) were counted as separate types. Spelling errors were not corrected as it was deemed challenging to try to interpret what each participant meant to write; hence all unique word types were counted (whether they included spelling errors or not) with an understanding that this approach might potentially skew the overall number of word types. Note, however, that the same approach was applied to all tasks and participants; thus, no substantial impact on comparisons made within the study was anticipated.

Lexical diversity was also assessed in terms of how much lexis was produced by the participants themselves, rather than potentially copied or mined from the task prompt and input, by calculating the ratio of the number of original words to the total number of words produced. Thus, the number of words that were *not* included in the task prompt and input seen on the screen while the participants responded to each task was divided by the total number of words produced. Table 2 shows the total number of words included in the task prompt and input in each version of the task types.

Syntactic Complexity

Syntactic complexity was measured by calculating the MLT. First, each response was segmented into T-units. In this study, following Hunt's (1970) definition, a T-unit was defined as "a main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it" (p. 4). In other words, by definition, a T-unit had to have one and only one main verb, except when a compound verb (two or more verbs connected by a conjunction: e.g., "This place is one of the best rated in the city and is close to the bus stop") was used with a single subject. In determining what constitutes a T-unit, some level of inference was required given (a) the genre of writing and (b) the learner population. In the writing genre of social media posts, for example, it is quite natural for a writer to use run-on sentences (although the task prompt specifically asked the participants to write in complete sentences). For instance, the response below is composed mostly of run-on sentences:

> *Good vacations this year!*
> *Happy to be here.*
> *The most beautiful view you can have in your life.*
> *Great food next to a great view.*
> *With the best person, eating my favorite food in my favorite place.*
> *One of the things I always wanted to do.*
> *Traveling you can know more of yourself .*
> *Even the smallest things can make a difference in your life.*
> *Life is most beautiful if we are in the correct place.*
> *Loving this view and this place.*
> *I couldn't be more happy.*
> *I'm grafetul to be here.*

Although those sentences do not entirely fulfill the definition of a T-unit in the conventional sense, given the genre, the main subject and verb were considered to be ellipted and each line was considered as a stand-alone T-unit.

Similarly, the target population and in particular what they can produce in writing made T-unit segmentation less straightforward. Given that our participants were low-proficiency learners of English, many of the responses contained ungrammatical sentences, and possibly the boundaries of idea units were not indicated by an appropriate use of punctuation. In the response below, T-units discerned by the coders are indicated by the square brackets.

> *[hey i going to see a movie] [i wanted to ask you if you wanted ton join me] [the name of the film is office party] [is an action film and will started at 3.30 p.m.]*

In this case, the first T-unit lacks the main verb, but it can easily be inferred that the verb *am* is missing, and otherwise it fits the definition of a T-unit well. Given the target learner population that we were dealing with in this study, we counted such an instance as a legitimate T-unit. The last T-unit comes with no main subject, but it can be asserted that "it [the film]" is missing, and thus it was considered to constitute a stand-alone T-unit. In this example, making those assertions and determining where a T-unit ended were made more difficult due to a lack of punctuation.

On the other hand, a list of words and phrases was *not* considered as a T-unit, unless there was a reason to believe that a verb was missing. The example response below was considered to include no T-unit:

> *Restaurant near sea*
> *white chairs*
> *mountain landscape*
> *city of Santorini*

When coding for T-units, only the body of the message was taken into account. Thus, openers and closers (e.g., "Hi Lisa!" "Xoxo," "Best") were excluded from the total number of words counted. Discourse markers and interjections (e.g., "Oh!" "Yes, of course") were counted as part of the subsequent T-unit unless they formed their own T-unit (e.g., "Thanks!" "You know"). Given the high inference required to segment a response into T-units, the coders participated in a training session, led by the first author, before they embarked on the task of T-unit segmentation. In the training, they reviewed the study-specific definitions of a T-unit and did a practice segmentation of T-units. Subsequently, they each rated 20% of the total number of responses (i.e., 136 responses) chosen randomly. The interrater reliability reached 0.93 overall across the four task types[3]. In the follow-up meeting, we discussed the discrepancies and came up with specific rules to follow in ambiguous cases. Then, the two coders segmented the rest of the responses into T-units. The MLT was calculated by dividing the total number of words by the total number of T-units for each response.

C-Test

The C-test used in the study had 20 blanks, and responses were scored using an exact-response approach for each blank. Thus, a participant was awarded one point for each blank where all missing letters were correctly supplied, with 20 being the maximum score possible.

Perception Question

Participants' responses to the question about whether they would have liked to have had a choice of the picture to describe (Form A) or whether they liked having the choice (Form B) were analyzed by tallying the number of responses for each answer choice (i.e., *Yes, No, No preference*) separately for each form.

Statistical Analyses

In order to discern patterns among the various tasks and measures in this data set, we focused on comparisons of mean values and 95% confidence intervals between the tasks on each measure (Norris, 2015). Prior to making these comparisons, we utilized experiment-wise inferential tests to examine whether task or measure effects were robust enough to detect differences beyond reasonable levels of error. An initial multivariate repeated-measures analysis of variance was conducted with the four tasks serving as the repeated factor and the four linguistic indices plus performance scores serving as dependent variable measures, with an overall alpha level set at $p < .05$. Subsequent univariate analyses were conducted for each measure, followed by graphic and descriptive statistical comparisons between the four tasks on each measure. In conducting inferential analyses using SPSS v. 27, participants with missing data (i.e., those who responded to all or some of the tasks in a non-English language, or provided no answer) were eliminated from the analyses. In the end, $N = 162$ participants' data were included in the inferential and means-comparison analyses. In order to examine capacity of the tasks

Table 3 Interrater Reliability for Each Task Type

	Task 1 (%)	Task 2 (%)	Task 3 (%)	Task 4 (%)
Exact match	51	59	59	55
Adjacent (±1)	44	40	39	41
Discrepant (±2 or more)	4	1	2	4

to discriminate among learners at different proficiency levels, Pearson correlation coefficients were calculated for each task between performance scores and C-test scores. Finally, performances on the two forms of each task on each measure were compared by inspecting the magnitude of mean differences between each form descriptively and statistically, using Bonferroni-adjusted analysis of variance techniques.

