ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-96
ETS Research Report No. RR–21-24

Developing an Innovative Elicited Imitation Task for Efficient English Proficiency Assessment

Larry Davis

John 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.

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Developing an Innovative Elicited Imitation Task for Efficient English Proficiency Assessment

Larry Davis & John Norris

ETS, Princeton, NJ

The elicited imitation task (EIT), in which language learners listen to a series of spoken sentences and repeat each one verbatim, is a commonly used measure of language proficiency in second language acquisition research. The *TOEFL® Essentials™* test includes an EIT as a holistic measure of speaking proficiency, referred to as the "Listen and Repeat" task type. In this report, we describe the design considerations that informed the development of the EIT for TOEFL Essentials. We also report the results of a series of investigations conducted during the prototyping and pilot phases of test development, which were undertaken with the goal of confirming task design specifications, evaluating scoring performance, and obtaining initial validity evidence to support score interpretation and use of the EIT in the TOEFL Essentials test. We found that task design variables generally performed as expected. The length of input sentence was strongly associated with performance (Pearson $r = .88$), consistent with the construct measured by the EIT, while other task variables not directly related to the EIT construct did not impact performance (e.g., graphics, speaker accent, and response time). Scorers drawn from TOEFL iBT test raters were able to score responses consistently with over 98% exact or adjacent interrater agreement on a 6-point scale, and scores on the pilot version of the EIT were highly reliable (Cronbach's $\alpha = .93$ on the 15-item pilot version). Correlations between EIT scores and other measures were generally as expected: Correlations with other speaking tasks were high (.78 – .84) and slightly to somewhat lower for other language measures (.73 for writing, .68 for listening, and .57 for reading). Correlation with an independent measure of holistic language proficiency (C-test) was moderately high (.69), as expected. We discuss the study findings in terms of the TOEFL Essentials test validity argument and point out limitations to the current results along with future research needs. Overall, we believe that the findings provide initial support to warrant the use of the EIT as operationalized in the TOEFL Essentials test.

Keywords Speaking assessment; second language proficiency; foundational language skills; validity argument; TOEFL Essentials test

doi:10.1002/ets2.12338

Background

Elicited Imitation Tasks as Measures of Second Language Proficiency

Second language (L2) proficiency can be measured in a variety of ways, depending on the purposes, audiences, and uses for assessment. L2 proficiency is also a complex and multifaceted construct, the specific definition of which has long been subject to debate (e.g., Cummins, 1979; Hulstijn, 2015; Long & Richards, 1990) and the components of which include aspects of explicit knowledge (e.g., of grammar rules, vocabulary, pragmatics), strategies for communication, abilities for using language to get things done, and many additional refinements (e.g., Bachman, 1990; Bachman & Palmer, 2010; Norris & Ortega, 2003). As a result of these complexities, assessment options for measuring L2 proficiency vary considerably, from selected-response tests of linguistic knowledge to performance assessments of situated communication tasks and a host of others (Brown & Hudson, 1998; Norris & Ortega, 2012). A key challenge in deciding among the available measurement options has to do with the need to balance precision and coverage of a specific proficiency construct with the need for measurement efficiency that is often required in practical assessment circumstances.

One approach to L2 proficiency assessment — reduced-redundancy testing — has directly addressed this goal of balancing efficiency and the adequate measurement of an L2 proficiency construct. Over the past half century (e.g., Caulfield & Smith, 1981; Gradman & Spolsky, 1975; Oller Jr., 1973), reduced-redundancy tests have been developed, investigated, and used, with a particular focus on the rapid estimation of global L2 proficiency in a variety of languages. These tests expose L2 learners to generally small amounts of atypical language input (e.g., mutilated written texts; noisy, quick, or

Corresponding author: L. Davis, E-mail: ldavis@ets.org

decontextualized aural stimuli—thus a reduction of the redundancy typical in authentic language use), then require the learner to process the input simultaneously for meaning and form to produce a response (e.g., rewrite the complete text accurately, repeat the full aural stimuli). Especially popular exemplars of this approach to assessment are literacy-based measures, including the Cloze and C-test, both of which have experienced robust investigation and operational test use (e.g., Bachman, 1982; Brown, 2013; Grotjahn et al., 1992; Klein-Braley, 1985; Norris, 2018; Trace, 2020). As indicated in accumulated research on these assessments, the advantages of reduced-redundancy tests include (a) rapid test development, (b) reduced time on task required of test takers, (c) promisingly high levels of reliability and discrimination across a broad proficiency spectrum, and (d) high criterion-related validity with other measures of L2 proficiency. Where efficiency, reliability, and a focus on global or holistic L2 proficiency are prioritized, these types of assessment may offer an attractive option.

The Elicited Imitation Task

The elicited imitation task (EIT) presents an interesting variation on the reduced-redundancy theme. Within an EIT assessment, test takers are presented with a series of sentences that are spoken aloud one by one, and the test taker must orally repeat each sentence as precisely as possible. The fundamental assumption underlying the use of EITs as measures of L2 proficiency is that learners must rapidly process what they hear (nearly simultaneous to the hearing of the stimulus), understand the meaning, and parse the stimulus linguistically to be able to reproduce it with full accuracy (e.g., Bley-Vroman & Chaudron, 1994; Gaillard & Tremblay, 2016; Markman et al., 1975; Vinther, 2002). Language learners with a more highly developed internal language system (including native speakers) are expected to more efficiently process language input for form and meaning and so should be able to repeat the stimulus sentences more accurately, and also should be able to perform better on longer sentences, than learners with lower proficiency. The fundamental construct underlying EIT performance focuses on the integrated abilities of a language learner to (a) process sentences aurally for meaning and form, (b) orally reproduce the sentences with accurate meaning and form, and (c) do so for sentences that increase in length and grammatical complexity. Within a relatively brief test administration time (i.e., a handful of sentences), then, it is possible to separate learners into distinct overall proficiency levels based on the EIT approach. Another advantage of the EIT is the focus on aural/oral language processing in an online (i.e., immediate) context, which suggests that the task taps into underlying or implicit L2 competence (see Erlam, 2006; Hulstijn, 2015; Sarandi, 2015), as opposed to declarative linguistic knowledge, which might play more of a role in test formats like the Cloze or C-test. Furthermore, the EIT calls upon the integration of various components of language competence, including lexical, morphosyntactic, phonological, and prosodic knowledge and abilities, suggesting that it may serve as a useful predictor of global proficiency rather than measuring discrete components of L2 ability.

Because of these apparent advantages, EITs have been developed and used for measuring L2 proficiency for a variety of purposes, with learners from beginning to advanced proficiency levels and across numerous languages (e.g., Deygers, 2020; Gaillard & Tremblay, 2016; Lever & Lonsdale, 2015; Mozgalina, 2015; Ortega, 2000; Tracy-Ventura et al., 2014; Wu & Ortega, 2013). They also have been investigated to the extent that at least three meta-analyses have been conducted on EIT research to date (Kostromitina & Plonsky, 2021; Yan et al., 2016; Zhou, 2012). Overall, findings provide substantial support for the use of EITs as measures of global L2 proficiency (Kostromitina & Plonsky, 2021). Average reliabilities across EITs are very high ($r = .93$), suggesting excellent internal consistency for these typically quite brief tests. EIT scores also correlate substantially with criterion measures of L2 proficiency (average $r = .75$), including, in particular, self-assessments (average $r = .81$) and standardized proficiency assessments (average $r = .74$). Research has also indicated that, although other factors may influence EIT performance to some degree (e.g., working memory [Kim et al., 2016; Park et al., 2020]; educational background [Deygers, 2020]), EIT measures consistently demonstrate stronger relationships with criterion measures of L2 proficiency compared to these other factors.