Results

In this section, we report on the results of the study in terms of (a) rater performance, (b) learner task performance, (c) discrimination, (d) comparisons between the forms, and (e) the perception question.

Rater Performance

In rating performances on all four task types, the raters were largely in agreement. As seen in Table 3, on the 6-point scale (0–5), the scores given by pairs of raters were either exactly the same or adjacent (i.e., ±1 point different) for 96% or more of all the responses scored. Nonadjacent/exact ratings were extremely rare across all task types.

Learner Task Performance

Overall, the tasks elicited substantial language use from the low-proficiency writers who participated in this study. Of the 169 participants, 162 of them were able to attempt a response to all four task types. Of those seven who were not able to, one provided no answer on Task 3, and the remaining six participants provided answers to either all or some (Task 2 or Task 3) of the tasks in their native language.

Prior to examining and interpreting differences among the four task types in terms of the performance ratings as well as the linguistics measures, an initial multivariate repeated-measures analysis of variance was conducted for all measures on all tasks, with task type serving as the repeated factor. The analysis indicated an overall statistically significant difference across the four tasks on the five measures (Wilks' Lambda $F = 147.595$ [15, 147], $p = .000$). Follow-up univariate tests for each measured variable indicated statistically significant differences across the four tasks on the total number of words produced ($F = 55.70$ [3, 159], $p = .000$), total number of unique word types produced ($F = 94.40$ [3, 159], $p = .000$), percentage of original words produced ($F = 450.128$ [3, 159], $p = .000$), and average performance rating ($F = 13.53$ [3, 159], $p = .000$), but not for average number of words per T-unit ($F = 1.06$ [3, 159], $p = .365$).

Table 4 shows the total number of words produced, the number of word types, the ratio of original words to all words produced, and the MLT. As can be seen in Table 4, all tasks elicited close to 50 words or more on average, with Task 4 eliciting the greatest average number of words. In fact, Task 4 elicited on average 20 words more per performance than did the other three task types. On the other hand, Task 3 elicited the fewest number of words on average. Figure 11 displays

Table 4 Total Number of Words Produced, Total Number of Word Types, Ratio of Original Words, and Mean Length of T-unit (MLT) for Each Task Type

	Task 1 ($n = 165$)		Task 2 ($n = 163$)		Task 3 ($n = 162$)		Task 4 ($n = 165$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Total words	48.70	29.46	49.40	38.82	44.64	28.81	68.55	34.90
Total word types	34.45	16.93	34.70	16.99	33.52	17.55	46.05	18.85
Rat. original words	0.85	0.07	0.55	0.10	0.58	0.09	0.57	0.08
MLT	9.34	3.83	8.93	3.86	8.73	3.50	9.17	3.23

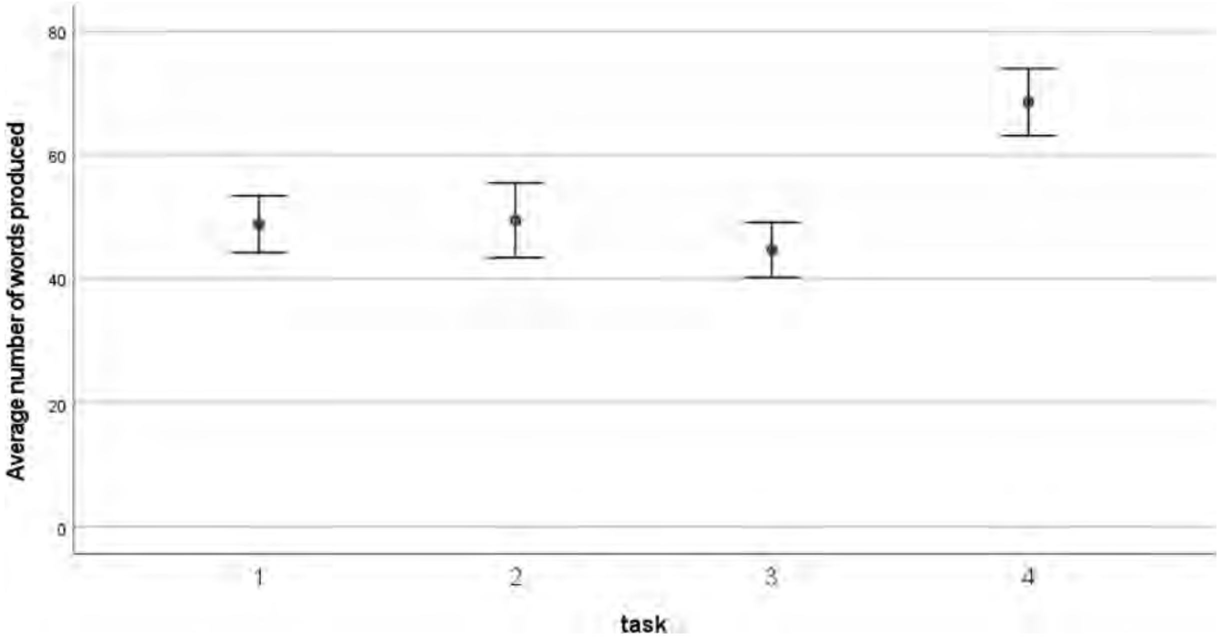


Figure 11 Total number of words produced for each task type.

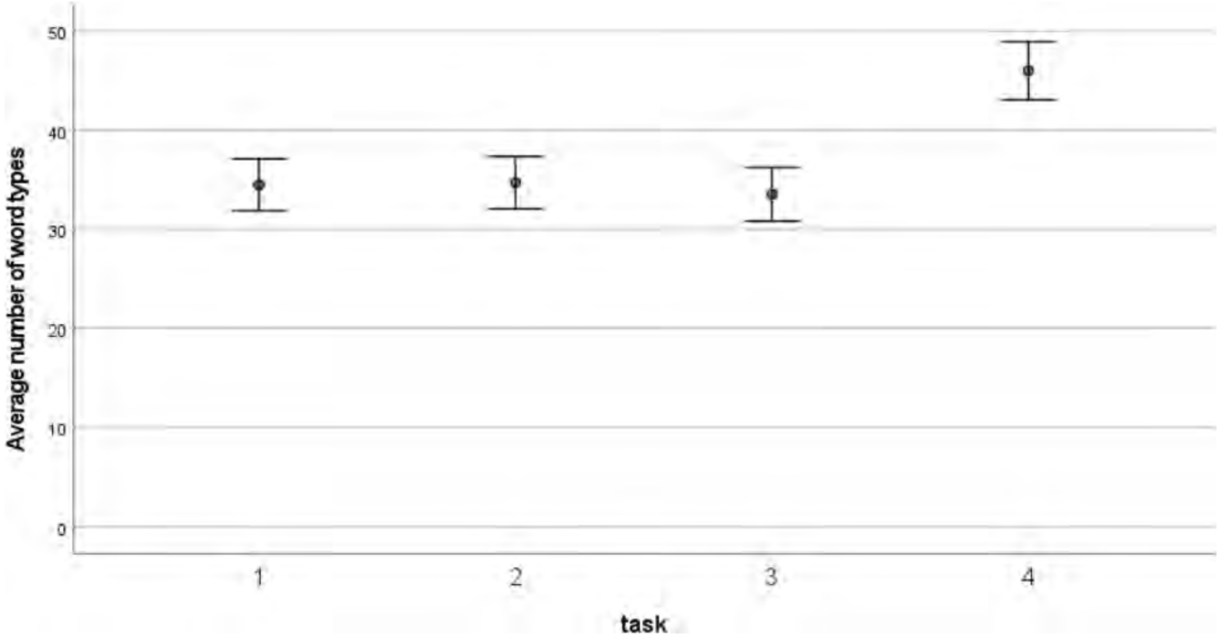


Figure 12 Total number of word types for each task type.

the means and 95% confidence intervals for total number of words on each task. Note that the 95% confidence interval for Task 4 in Figure 11 does not overlap with any of the other task types.

Not only in terms of total words but also for lexical diversity, Task 4 elicited more unique words than did the other three task types. On average, Task 4 elicited 10 more unique words per performance than did the other task types. Again, the 95% confidence interval for Task 4 does not overlap with any of the other tasks (see Figure 12). As was the case for total number of words, Task 3 elicited the fewest number of word types on average.

Now, looking at the ratio of original words produced on each task (see Figure 13), it is clear that Task 1 elicited a much greater proportion of original words, on average around 30% more than the other three task types.

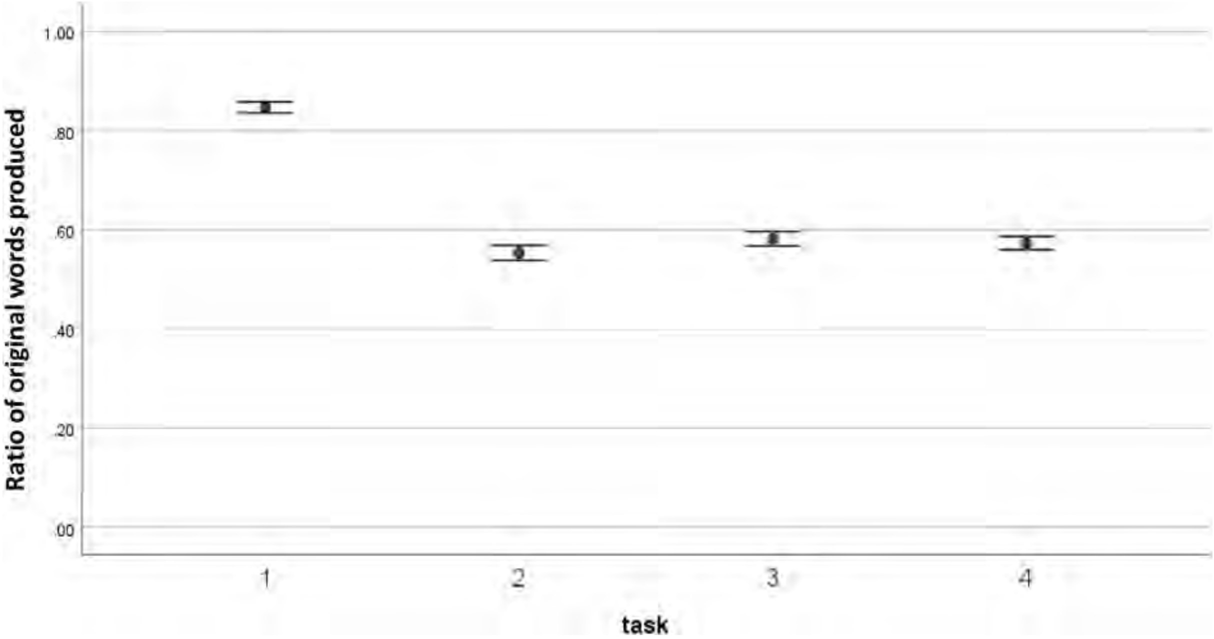


Figure 13 Ratio of original words for each task type.