Although research to date has suggested encouraging patterns of psychometric performance by EIT tests, a variety of design possibilities have been proposed that may influence the resulting effectiveness of EITs in estimating global L2 proficiency. One design factor has to do with whether test takers are asked simply to repeat each stimulus sentence or whether they are tasked with other activities as well, for example, answering a comprehension question about the stimulus before repeating it (e.g., Papadopoulou & Clahsen, 2003). Another design factor has to do with whether the stimulus sentences are grammatically correct or contain a mix of ungrammatical and grammatical exemplars (e.g., Erlam, 2006; Spada et al., 2015). Meta-analytic findings to date (Kostromitina & Plonsky, 2021) have indicated that both of these EIT

designs lead to a decrease in the average correlation with measures of global proficiency compared with EIT designs that simply require repetition of grammatically correct sentences.

Another design factor addresses the potential problem of test takers "parroting" the stimulus sentences without having to process them for meaning and form (Vinther, 2002). An early suggestion for resolving this problem (Ortega et al., 1999) was to introduce a pause of some duration between the playing of an audio stimulus and the beginning of the test taker response (e.g., as introduced by a tone sound after a few seconds, indicating that the test taker could repeat the stimulus). Meta-analytic findings to date (Kostromitina & Plonsky, 2021) indicate that EIT tests that adopt a pause between the stimulus and the response outperform those that allow immediate repetition (i.e., in terms of correlation with L2 proficiency criterion measures: $r = .76$ with a pause, $r = .63$ without a pause).

Final important design factors have to do with how many items should comprise an EIT test and how long the stimulus sentences should be. Early EIT designs were agnostic to varying stimulus length, focusing instead on the presentation of targeted grammatical forms within the tested sentences (e.g., Larsen-Freeman, 1975; Naiman, 1974). However, whereas the goal of the EIT is to predict global L2 proficiency, the tradition of practice since the early 2000s (e.g., Ortega, 2000) has been to compile stimulus sentences that gradually increase in length (as measured by number of syllables). A prototypical EIT design contains 20–30 sentences, with each sentence ranging from a low of 7 syllables to a high of 19 or so (e.g., Tracy-Ventura et al., 2014; Wu & Ortega, 2013). In general, meta-analytic findings (Kostromitina & Plonsky, 2021) have pointed to a moderate positive effect for both longer EIT tests (i.e., with more items) and tests with a wider range of stimulus length (i.e., from few to many syllables). However, the questions of ideal test length and range of stimulus length have not been investigated systematically to date and are likely confounded with issues like the linguistic characteristics of the target language, the population of test takers and their range of proficiency levels, and other factors (see discussion in Mozgalina, 2015).

On the basis of research to date, then, EIT tests show promise as efficient and reliable predictors of global L2 proficiency. Best practices in EIT design indicate that tests should (a) feature multiple stimuli of varying lengths; (b) include a pause between stimulus and response, but not any other activity (such as a comprehension question); and (c) incorporate only grammatical (not ungrammatical) stimulus sentences. Furthermore, meta-analytic findings and traditions of practice have indicated that EIT tests are most appropriately scored using a holistic scale and rubric that emphasize degree of accuracy in response to each stimulus (Kostromitina & Plonsky, 2021).

Although EIT tests offer a potentially useful option in efficiently measuring the global language proficiency of L2 learners, a handful of important issues merit consideration before this approach is adopted for operational test use. First, the format of the EIT is artificial when compared with real-world language tasks and likely to be unfamiliar to many test takers. Repeating multiple overheard sentences aloud is not a typical communication task, so test taker perceptions of and responses to the test format itself may affect how the test taker performs the task (see discussion in Norris, 2018). Second, the stimulus sentences on EIT tests are typically collections of random or otherwise unrelated ideas and topics. This approach simplifies the creation of items and facilitates automated generation of content, but it does not provide test takers with any purpose or context for listening and repeating (i.e., other than the purpose of the test itself; see Papageorgiou et al., 2021, for a discussion of the EIT design found in the *TOEFL® Essentials™* test). Third, listening and repeating multiple times can be challenging, especially for lower proficiency learners, thereby raising questions related to both test taker motivation and test taker fatigue, either of which may affect performance (see discussion in Bachman & Palmer, 2010). Last, depending on how the test is going to be used and what kinds of interpretations are intended, the EIT approach may or may not be warranted (Kane, 2013; Norris, 2008). In particular, considerations that should drive the design and operational implementation of EIT tests include whether the EIT is used in isolation or in combination with other test tasks, how test performance is interpreted in relation to ability to use the language for specific communication purposes, and what consequences may ensue from test use and associated decisions.

Development of an Innovative Elicited Imitation Task

With the EIT's robust history of use, promising psychometric qualities, and associated recommendations and caveats in mind, we developed a particular instantiation of the EIT format in the context of a larger test development initiative. This initiative focused on creating a new English language proficiency test within the TOEFL Family of Assessments, the TOEFL Essentials test. Overall, this test is intended to measure "foundational language skills and communication abilities in academic and general (daily life) contexts" (Papageorgiou et al., 2021, p. 2). A major goal of this new test was to offer an

assessment of English language proficiency that was efficient (in terms of test-taking time required) and rigorous (in terms of construct coverage and psychometric quality) and at the same time provided test takers with meaningful and engaging opportunities to display their language skills. Another important dimension of the new test was that it was intended to be used both for making admissions decisions in higher education contexts (see Xi & Norris, 2021) and for other possible decisions related to the identification of English learners' proficiency levels spanning the full range of abilities (i.e., from A1 to C2 on the Common European Framework of Reference scale; Council of Europe, 2001). In light of these goals and intended uses, the design of the TOEFL Essentials test called for a combination of authentic communication tasks across the four skills plus several "efficient" task types that would enable quick and trustworthy discrimination among test takers across the full range of proficiency levels.

The EIT task type was identified as a likely candidate of the efficient variety for the reasons outlined previously and in particular because it seemed capable of rapidly estimating the overall oral proficiency of test takers within the speaking modality. From the outset of test development, we adopted the aforementioned EIT best practices. First, the test included a variety of shorter to longer sentence stimuli, ranging from 6 to 29 syllables, with each iterative sentence containing a few syllables more than the previous. Second, each sentence adhered to standard grammatical norms of English and represented a variety of syntactic structures that typify longer and longer sentences. In addition, sentences excluded excessively long words (i.e., those of more than five syllables) as well as proper nouns, jargon, and highly infrequent vocabulary terms. Third, prior to repeating each sentence, test takers encountered a brief pause of 2 seconds, followed by a tone sound that indicated they should repeat. This last step was taken to minimize the possibility of a parroting effect.

In addition to these standard practices, we adopted several design innovations to provide test takers with a maximally meaningful opportunity to engage in the task. Namely, we designed the EIT as a conceptually coherent set of items (rather than as unrelated individual items), all of which were embedded within a simulated communication context that gave learners a purpose for listening and repeating. This approach provided a context (e.g., the test taker is being trained to lead a college campus tour), a reason for carefully listening and repeating (e.g., the test taker needs to practice exactly what to say at different points on the tour), and a visual support (e.g., the test taker sees a map of the campus on the computer screen, with points highlighted as the tour progresses). To our knowledge, this effort represented the first attempt to operationalize a contextualized version of the EIT in an assessment or elsewhere.