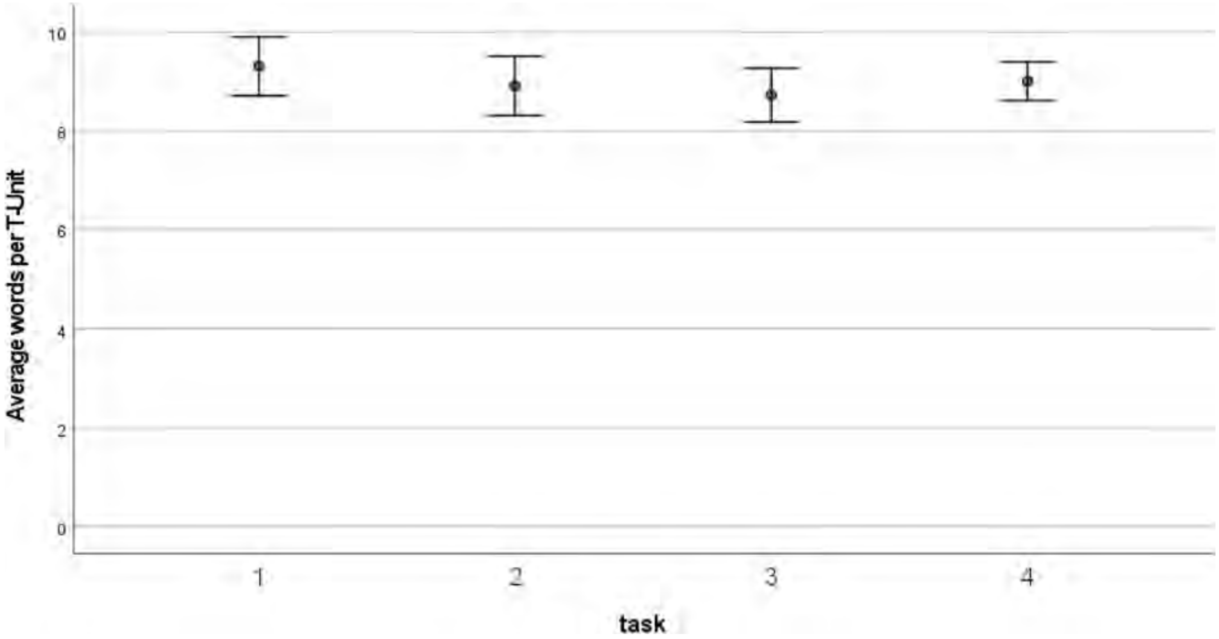


Figure 14 Mean length of T-unit for each task type.

Last, in terms of MLT, Task 1 elicited the longest T-units, and on average more than half a word longer than Task 3, which elicited the shortest T-units on average (see Figure 14). Although the overall differences across the tasks were not statistically significant, the actual difference of 0.5 words between the top and bottom tasks should not be underestimated.

To summarize, Task 4 elicited the most language use and greatest lexical diversity, whereas Task 1 elicited the greatest proportion of original words as well as the longest T-units on average. It is also noteworthy that Task 3 elicited the least language use on three of the four linguistic measures. Additionally, there was good variability (*SD*s) for all linguistic measures, indicating that the tasks elicited broadly differing patterns of language use within the participant population.

Table 5 Average Performance Rating for Each Task Type ($n = 169$)

	Task 1		Task 2		Task 3		Task 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Performance rating	2.52	1.06	2.37	1.03	2.29	1.09	2.66	1.04

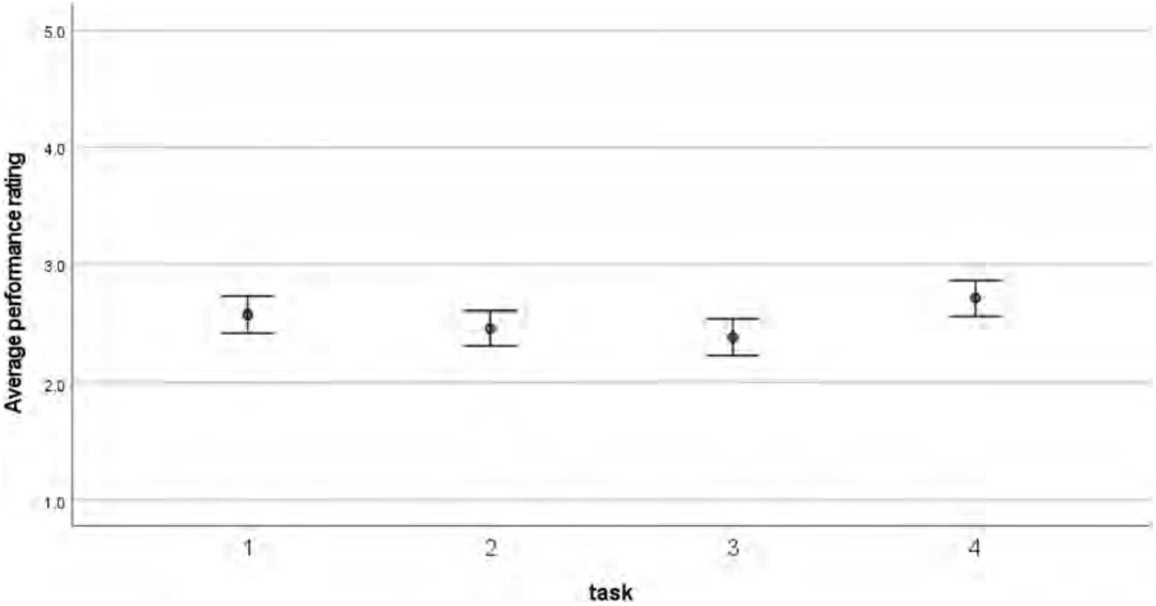


Figure 15 Average performance rating for each task type.

Turning to the holistic performance scores, Table 5 shows the average performance ratings for each of the four task types. The rater scores are well centered in the scale, and they show on average 1 point of standard deviation, which means that the performances varied substantially across the 169 participants. Task 4 elicited performance ratings that were on average 0.37 rating points higher than the lowest-rated Task 3. Note that in Figure 15, the 95% confidence interval for Task 4 does not overlap with those of Tasks 2 and 3.

Discrimination

In order to investigate the extent to which the four task types were more or less capable of discriminating among the low-proficiency test takers, bivariate correlations were calculated between performance ratings on each task and each test taker's score on the C-test. Note that the magnitude of correlation coefficients was anticipated from the outset to be affected by the truncated range of proficiency levels represented among the test-taker sample (i.e., all low-proficiency learners) in the current study. Nevertheless, to the extent that variability in scores existed on both the C-test and the task performance ratings, correlations between the two measure types would serve as one indicator of the capacity of these tasks to discriminate higher- and lower-proficiency learners within the lower end of the proficiency spectrum.

As seen in Table 6, moderately strong relationships were identified between test takers' performance ratings on each of the four tasks and their corresponding C-test scores. Given the overall limited proficiency range in the low-proficiency test-taker sample, these correlations suggest appropriate levels of discrimination in general. Performance ratings on Task 4 demonstrated a noticeably stronger relationship with the C-test scores compared to the other three task types.

Comparisons Between the Forms

Comparing the two versions of the four task types, some different patterns in the linguistic and performance indices were observed. Table 7 shows the means and standard deviations for the four linguistic measures as well as the performance

Table 6 Pearson Correlation Coefficients Between Performance Ratings and C-test Scores

	Task 1	Task 2	Task 3	Task 4
R with C-test scores	.502	.505	.536	.572

Note. All correlations statistically significant, $p < .001$.

Table 7 Comparisons Between the Forms by Task on the Linguistic Indices and Performance Ratings

	Total words		Total word types		Rat. original words		MLT		Performance rating	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
T1 Form A ($n = 83$)	49.19	30.19	34.70	17.84	0.82	0.07	9.23	4.01	2.49	1.06
T1 Form B ($n = 82$)	48.20	28.87	34.20	16.07	0.88	0.06	9.44	3.67	2.54	1.07
T2 Form A ($n = 81$)	55.41	48.95	37.26	19.12	0.56	0.09	8.82	4.27	2.44	1.11
T2 Form B ($n = 82$)	43.46	23.99	32.17	14.25	0.55	0.10	9.03	3.43	2.30	0.94
T3 Form A ($n = 81$)	48.37	32.19	36.75	20.11	0.56	0.08	8.99	3.58	2.48	1.18
T3 Form B ($n = 81$)	40.91	24.60	30.28	13.93	0.60	0.10	8.47	3.43	2.10	0.97
T4 Form A ($n = 83$)	63.72	34.16	43.87	19.32	0.59	0.09	9.45	4.02	2.70	1.13
T4 Form B ($n = 83$)	73.44	35.17	48.26	18.22	0.56	0.08	8.89	2.15	2.61	0.94

Note. T1 = Task 1, T2 = Task 2, T3 = Task 3, T4 = Task4. Shaded cells indicate comparisons that were found to be statistically significantly different, $p < .05$, Bonferroni adjustment made for multiple comparisons.

ratings separately for Form A and Form B. Note that statistically significant differences are indicated by the shaded cells. In order to take full advantage of the information provided by these comparisons, in addition to statistically significant differences, we also took into account the magnitude of all observed differences in the mean values.

On Task 1, no perceptible differences between Forms A (choice of a picture to describe *not* provided) and B (choice of a picture provided) were identified, except for ratio of original words produced, where a small difference in favor of Form B was detected (though overall both forms showed a much higher proportion of original words than did the other three tasks). On Task 2, Form A (a café review) clearly elicited more words and more word types than did Form B (a tablet review) though the difference in word types was not statistically significant. On Task 3, Form A (chat about a trip) elicited distinctly better performances than did Form B (chat about a new bike): 6.47 more word types and 0.38 points higher performance ratings (statistically significant differences), as well as statistically nonsignificant—though apparent—differences on total number of words (7.46 more words) and MLT (0.52 words longer T-units). A small difference in the opposite direction (in favor of Form B) was noted for ratio of original words. Last, on Task 4, although the differences were not statistically significant, Form B (which asked the participants to invite a friend to a movie, describe the two movies, and explain which one they prefer and why) elicited noticeably more words and word types than Form A (which asked the participants to simply invite a friend to a movie and provide some information about it). However, Form A elicited T-units that were 0.56 words longer than Form B. We return to these interesting differences in the Discussion section.

Perception Question

Finally, the perception question posed after Task 1 revealed an interesting difference in preference for choice between the participants who were assigned to the two distinct forms. As can be seen in Table 8, just a little more than half the participants who were assigned to Form A, with no choice, expressed their desire to have been able to choose which picture to describe when engaging in Task 1. On Form B, about 88% of the participants who did experience having a choice of the pictures to describe indicated they liked having that choice.

Discussion

In this section, we discuss the findings to answer our research questions, with a particular focus on which of the four task types, and which versions, seem to be best suited to meeting the assessment purposes and design expectations.

Table 8 Number of Participants Who Chose *Yes, No*, or *No preference* about the Question Related to Choice on Task 1

	Yes	No	No preference
Form A ($n=85$)	47	13	25
Form B ($n=84$)	74	1	9

Which Task Types?

Overall, the four task types performed fairly well, though some substantial differences were observed among them. First, in terms of accessibility to the target population (i.e., low-proficiency L2-English writers), they all generated substantial language use (close to 50 words or more on average) within the allotted times, and there were only seven participants out of 169 who either did not enter anything or provided their responses in their native language. Thus, it is probably fair to say that all four task types were accessible to most of the low-proficiency participants in this study, including the ones who scored very low on the C-test proficiency measure. Similarly, all four task types were rated reliably, with discrepancies (difference of more than two scale points) of less than 4%. Clear differences between the interrater reliabilities among the different task types were not observed. In other words, all tasks performed equally well when it came to interrater reliability.

What were the main differences observed among the four task types? Differences were identified in terms of the linguistic indices, performance ratings, and the degree of discrimination. Looking at the linguistic indices first, Task 4 elicited the greatest number of words as well as word types, whereas Task 1 elicited the greatest proportion of original words that were not included in the task prompt/input and the longest T-units (although the MLT differences with the other tasks were not statistically significant). It is also notable that Task 3 elicited the least language use across all linguistic measures but ratio of original words.

Task 4 elicited the longest responses of the four task types, most likely because its task prompt provided the participants with the greatest degree of scaffolding or guidance on how to respond to the task as well as ample time to construct the response. The task required participants to perform several speech acts (i.e., invite and describe in Form A; invite, describe, and justify in Form B), and it was scaffolded by providing them with step-by-step instructions on how to write an effective e-mail message in the given context (see Figures 8 and 9). By contrast, Task 1 simply asked the participants to write about the picture without further guidance on what to include in the response or how to write an effective response. Task 2 provided some degree of guidance on what to be included in the response, by instructing test takers to include all information provided about the café or the tablet. Task 3 provided a little more scaffolding in that it asked three questions that the participants could write about; however, the questions were fairly brief, which seems to have made it challenging for test takers to elaborate on their answers. Especially for low-proficiency learners, there may be an important tension between providing scaffolding in hopes of making the task *less* demanding and inadvertently increasing the reading load. In the case of Task 4, it seems that the scaffolding (coupled with extended response time) had a positive effect and helped the participants write more and show what they could do in English more extensively than did the other three tasks.