An early concern was whether enough scenarios and accompanying sets of graphics could be produced to supply adequate numbers of items for operational use. A variety of different scenarios have subsequently been created, and within a given scenario and accompanying set of graphics, we have found that large numbers of input sentences can be created by varying the length and propositional content of sentences associated with a specific part of the graphic. In the campus tour example, separate test forms may identify a particular location on the map as a different building, what is said about each building can vary, and so on.

The stimulus sentences were audio-recorded by professional voice actors; in the operational assessment, each set features a standard accent from the United States, the United Kingdom, Australia, or New Zealand, thereby increasing the range of English varieties represented on the test. The instructions, visual support, and audio stimuli were all packaged and delivered via a computerized environment such that the test proceeded as a seamless scenario within which the test taker was embedded. Last, the EIT was renamed as "Listen and Repeat" within the TOEFL Essentials test to provide a more meaningful name for test takers.

A final development step for the Listen and Repeat task addressed the need for scoring of test taker responses to each sentence. Following traditions of practice (e.g., Ortega et al., 1999; Tracy-Ventura et al., 2014), we adopted a rubric-based approach for initial scoring by human raters, with the overall goal of establishing a reliable rating process that could be generalized across items of differing length and composition. Initial development of a scoring rubric was carried out in concert with early-stage prototyping and was informed by evaluation criteria used in second language acquisition (SLA) research (Ortega, 2000). Scoring criteria primarily focused on the completeness and accuracy of repetition, with intelligibility also considered given that a determination of accuracy depended in part on being able to understand the response. Responses were scored on a 5-point scale ranging from 4, assigned to responses that fully and accurately repeated the stimulus, to 0, given to responses that contained only silence, were unintelligible, or contained no more than one correct content word.

This initial scoring rubric was revised over several iterations to produce the final operational rubric for the test. Major revisions included expanding the scale to 0 – 5 by splitting the zero category; in keeping with ETS practice in other speaking

Table 1 General Scoring Criteria Used in TOEFL Essentials for the Elicited Imitation Task

Score	Description
5	The response exactly repeats the prompt
4	The response captures the meaning expressed in the prompt, but it is not an exact repetition
3	The response is a full statement but does not accurately capture the original meaning
2	The response is missing a significant part of the prompt and/or is highly inaccurate
1	The response captures very little of the prompt or is largely unintelligible
0	There is no response, or the response is unintelligible, not in English, or off topic (unconnected to the prompt)

Note. Adapted from *TOEFL Essentials Test Speaking Scoring Guide* by ETS (2021), https://www.ets.org/s/toefl-essentials/rsc/pdf/speaking-rubric.pdf. Copyright 2021 Educational Testing Service.

tests, a score of zero was reserved for responses that did not contain any recognizable or appropriate content. Minimal responses that included recognizable content from the stimulus were awarded a score of 1, and scores for all other scoring categories were increased by 1 so that a fully accurate response was now awarded a score of 5. Specific scoring criteria were refined following piloting and field testing to clarify the degree of accuracy and completeness required for the score in question, and scoring rules were developed to deal with issues like self-repair while responding and test takers repeating the response more than once. The scoring rubric used in TOEFL Essentials is available online (ETS, 2021); the general criteria used for scoring are listed in Table 1.

The Current Study

This report describes a variety of investigations carried out on the EIT that were done during the development of the TOEFL Essentials test. Not including initial small-scale design and usability studies, development of the EIT proceeded in the three major phases of prototyping, piloting, and field testing (Kenyon & MacGregor, 2012; Nissan & Schedl, 2012). The prototyping phase focused on usability issues related to the EIT and the impact of specific design features. The current report describes findings from an experiment done at this stage to evaluate one of the major innovations of the EIT, the addition of visual input (research question [RQ] 2). The pilot phase focused on refinement of scoring, investigations of the psychometric qualities of the task, and further evaluation of selected usability issues. The current report focuses primarily on findings from research done during this pilot phase. The purpose of the field test was to evaluate the testing infrastructure to be used for the operational test, create and pretest items for operational use, and make final refinements to scoring materials. Work associated with the field test was mainly confirmatory in nature and so is not described in this report, where the emphasis is on test development.

Research Questions

Our goal was to investigate key issues in EIT design and function, with the findings intended to serve as evidence for specific claims (or warrants) listed in the TOEFL Essentials validity argument (Papageorgiou et al., 2021). The investigations addressed a variety of questions that were felt to be important for supporting the use of the EIT in the TOEFL Essentials test and, more broadly, for making inferences regarding English language proficiency. The RQs investigated are listed here, organized by the associated warrants and inferences (in parentheses):

- *Warrant*: Test tasks measure relevant aspects of language proficiency (domain description).
    - RQ1: Do test takers perceive the EIT as a valid measure of their speaking ability?
- *Warrant*: Task administration conditions are appropriate for providing evidence of targeted language abilities (evaluation).
    - RQ2: Does the presence of animation impact task performance?
    - RQ3: Do production features (topic, speaker accent, type of visual) influence performance?
    - RQ4: Are test takers able to produce a response in the allowed time frame?
- *Warrant*: Scores for constructed-response tasks reflect the targeted language abilities and skills (evaluation).
    - RQ5: Were raters able to consistently score responses?

- *Warrant*: Test tasks distinguish among examinees with varying degrees of proficiency (evaluation).
    - RQ6: Do EIT scores separate test takers into distinct levels of proficiency?
- *Warrant*: A sufficient number of tasks are included on the test to provide stable estimates of test takers' performance (generalization).
    - RQ7: What is the optimal number and composition of items for the EIT?
- *Warrant*: Task features impact performance in expected ways (evaluation).
    - RQ8: Does EIT item difficulty increase as stimulus length increases?
- *Warrant*: The internal structure of the test scores is consistent with a theoretical view of language proficiency as a number of highly interrelated components (explanation).
    - RQ9: Are EIT scores associated with measures of related abilities?
- *Warrant*: Performance on the test measures relates to performance on other test-based measures of language proficiency, as expected theoretically (explanation).
    - RQ10: Are EIT scores associated with a global measure of proficiency (C-test)?

Methodology

Participants

As mentioned previously, the EIT was investigated as part of the prototyping and pilot phases of the development of the TOEFL Essentials test. Participants in each phase are described in the following sections.

Prototyping Study

Although most of the RQs were investigated in the pilot study, RQ2 (effect of animation) was evaluated in the prototyping phase, which was conducted from November 2019 to March 2020 at three large universities, one each in Colombia, Turkey, and the United States. Only students from Colombia and Turkey participated in the animation-related experiment. Participants were recruited from English preparatory courses by a local coordinator in each location and paid US$25–30 for participating in a data collection session of 60–90 minutes. All participants in Colombia reported Spanish as their first language, and nearly all individuals in Turkey indicated a first language of Turkish (Table 2). Participants were roughly equally divided by gender (45% female, 55% male). To the extent possible, local coordinators recruited participants across a broad proficiency range, and participants demonstrated a wide range of proficiency, as indicated by C-test scores (Figure 1).