Task 1 came with little scaffolding, not only in terms of guidance on how to respond to the task, but also in terms of the language that participants could mine in writing their responses. That is, Task 1 provided little *linguistic* scaffolding: Except for the task prompt, there was virtually no language that the participants could mine in writing their response. Tasks 2 and 4, on the other hand, provided a lot of linguistic scaffolding or language to mine, in an effort to make the tasks more accessible to the low-proficiency writers. For example, in Task 2, the information about Café Choli or the Axis tablet was provided in the input. Although we included only the language that was minimally needed (e.g., words instead of sentences), in some cases, all the test takers needed to do was put together what was provided in the input, like "*location* is *close to bus stop*" (the words provided in the input are italicized). Task 1 is unique in the sense that the scaffolding was only pictorial in nature, and this characteristic worked to its strength as it provided greater evidence of what the low-proficiency writers could do on their own when linguistic scaffolding was not provided. Interestingly, as the test takers worked to produce their own language in response to Task 1, the syntactic complexity of their performances may also have been positively affected. That is, rather than building from linguistic information in the input of the other tasks or following the scaffolded guidance provided, test takers had to do it alone on Task 1, and this factor may have led them to stretch their lexical and syntactic abilities. In summary, based on evidence from the pilot study, Task 4 was the task that elicited the most substantial and diverse language use from the low-proficiency learners, and Task 1 was the task that elicited the greatest proportion of original language use as well as syntactically most complex language. Together, these two tasks were deemed to provide adequate evidence about low-proficiency learners' writing abilities in English from two distinct perspectives.

Turning to the performance ratings, Task 4 performances were rated highest on average (followed by Task 1), and these were rated statistically significantly higher than performances on Tasks 2 and 3. Note that the scoring guide was designed to award a higher score to a response that was elaborated and thus demonstrated test-taker abilities to a greater extent. With Task 4 being the task that elicited the most language use, this finding may be expected. Task 3 performances, on the other hand, were rated the lowest. This finding is interesting in that Task 3 was designed to be easiest and most accessible to the low-proficiency learners. All they needed to do in Task 3 was answer the three simple questions about their hypothetical trip (Form A) or bike (Form B). Given how the scoring guide was designed, it is most likely that the participants received the lowest average score because the task did not do a good job of eliciting elaborated responses, which overall made the task more difficult to get a higher score on with respect to the generic scoring guide. Here again, Tasks 4 and 1 seemed to provide the greatest opportunities for test takers to demonstrate their abilities to communicate in attempting to accomplish the tasks.

Focusing now on discrimination, all four tasks showed moderately strong relationships between test takers' performance ratings and their corresponding C-test scores. It can be argued that these correlations suggest appropriate levels of discrimination in general, given that the target population of interest had the restricted range of lower-proficiency levels. Among the four tasks, again, it was Task 4 that showed the strongest correlation with C-test scores and hence was deemed best at discriminating among different proficiency levels at the lower end of the spectrum.

All in all, Task 4 was an easy choice for inclusion in the low panel of the Writing section in the operational test: It elicited the most substantial language use and was best able to discriminate among different proficiency levels. Our second choice for inclusion was Task 1, as justified by the fact that it encouraged the test takers to stretch their interlanguage and produce syntactically more complex language, as well as to use words and phrases that were not provided in the task prompt/input. Task 1 also showed sufficient discrimination.

Which Version?

To what extent did different variations of the same task types lead to observable differences in test-taker performances? First, Form B of Task 1, Describe a Photo, offered the test takers a choice of three pictures to describe, but Form A did not offer such a choice. From a perceptual point of view, it is interesting that 88% of the participants who were assigned to Form B liked having the choice of the pictures, but only 55% of the participants in Form A said they would have liked to have had a choice of the pictures to describe. In terms of performance, however, Form A and Form B showed almost no perceptible differences for this task type. Only a small difference in ratio of original words was observed, although both forms elicited high proportions of original words. Despite the fact that choice was seen preferably by participants assigned to both Form A and Form B, aspects of participants' performances seem to have been affected quite minimally by having or not having a choice of pictures to describe. Given these findings, the decision was made *not* to provide a choice in the operational test, in order to maintain standardization and consistency in new task development, while keeping that possibility in mind for future iterations of the test.

Turning to Task 2, Write a Review, Form A asked the participants to review a café, and Form B asked them to review a tablet. Interestingly, Form A elicited more words and word types on average than did Form B. Looking at the participants' responses, it appears that they had more to say about a café, and some even added extra information that was not included in the input, namely the location's nearness to the bus stop, the high-quality coffee they serve, and people being friendly (see example responses below).

Participant 231:

Cafe Choli has a lot of seats. It may be famous by women. It locates close to bus stop. So you can go easily. It gets parfect score. Customer of this cafe cn talk with each other. Shall we go to there to have lunch?

Participant 001:

it is a coffee that have a good reputation. its close to te bus stop, easily to find it. it have a big inner space. its design is cool and the people are friendly.

Participant 021:

I think that's one of the best places I've ever visited with my best friend Cristian. The attention was so good and everything came out better that we expected. Also, there is a variety of food and different types of coffee. Another thing to add about this place was the environment, the music wasn't too loud or too low, it was the perfect level. I think it's better if you in the afternoon or at night because you can appreciate in a fantastic way the view.

For Form A, it seems that many participants were mining the content of their response from the picture provided, in which several women are enjoying their coffee and working on a computer in a relatively large space, with one woman facing a large window. As seen in the response by Participant 021 above, some also elaborated by including creative personal stories. This may be because it is a common experience to go to a café or even to talk with friends about which café is good and why.

Writing about a tablet, on the other hand, elicited less language use. Most of the participants did not include any extra information other than what was provided in the input (see the example response by Participant 262 below), and if included, it was fairly minimal (e.g., the color, things one can do with a tablet in general).

Participant 262:

A Axis tablet's screen size is small.
And, it has only 1 camera.
But it cost you $800!
I think it is too expensive.
You should not buy it.

Participant 010:

its price is $ 800, has a small size screen and a single camera and coes in white

Participant 068:

this is a good product because is cheap.
this produce have one camera for you take a photo.
and is easy have in your hand because is small.
is really cheap, Axia tablet have a price of only 800 dolars.
with this tablet you can read books, connect to social media, and you can do homework.
have a good camera.

In sum, Form B did not elicit as much language use as did Form A, perhaps due to the nature of the picture or the product being reviewed. Unlike the picture included in Form A, the picture of the tablet did not provide any new information, except for the color, and tablets per se or talking about pros and cons of a tablet may not be as common as going to or talking about cafés for this group of participants. Although we decided not to include Task 2 in the operational test, findings regarding the likely impact of pictorial input were useful in offering additional guidance for other tasks during the assessment development process.

It is curious that a similar trend was observed for Task 3, Chat With a Friend. Form A asked the participants to write a response about a hypothetical trip, and Form B asked them to write about a new bike that they supposedly just acquired. In Task 3, Form A elicited not only more words and word types, but also longer T-units and higher performance ratings than did Form B. On the other hand, a small difference was observed in favor of Form B for ratio of original words. Similarly to Task 2, a more common topic, a trip, might have elicited more substantial and syntactically more complex language use compared with a bike that not everyone has or enjoys riding. Alternatively, it is plausible that talking about or expanding on an object like a tablet or a bike is harder to do than talking about a place, such as a café, or a journey destination and associated experiences. Although these speculations may help with future item development, more research is required to come to a firm conclusion about the relationship between a topic, its visual scaffolding, the task prompt/expectations, and learner performance, especially for low-proficiency writers. It is also unclear why Form B elicited a slightly (but statistically significantly) higher proportion of original words than did Form A. In any case, given the discrepancies in performances observed between Forms A and B of Tasks 2 and 3, we were further convinced to not include these tasks in the operational test, at least until additional evidence can be generated to resolve the above uncertainties.

Last, looking at Task 4, Write an E-mail, Form A asked the participants to choose one movie out of the two provided in the input, invite a hypothetical friend to the movie, and tell the friend the name and the start time of the movie. On the other hand, Form B challenged the participants a bit more by asking them to invite a hypothetical friend to a movie, talk about the two movies provided in the input, and in particular explain which movie they prefer and why. This design difference had a noticeable impact on the participants' written performance. Form B elicited more words and word types than Form A; however, Form A elicited T-units that were half a word longer than Form B. Form B elicited a longer response with more word types, most likely because the participants needed to describe the two movies and explain why they prefer one over the other as intended by the design of the task. Although the requirement to discuss the two movies did elicit more substantial language use, the responses tended to be like a list, as seen in the example response below:

Participant 236:

I'm going to see movies and invite you.
Before going to theater, I would like to dicide which we watch.
There are two movies.
One's name is ""Fight for the Stars"". It's category is ""Action"" and it starts at 1:00. Another's name is ""Office Party"".
It's category is ""Comedy"" and it starts at 3:30.
I prefer see ""Fight for the Stars"" because we love action and we would enjoy this time.

This trend probably explains why Form B elicited syntactically less complex structures than did Form A. The unconnected list of movie information did increase the word counts, but it was based, by and large, on somewhat shorter sentences. Given these findings, in the operational version of the test a decision was made to ask test takers to (a) perform a social act (e.g., invite a friend to a movie) by choosing one of the two options provided in the task input, and (b) justify their decision (e.g., explain why they prefer one movie to the other).

Conclusion

This report described the development of L2 writing assessment tasks that aimed to efficiently assess low-proficiency English learners' communicative writing ability and discriminate effectively among learners at distinct proficiency levels. The desired characteristics of the tasks were that they needed not only to get at key aspects of language knowledge and skills expected at the A1, A2, and B1 levels of proficiency on the CEFR, but also to replicate real-world scenarios and purposes for writing while distinguishing among low-proficiency levels (A1 through B1) efficiently. The four task types that were thought to meet those criteria and were developed in this project included (a) Describe a Photo, (b) Write a Review, (c) Chat With a Friend, and (d) Write an E-mail. Based on the results of a pilot test with 169 low-proficiency learners of English, the Describe a Photo and Write an E-mail tasks were chosen to be included in the operational version of the TOEFL Essentials test.

The analyses revealed that the Write an E-mail task elicited the most substantial language use (i.e., the highest total number of words and types, second longest MLT), which led to the greatest performance scores on average and the best discrimination. This finding suggests that, generally speaking, when designing L2 writing assessment tasks, the more language a task can elicit, the better evidence it can provide about test takers' writing abilities, and in particular what they can *do* in the target language. Importantly, extended elicitation of language use was associated with substantial scaffolding in the form of goals for what to accomplish in the e-mail task. This design aspect is also important from the task-based assessment perspective. If the goal of assessment tasks is to elicit real-world language use, it is essential to provide a context where test takers can actually demonstrate what they can do in real-world communication in response to expectations of the task and audience.

The study findings also revealed that the Describe a Photo task was best at measuring what test takers can do on their own without much scaffolding. To reduce the reading load, the Describe a Photo task was presented with very little guidance on how to write a response, and thus there was very little language to be mined, whereas the other tasks provided more guidance and linguistic scaffolding. As a result, the Describe a Photo task elicited the greatest proportion of words that were not included in the task prompt, which provided important additional evidence about the L2 learners' writing ability.

Clearly, in designing language assessment tasks, it is essential to consider what evidence the tasks need to elicit in order to accurately measure test takers' L2 proficiency. In response to the assessment mandate for the TOEFL Essentials test, we identified a handful of test tasks that are able to reflect key dimensions of writing ability among low-proficiency English learners, as described in the CEFR levels A1 – B1. By investigating how actual target test takers responded to and performed on these prototype test tasks, we were able to narrow down the candidate tasks to two distinct types which, combined, provide sufficient coverage of different aspects of communicative language ability. We were also able to optimize our task designs by examining how important aspects of test instructions, prompts, and graphical support affected target test takers' performance. The results of the assessment task design and pilot-testing process described here provided a sufficient procedural and evidentiary basis for moving ahead to developing test tasks for the operational version of the low-proficiency panel on the TOEFL Essentials test.

Acknowledgments

We would like to thank our colleagues Larry Davis, Phil Everson, Brad Daniels, and Spiros Papageorgiou for their support in the development and evaluation of the writing tasks as well as the scoring guide. We would also like to thank our research assistants, Mike Suhan, Michelle Kim, and Jeremy Lee, for their assistance in scoring and data coding, as well as ETS technical reviewers for their feedback on the manuscript.