Pilot Study

The pilot study was conducted from July to August 2020 with 701 participants in 10 countries. As with the prototyping study, local coordinators in each country recruited college-aged participants, who were paid approximately US $30–50, depending on the country, for participating in a session lasting approximately 2 hours. Participant demographic data are reported in Table 2. Effort was made to recruit participants who would reflect the anticipated candidature for the TOEFL Essentials test, resulting in an adequate targeted sample of countries and language groups. However, recruiting low-proficiency participants was a challenge, and compared to the prototyping study, the proficiency distribution was shifted somewhat toward the higher end of the scale (Figure 2). Nonetheless, pilot participants showed a broad range of English proficiency in terms of C-test scores.

Materials and Data Collection Procedures

The basic design of the EIT was described earlier and is also documented in Papageorgiou et al. (2021). Two sets of EIT items were used in both prototyping and piloting; one set was based on a scenario where the participant was a university student learning what to say while taking visitors on a tour of the campus (henceforth the "Campus Tour" set), while in the other set, the participant played a student who was working with a teacher to practice useful phrases for giving oral

Table 2 Demographic Information for Study Participants

	Prototyping (RQ2), *n* (%)	Pilot (RQ1, 3–10), *n* (%)
Total participants	419	701
Country		
Colombia	233 (55.6)	
Turkey	186 (44.4)	26 (3.7)
PR China	—	245 (35.0)
Korea	—	90 (12.8)
United States	—	70 (10.0)
Japan	—	69 (9.8)
Mexico	—	66 (9.4)
Brazil	—	41 (5.8)
Germany	—	35 (5.0)
India	—	34 (4.9)
Jordan	—	25 (3.6)
Gender		
Female	190 (45.3)	475 (67.8)
Male	229 (54.7)	221 (31.5)
Other responses	—	5 (0.7)
Age (years)		
<18	50 (11.9)	9 (1.3)
18–22	345 (82.3)	369 (52.6)
23–30	22 (5.3)	234 (33.4)
31–40	0 (0.0)	67 (9.6)
>40	2 (0.5)	22 (3.1)
Educational level		
Secondary school	7 (1.7)	32 (4.6)
Undergraduate student	401 (95.7)	386 (55.1)
Bachelor's degree	4 (1.0)	137 (19.5)
Graduate student	4 (1.0)	92 (13.1)
Graduate degree	3 (0.7)	43 (6.1)
None of the above	—	11 (1.6)
Years of English study		
≤2	93 (22.2)	49 (7.0)
3–5	124 (29.6)	73 (10.4)
6–8	87 (20.8)	133 (19.0)
9–11	64 (15.3)	192 (27.4)
≥12	51 (12.2)	254 (36.2)
First language		
Chinese	—	288 (41.1)
Spanish	233 (55.6)	66 (9.4)
Turkish	184 (44.4)	9 (1.3)
Korean	—	90 (12.8)
Japanese	—	60 (8.6)
Portuguese	—	41 (5.8)
Hindi	—	33 (4.7)
Arabic	1 (0.2)	31 (4.4)
German	1 (0.2)	29 (4.1)
French	—	15 (2.1)
Other[a]	—	27 (3.9)
Invalid[b]	—	12 (1.7)

[a] Languages mentioned by three or fewer participants: Amharic, Bantu, Bengali, Burmese, Creole, English, Indonesian, Malayalam, Maya, Nepali, Polish, Punjabi, Russian, Somali, Thai, Ukrainian, Wolof. [b] The online system returned a result of "invalid" for these participants, for unknown reasons.

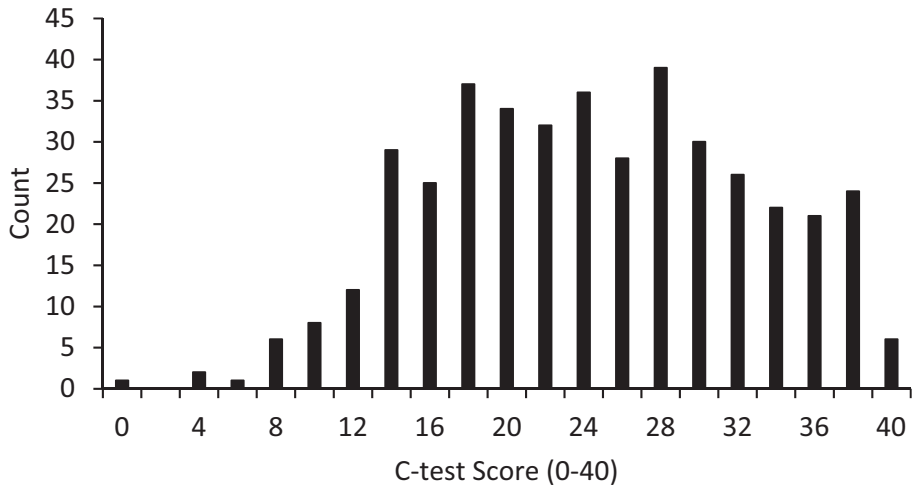


Figure 1 English proficiency level of prototyping study participants included in RQ2 ($N = 419$).

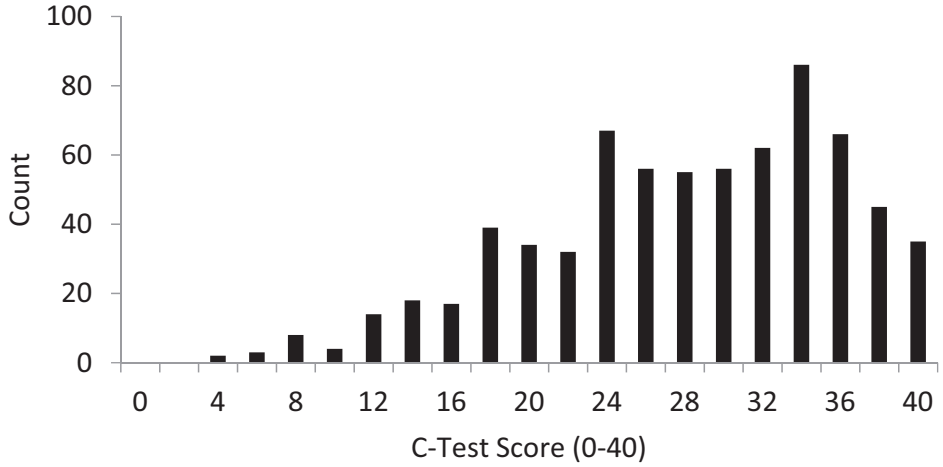


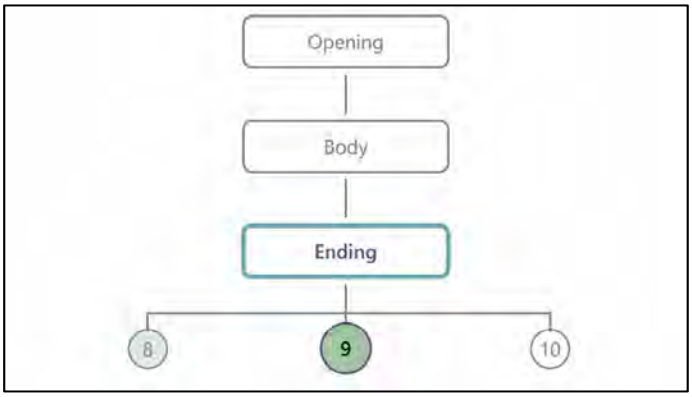
Figure 2 English proficiency level of all pilot study participants ($N = 699$). *Note*. C-test scores were missing for two participants.

presentations (henceforth the "Presentations" set). Sets of 10 sentences each were used in prototyping and then expanded to 15 sentences for the pilot to allow further evaluation of the appropriate number and length of sentences for the operational test (Appendix A). For the prototyping study, the stimuli for both sets were recorded by a female L1 speaker of U.S. English, while in the pilot, the Campus Tour set was recorded by a male L1 speaker of U.S. English and the Presentations set was recorded by a female L1 speaker of British English.