Notes

[1] Note that in the operational version of the TOEFL Essentials test, all writing tasks are scored according to task-dependent scoring guides.

[2] Although Task 4 does contain a picture (a motion-picture projector) in the input, it is more for decorative purposes, and it does not provide any new information that is not available in the textual input.

[3] First, the rate of agreement between the two raters was calculated for each response by dividing the number of T-units that both raters agreed upon by the total number of T-units. Then, the overall interrater reliability was calculated by averaging the interrater agreement across all responses on each task type.

References

Chapelle, C. A., Enright, M. E., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume*. Council of Europe Publishing.

Crossley, S. A., & McNamara, D. S. (2010). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115 – 135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243 – 263. https://doi.org/10.1177/0265532211419331

Cumming, A., Cho, Y., Burstein, J., Everson, P., & Kantor, R. (2021). Assessing academic writing. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 107 – 151). Routledge. https://doi.org/10.4324/9781351142403-4

Davis, L., & Norris, J. M. (2021). *Developing an innovative elicited imitation task for efficient proficiency assessment* (TOEFL Research Report No. 96). ETS.

Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*(2), 414 – 420. https://doi.org/10.2307/3587446

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*(2), 123–145. https://doi.org/10.1016/S1060-3743(00)00019-9

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development*, *35*(1), 1–67. https://doi.org/10.2307/1165818

Jiang, J., Bi, P., & Liu, H. (2019). Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, *46*, 100666. https://doi.org/10.1016/j.jslw.2019.100666

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference Levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532. https://doi.org/10.1093/applin/amy064

Long, M. H., & Norris, J. M. (2000). Task-based language teaching and assessment. In M. Byram (Ed.), *Encyclopedia of language teaching* (pp. 597–603). Routledge.

Mislevy, R., Steinberg, L., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496. https://doi.org/10.1191/0265532202lt241oa

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and some solutions. In J. M. Norris, S. Ross & R. Schoonen (Eds.), *Improving and extending quantitative reasoning in second language research* (pp. 95–124). Wiley-Blackwell.

Norris, J. M. (2018a). Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 7–33). Peter Lang.

Norris, J. M. (2018b). Task-based language assessment: Aligning designs with intended uses and consequences. *JLTA Journal*, *21*, 3–20. https://doi.org/10.20622/jltajournal.21.0_3

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the* TOEFL® Essentials^{TM} *test 2021* (Research Memorandum No. RM-21-03). ETS.

Sasayama, S. (2015). *Validating the assumed relationship between task design, cognitive complexity, and second language task performance* [Unpublished doctoral dissertation]. Georgetown University.

Tack, A., Francois, T., Roekhaut, S., & Fairon, C. (2017). Human and automated CEFR-based grading of short answers. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudaki (Eds.), *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 169–179). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5018

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327.

Xi, X., & Norris, J. M. (Eds.). (2021). *Assessing academic English for higher education admissions*. Routledge. https://doi.org/10.4324/9781351142403

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Appendix A

This appendix presents the final scoring guide used in the current study

Scoring Guide	
5	**A fully successful response** The response is well elaborated and shows consistent facility in language use A typical response consistently displays all of the following: ● The response is fully elaborated with relevant supporting detail ● A range of grammar and vocabulary is used effectively (e.g., to create a degree of fluidity, precision, idiomaticity, or expressiveness) ● Language and mechanics are accurate, although minor errors or non-idiomatic uses may occur However, the intended meaning is fully clear and cohesive throughout

Scoring Guide	
4	**A generally successful response** The response is elaborated, although minor inconsistencies in language use are present A typical response shows most of the following: ● Suitable elaboration is provided ● A variety of grammar and vocabulary is used appropriately, but range, cohesion, or precision may not be sustained throughout the response ● Inaccuracies in language use or mechanics may be noticeable and cause minor ambiguities in sentence-level meanings, but the overall response is clear
3	**A partially successful response** The response shows some elaboration and/or facility in language use, but is limited in one or more aspects A typical response partially accomplishes the task, and shows either of the following patterns: ● The response contains a reasonable degree of elaboration, with limitations elsewhere ○ A somewhat limited range of grammar and vocabulary is used; attempts at more complex structures or vocabulary are not entirely successful ○ Limitations in mechanics, range, or accuracy may impact the readability or clarity of part or all of the response. The response may lack cohesion and be visibly list-like ● The response shows limited elaboration ○ An adequate range of grammar and vocabulary is used; limitations in language use do not greatly impact the overall clarity of the response
2	**A mostly unsuccessful response** The response is limited in content and in facility of language use A typical response shows some or all the following: ● Very limited elaboration is present, which significantly impacts the effectiveness of the response ● A limited range of grammar and vocabulary is used, although some attempt is made to produce sentence-level language beyond simple clauses or basic vocabulary. The response may consist of a list of relatively disconnected items ● Errors in language use or mechanics may cause the intended meaning to be unclear
1	**An unsuccessful response** The content of the response is very limited, with little evidence of ability to produce extended text A typical response shows some or all of the following: ● Very little elaboration is provided ● The response consists of telegraphic language (short and/or disconnected phrases and simple sentences) ● Serious errors in language use may be present
0	The response is blank, rejects the topic, is not in English, is entirely copied from the prompt, is entirely unconnected to the prompt, or consists of arbitrary keystrokes

Scoring notes: Even though a response does not fully follow the task instructions (e.g., it does not include all information provided about Café Choli/Axis tablet in the Write a Review task, it does not answer all three questions in the Chat With a Friend task), it can receive up to Score 4 as long as the response fits into the description of a given score band. For a response to be awarded Score 5, the response should fully follow the task instructions.

Suggested citation:

Sasayama, S., Garcia Gomez, P., & Norris, J. M. (2021). *Designing efficient L2 writing assessment tasks for low-proficiency learners of English* (TOEFL Research Report No. 97). ETS. https://doi.org/10.1002/ets2.12341

Action Editor: Jonathan Schmidgall

Reviewers: Ikkyu Choi and Jennifer Sakano