Audio-recorded input sentences were accompanied by a series of graphics that provided an animated and contextualized representation of the participant's progress through the task (Figure 3). To answer RQ2, a subset of participants in the prototyping study completed one set in the usual "animated" version of the task, followed by a version of the other set where a simplified and decontextualized image was presented. Two randomly selected groups of participants completed counterbalanced combinations of set and delivery version, that is (Campus Tour + animated) followed by (Presentations + simplified) or, alternately, (Campus Tour + simplified) followed by (Presentations + animated).

In addition to the EIT, the participants completed a C-test as a separate measure of global English language proficiency, consisting of two passages of 20 blanks each, with each blank scored correct or incorrect to produce a total score of 0–40. Other materials used in the study included a consent form, a background questionnaire (Appendix B), and user perception questions, all translated into the primary local language for each data collection location (Spanish or Turkish). In addition, data collection sessions included a variety of other assessment tasks that were also being evaluated. The EIT and other materials were delivered using an online platform developed for the study. A unique link was

Animated version of the Campus Tour set; view when completing Item 9 (of 10)



Animated version of the Presentations set; view when completing Item 9 (of 10)

Simplified version for both sets, showing the image displayed for all items

Figure 3 Screenshots of visual input during animated and simplified versions of the elicited imitation task used in prototyping. *Note*. For the animated version of Campus Tour, name labels (e.g., "Music Building") were used in the prototyping study but were discarded afterward to reduce the potential for distraction. For the simplified version, the headset image was colored green during the response time and was colored gray otherwise.

provided to each participant; clicking on this link opened the test content in the participant's Web browser. Participants were required to complete microphone and audio checks at the start of the session, and sample tests were checked by local coordinators prior to data collection to confirm that test content functioned as intended where the test would be administered.

In the prototyping study, nearly all participants completed the session in groups within a university computer lab and were supervised by lab staff. For the pilot, participants generally accessed the materials independently, given that most students were on summer break and face-to-face instructional activities had been disrupted by the COVID-19 pandemic. To support at-home delivery, the online research platform was modified for improved functionality in situations of limited Internet bandwidth, and prior to their session, pilot participants were required to pretest their systems by logging in to the platform and completing a system check. (Participants also filled out a consent form and a background questionnaire at this stage.) Nonetheless, there were instances in which one or more responses to the EIT were either not recorded or unintelligible due to audio problems. In the pilot, full sets of 15 responses were obtained from 621 participants, or 89% of the total. Of the remaining 80 participants, 49 were from China and 10 were from Mexico; other countries having 5 or fewer individuals with missing responses included Brazil, India, Japan, Jordan, Korea, Turkey, and the United States. In all but 9 cases, at least one response was captured by the online research platform, and 46 instances (58%) involved the loss of 5 or fewer responses out of the 15 responses elicited.

Scoring of Elicited Imitation Task Responses and Data Analyses

Responses to the EIT were scored using the scoring rubric, as described previously. For the prototyping study, responses were double-scored by four ETS researchers who were familiar with the EIT; prior to scoring, the group collectively reviewed the rubric and a sample of responses to standardize their perceptions. For the pilot study, responses were double-scored by a group of 20 ETS raters and scoring leaders who had experience scoring the TOEFL iBT speaking test. The raters and scoring leaders completed an online video-mediated group training session in which the EIT and scoring rubric were introduced and exemplar responses and topic-specific scoring notes were reviewed. This was followed by scoring of a practice set of responses, where the group could compare and discuss their scores. Procedures for scoring responses were also discussed, and the training session was recorded for later review, as desired. Scoring was done using an online system developed for research purposes that is similar in functionality to the system used for operational scoring of the TOEFL iBT test and other ETS assessments. Raters could replay a response as many times as desired and, in addition to scores, could input comments or indicate one of several diagnostic codes to categorize responses that were awarded a score of 0 or were unscorable because of technical problems. Raters worked individually, with ETS R&D staff available to answer questions.

The combination of scoring and other data collection methods resulted in a variety of data sources. In keeping with the exploratory nature of the study, analyses of the data consisted primarily of various types of descriptive and relational statistics. Specific analyses are described in the results section for each RQ.

Results

RQ1: Do Test Takers Perceive the Elicited Imitation Task as a Valid Measure of Their Speaking Ability?

In the pilot study, immediately following the EIT, participants were asked the question "How well does this task show your speaking ability?" A majority (60%) of test takers felt that the EIT was a "very good" or "good" example of their ability (Figure 4); we consider this to be an encouraging result given the relatively abstract nature of the EIT and the likelihood that it was unfamiliar to most participants. However, we note that the frequency of positive responses was lower than for a simulated interview task also administered in the pilot, where the approval rate was 92% (Figure 4). This difference is not unexpected given that the Virtual Interview consists of questions related to the test taker's personal experiences and opinions and is more obviously communicative in nature. Overall, these findings provide support for combining the EIT with tasks that elicit spontaneous meaning-focused speech, as is done with the TOEFL Essentials test.

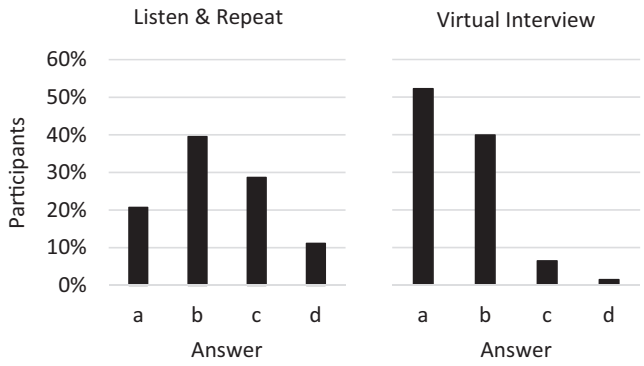


Figure 4 Test taker answers to the question "How well does this task show your speaking ability?" ($N = 699$). *Note*. a = It's a very good example of what I can do. b = It's a good example of what I can do. c = It's a poor example of what I can do. d = It's a very poor example of what I can do.

Table 3 Mean Score on the Elicited Imitation Task Across Task and Administration Conditions

	Campus Tour (presented first)				Presentations (presented second)			
	Condition	*n*	Mean	*SD*	Condition	*n*	Mean	*SD*
Group 1	Animated	201	16.3	6.91	Simplified	201	17.5	6.86
Group 2	Simplified	218	18.7	7.98	Animated	218	19.4	7.78

Note. Responses were scored on a 0–4 scale; total scores ($k = 10$ items) are on a 0–40 scale.

RQ2: Does the Presence of Animation Impact Task Performance?

Exact repetition of the stimulus sentence requires mental focus while listening to and processing the input, and a concern in the development of the EIT was that the visual content accompanying the task might be a distraction. As mentioned, in the prototyping phase we conducted an experiment in which participants completed tasks delivered with both the full "animated" graphics and a simplified graphic. We found little difference in mean scores for tasks delivered with either animated or simplified visual input (Table 3). Differences in mean scores across animated and simplified conditions were 2.4 points (Campus Tour set) and 1.9 points (Presentations set) on a 40-point scale; as a practical matter, such differences would be expected to have minimal impact on overall test results. Nonetheless, mean scores within experimental groups were statistically significantly different as determined using a paired t test (Group 1, $t(200) = -5.761$, $p < .01$; Group 2, $t(217) = 2.902$, $p < .01$; p-values adjusted with a Bonferroni correction for multiple comparisons). In both cases, the mean score was slightly higher for the task presented second, which may suggest an influence of practice, but the effect sizes were small (Group 1, Cohen's $d = -0.18$; Group 2, Cohen's $d = 0.08$). Pearson correlations between sets of scores for each version, within groups, were also high (Group 1, $r = .90$, $r^2 = .81$; Group 2, $r = .92$, $r^2 = .84$), suggesting that different combinations of task and condition largely captured the same variance.

We also examined whether any effect of animation might be conditional on the proficiency of the test taker. In particular, we were concerned that any distraction associated with the visuals might have a larger impact on less-proficient individuals, who likely experience a greater cognitive load in performing the task. We did not observe evidence of any such effect. When participants were divided into quartiles based on C-test scores, very similar patterns in mean scores were seen across the different versions of the EIT (Figure 5). Performance on the EIT improved with increasing general proficiency, as would be expected, and within-group mean differences ranged from 0.2 to 1.4 score points across the proficiency spectrum.

Immediately following completion of each EIT, participants were asked to give their opinions regarding the difficulty of the task and their level of engagement; we found that perceptions were similar across task versions (Table 4). In all cases, over 90% of individuals responded that the EIT was "somewhat difficult" or "very difficult," with little difference across task version or sets of items. In terms of engagement, 70% of Group 1 test takers responded that the animated Campus Tour task was "very" or "somewhat" interesting, while the interest rating in other conditions was 54%–57%. The reason for this difference is unclear, but the graphics used for the Campus Tour set were somewhat more elaborate compared to

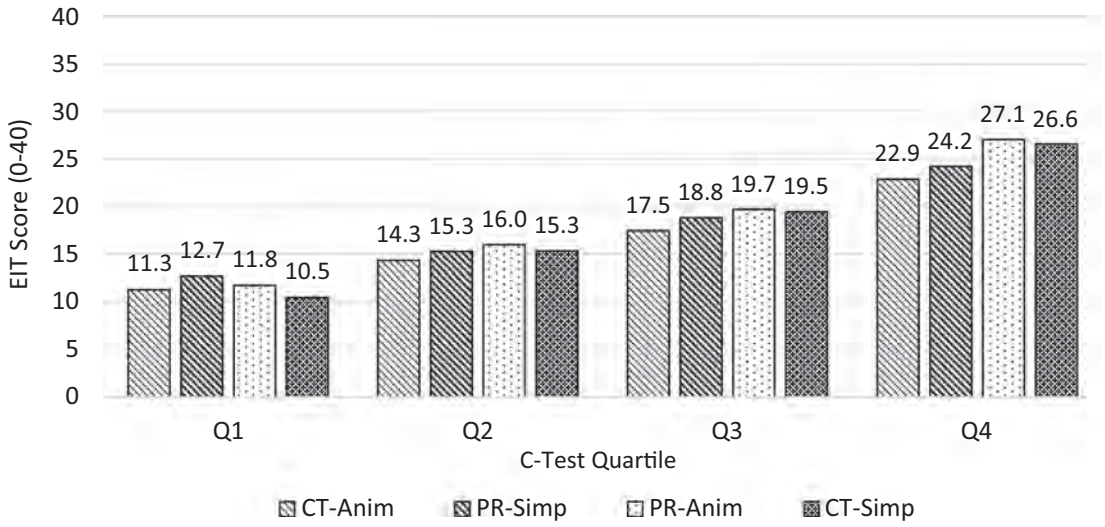


Figure 5 Mean score on the prototype elicited imitation task ($k = 10$) across task and administration conditions, by proficiency level. *Note*. The first two bars in each quartile group correspond to test takers in Group 1 ($n = 201$); the third and fourth bars represent test takers in Group 2 ($n = 218$). C-test quartile boundaries were determined using all participants in the prototyping study ($N = 571$). CT = Campus Tour set. PR = Presentations set. Anim = animated version. Simp = simplified version.

Table 4 Perceptions of Test Takers Toward Different Task and Administration Conditions

	Group 1		Group 2	
Presentation order	First	Second	First	Second
Set	Campus Tour	Presentations	Campus Tour	Presentations
Version	Animated	Simplified	Simplified	Animated
How challenging was this task for you?, *n* (%)				
Very easy	0 (0)	1 (1)	1 (0)	2 (1)
Somewhat easy	6 (3)	11 (6)	2 (1)	17 (8)
Somewhat difficult	85 (43)	80 (40)	67 (31)	93 (43)
Very difficult	108 (54)	107 (54)	148 (68)	106 (49)
How do you feel about doing this task?, *n* (%)				
Very interesting	42 (21)	31 (16)	44 (20)	45 (21)
Somewhat interesting	98 (49)	77 (39)	79 (36)	79 (36)
Somewhat boring	40 (20)	59 (30)	62 (28)	65 (30)
Very boring	19 (10)	32 (16)	33 (15)	29 (13)

Note. Group 1, $n = 201$; Group 2, $n = 218$.

those used for the Presentations set (Figure 1), so perhaps the Campus Tour set benefited more from the use of animation. In any case, the animation condition appeared to have little consistent effect on perceptions of task difficulty or interest. Animated graphics were retained as a feature of the operational test, however, given that the graphics provide additional context for the task and also serve as a visual indicator of the test taker's progress through the items.

RQ3: Do Production Features (Topic, Speaker Accent, Type of Visual) Influence Performance?

A requirement for the TOEFL Essentials test was that major varieties of English be represented in the spoken input provided to test takers. In the pilot, test sentences used in the Campus Tour set were delivered by a male speaker of U.S. English, while the prompts for the Presentations set were recorded by a female speaker of British English. The impact of accent on performance on the EIT was a potential concern given that comprehension of the input is required for successful repetition. We compared mean task-level scores (15 items, 0–75 scale) for these two conditions and found no difference, independent two-tailed *t* test, $t(581) = .866$, $p = .78$ (Table 5). This is only a preliminary finding given the uncontrolled nature of the comparison: Accent was confounded with the task content, which differed in both input sentences and

Table 5 Average Task-Level Score (0–75) for the Two Versions of the Task Used in the Pilot Study

Task	Accent	Topic	Visual input	*n*	Mean	*SD*
Set 1	U.S. male	Campus Tour	Campus map	297	49.61	12.31
Set 2	U.K. female	Presentations	Graphic organizer	286	48.72	12.48

Note. Individuals with missing scores for one or more items were excluded.

supporting graphics. However, we note that both speakers used in the pilot spoke a standard version of their respective English variety, and studies of listening assessment in a similar context found no impact of such accents on listening comprehension (Ockey et al., 2016; Ockey & French, 2016).

RQ4: Are Test Takers Able to Produce a Response in the Allowed Time Frame?

One challenge in the design of the EIT was determining an appropriate length of time to allow for making a response. Adequate time to repeat the stimulus sentence was required, but if too much time is provided, then some test takers may repeat the sentence one or more times, a behavior that complicates scoring of the response and interpretation of scores. Additionally, the response time must increase as the sentences grow longer throughout the task. In early prototyping of the EIT, we found that a reasonable response time was approximately double to triple the length of the time used to speak the stimulus sentence in the audio input.

In the pilot study, data to confirm this specification were obtained by providing participants with a "next" button that allowed them to advance to the next item, if desired. Participants were clearly instructed to repeat the input sentence only once and that if time remained after repeating the input, they could use the "next" button to go on. We then evaluated the appropriateness of response times by measuring the length of the audio file of the response captured by the system: An audio file length less than the maximum allowable response time was taken as evidence that the test taker had chosen to go on. Length of the audio file was also used as an indicator of how much time participants were actually taking to make responses.

We found that 63%–76% of participants chose to advance to the next item before the response time was expired, depending on the item (Table 6). Average response time was typically 2–3 seconds shorter than the maximum time allowed. We also found no consistent relationship between the length of the stimulus sentence and the percentage of individuals choosing to advance. The results suggest that the response time provided was more than adequate for at least two-thirds of participants to make a response. Moreover, this figure does not include individuals who completed repetition but chose to wait out the remaining few seconds of response time.

Although the results overall appeared to confirm the appropriateness of the response times used, there was still concern that participants of lower language proficiency might struggle to finish in time. Accordingly, we separately analyzed the behavior of lower proficiency individuals (as indicated by C-test score quartile) and found their behavior to be very similar to the overall pool of participants (Table 6). Lower proficient individuals may even have been slightly more likely to go on when encountering the very longest sentences, 20 syllables or more in length. We also anecdotally observed that low-performing individuals who were not able complete a full repetition, a common situation for longer sentences, tended to have either skipped content in the middle of the sentence or fallen silent by the end of the response time; there was little indication that extra time would have elicited a full repetition for these low-proficiency participants. Furthermore, the probability of making an accurate repetition likely declines as the interim between stimulus and response grows, making it unlikely that an otherwise high-quality response will be produced late and then be cut off.

In addition to the appropriateness of response times, we wanted to confirm that participants did not disadvantage themselves by using the "next" button. To do this, we compared the scores received for responses where "next" was used or not used. We observed little difference in mean scores associated with use of "next." Differences in mean scores ranged from −0.23 to 0.47 score points (on a 0–5 scale), with the corresponding effect ranging from low to medium in size (Cohen's *d* values of −0.27 to 0.46; Table 7). Overall, these results suggest that the choice to advance to the next screen most likely reflected a situation where the test taker had made what they felt to be a satisfactory response and had decided to go on.

Table 6 Proportion of Participants Choosing to Advance to the Next Item, and Average Response Times

		All test takers			C-test Quartile 1		
No. of syllables	Max. time (sec.)	No. of responses	*M* response time (sec.)	Less than max. (%)	No. of responses	*M* response time (sec.)	Less than max. (%)
6	7	337	5.4	63	79	5.5	62
7	7	340	4.9	74	80	5.0	74
8	7	695	5.1	72	171	5.1	71
9	7	1,050	4.8	76	261	4.9	74
10	7	354	5.1	74	90	5.3	69
11	7	690	5.4	67	169	5.5	66
13	9	1,041	6.3	74	258	6.3	74
14	9	334	6.7	69	78	6.9	65
15	9	1,363	6.7	71	335	6.6	72
16	9	331	6.8	68	77	6.7	71
17	9	350	7.0	67	89	6.9	64
18	11	676	8.1	72	164	7.8	73
19	11	328	8.1	70	76	8.0	72
20	11	1,024	8.2	72	253	8.0	74
22	14	326	10.1	75	76	9.9	76
23	14	325	10.1	75	75	10.1	72
24	14	344	10.5	72	87	10.1	75
29	14	342	10.5	73	87	10.2	76

Note. Responses include audio files that were unscorable, so *n* sizes are somewhat larger than reported in Table 7.

Table 7 Average Scores When the Test Taker Elected to Advance

			Resp. time < max.			Resp. time = max.				
No. of syllables	Max. time (sec.)	Item *n*	*n*	*M*	*SD*	*n*	*M*	*SD*	Mean diff.	Cohen's *d*
6	7	337	212	4.42	1.47	125	4.33	1.57	0.08	0.08
7	7	340	252	4.35	1.44	88	4.52	1.15	−0.17	−0.19
8	7	695	500	4.35	1.34	195	4.23	1.30	0.11	0.12
9	7	1,050	799	4.06	1.39	251	3.98	1.38	0.07	0.08
10	7	354	263	3.85	1.27	91	3.54	1.42	0.30	0.32
11	7	690	462	3.65	1.45	228	3.19	1.38	0.47	0.46
13	9	1,041	773	3.40	1.49	268	3.16	1.41	0.24	0.23
14	9	334	230	3.07	1.49	104	3.01	1.28	0.06	0.06
15	9	1,363	968	2.86	1.43	395	2.67	1.32	0.19	0.20
16	9	331	224	2.77	1.37	107	2.67	1.24	0.10	0.11
17	9	350	235	2.77	1.34	115	2.62	1.27	0.16	0.17
18	11	676	487	2.18	1.18	189	2.19	1.18	−0.01	−0.01
19	11	328	228	1.90	1.17	100	1.89	1.07	0.01	0.02
20	11	1,024	736	2.61	1.29	288	2.53	1.32	0.08	0.09
22	14	326	243	2.26	1.23	83	2.44	1.24	−0.18	−0.21
23	14	325	243	2.00	1.29	82	2.23	1.07	−0.23	−0.27
24	14	344	247	2.30	1.02	97	2.06	1.09	0.24	0.32
29	14	342	251	2.42	1.10	91	2.46	0.95	−0.04	−0.06

RQ5: Were Raters Able to Consistently Score Responses?

One goal of the pilot study was to refine the scoring rubric developed in the prototyping phase and evaluate the ability of raters to use the scoring materials to produce consistent scores. Accordingly, all responses were double-scored, and we generated a number of different rater agreement indices. Rater agreement was generally high; raters showed exact agreement for 67.5% of all responses scored ($N = 9{,}507$). Adjacent agreement (no more than 1 score point difference) was seen in an additional 30.8% of cases, and raters disagreed by 2 points or more in 1.7% of cases. For all responses, the overall interrater Spearman rho correlation was .91, indicating that raters were very similar in ranking performances. Table 8 is a confusion matrix that tabulates instances of agreement and disagreement at the response level.

Table 8 Confusion Matrix for Item-Level Interrater Agreement

Rater 1 score	Rater 2 score					
	0	1	2	3	4	5
0	223	44	2	1	0	0
1	35	567	323	28	3	2
2	2	349	942	380	18	4
3	0	39	451	1,168	387	13
4	0	1	26	436	1,166	241
5	0	0	3	19	284	2,350

Note. A total of 9,507 responses from 701 test takers were scored. Shaded values indicate instances of exact agreement between raters.

Table 9 Rater Agreement Versus Length of Stimulus Sentence

Syllables	*n*	ρ[a]	Agreement (Rater 1 – Rater 2), % (*n*)		
			0	±1	±2 or more
6	300	0.65	92.7 (278)	7.0 (21)	0.3 (1)
7	309	0.73	86.4 (267)	12.6 (39)	1.0 (3)
8	641	0.76	81.1 (520)	16.5 (106)	2.3 (15)
9	963	0.81	74.1 (714)	23.7 (228)	2.2 (21)
10	325	0.85	76.0 (247)	23.7 (77)	0.3 (1)
11	648	0.88	67.4 (437)	30.1 (195)	2.5 (16)
13	962	0.88	65.4 (629)	32.4 (312)	2.2 (21)
14	306	0.87	63.1 (193)	35.0 (107)	2.0 (6)
15	1,274	0.88	61.5 (784)	36.8 (469)	1.6 (21)
16	306	0.84	62.7 (192)	35.9 (110)	1.3 (4)
17	326	0.86	62.9 (205)	35.6 (116)	1.5 (5)
18	633	0.82	61.0 (386)	37.0 (234)	2.1 (13)
19	308	0.83	64.3 (198)	34.7 (107)	1.0 (3)
20	951	0.85	61.7 (587)	37.2 (354)	1.1 (10)
22	303	0.86	68.0 (206)	30.7 (93)	1.3 (4)
23	304	0.83	57.2 (174)	41.4 (126)	1.3 (4)
24	327	0.76	58.7 (192)	40.1 (131)	1.2 (4)
29	321	0.75	64.5 (207)	32.7 (105)	2.8 (9)

Note. Only responses with a valid score from both raters are included in the analysis. [a]Rater 1 versus Rater 2 correlation (Spearman rho).

Rater agreement was greatest for the shortest items (Table 9). Average scores for these items were close to the maximum (5 points), suggesting that many test takers were able to repeat the stimuli flawlessly, simplifying the rater's decision task. Agreement generally decreased as item length increased, reflecting the fact that longer sentences provide more opportunities for various sorts of inaccuracies in the response, requiring a greater degree of interpretation in applying the scoring criteria. However, exact agreement never fell below 57% (Table 9). Interrater correlations (Spearman rho) were also generally above .80, except for the shortest and longest input sentences, where correlations were .65–.76, likely due to restricted distribution of scores (i.e., a large proportion of high scores for short sentences and low scores for long sentences). Overall, the results suggest that raters were able to score the full set of EIT responses with a high degree of consistency.

RQ6: Do Elicited Imitation Task Scores Separate Test Takers Into Distinct Levels of Proficiency?

To examine the extent to which the EIT items as a group were able to separate test takers into broadly differing levels of ability, total scores were calculated for each test taker on the 15 items completed during the pilot test (for a total possible score of 75 points). Overall, the average score ($M = 49.17$) showed that the group of participants was capable of performing relatively well on the set of items, with average scores well above the midpoint in the range of possible scores. The standard deviation value ($SD = 12.39$) also indicated considerable variability in scores, providing an initial gauge of the extent to which the EIT was able to spread out test takers. The histogram in Figure 6 shows the distribution of scores for $N = 583$

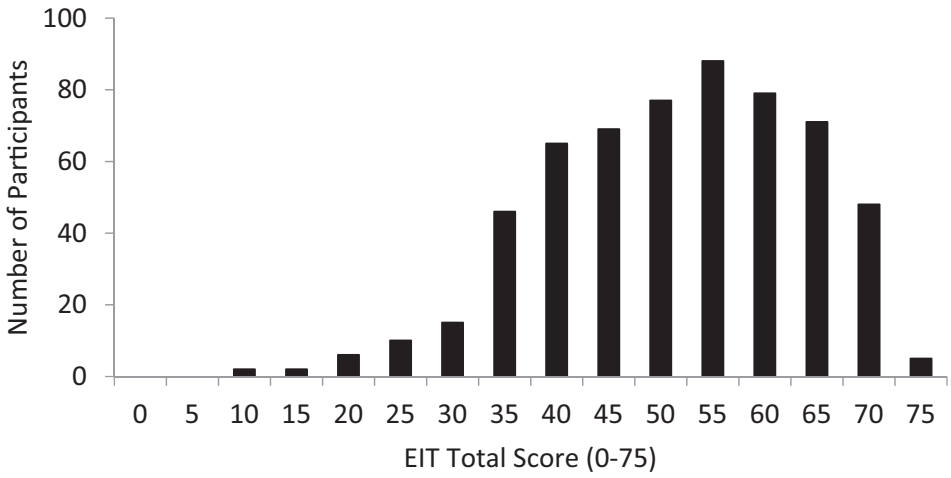


Figure 6 Distribution of total task scores ($k = 15$ items) for the pilot data ($N = 583$ test takers).

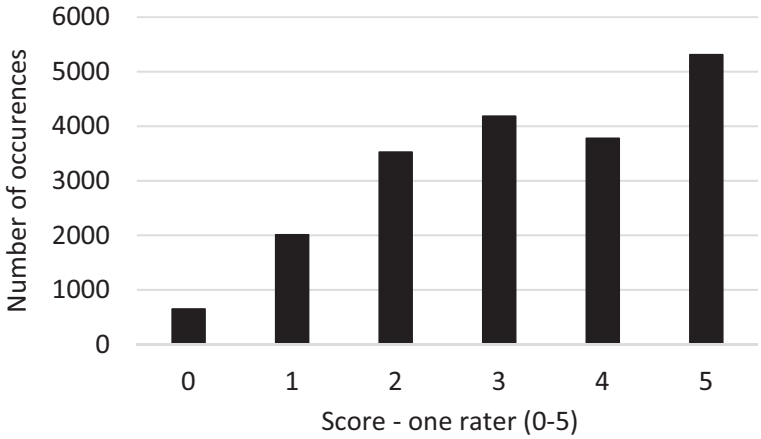


Figure 7 Distribution of raw scores for individual EIT responses. For double-scored responses, scores from each rater were tabulated separately.

pilot test takers, with total scores ranging from a low of 6 points to a high of 73 points. Although the distribution is apparently somewhat negatively skewed (skewness = −.421), this pattern likely reflects the prevalence of somewhat less difficult items in the 15-item sets experienced by participants. Note also that there is ample space on either side of the mean to accommodate more than 2 standard deviations' worth of scores, suggesting that the EIT in general can separate test takers into higher and lower ability across a broad range. Additional evidence for the wide and systematic spread of scores can be seen in the very high Cronbach's alpha reliability estimate ($\alpha = .93$).

The scoring patterns of individual raters also revealed the extent to which the full rating scale was utilized to score pilot test taker performances. Figure 7 shows the distribution of individual item responses that were scored at each of the 6 score points on the EIT rating scale. It is apparent that raters utilized rating scale points of 2–5 substantially, and to roughly equivalent degrees for scale points 2–4 (between 3,500 and 4,000+ responses per scale point). Scale point 5 was clearly the most used category, reflecting the prevalence of somewhat easier items in the beginning part of the test. Thus these four scale points were deemed useful by raters to capture the varying levels of performance test takers exhibited, indicating the clear need to differentiate among them. In other words, test takers differed broadly in how they performed on the items to the extent that raters required multiple scale points to account for their differing levels of performance. Rating scale points of 1 and 0 were utilized to a lesser, but nevertheless still substantial, degree, again indicating that at least some test taker performances required these lowest points to capture their abilities at repeating some of the items.

Finally, Figure 8 shows the distribution of performance ratings on the EIT for participants grouped by quartile according to their C-test proficiency scores. Note that the distribution curves and mean values are clearly distinct across the

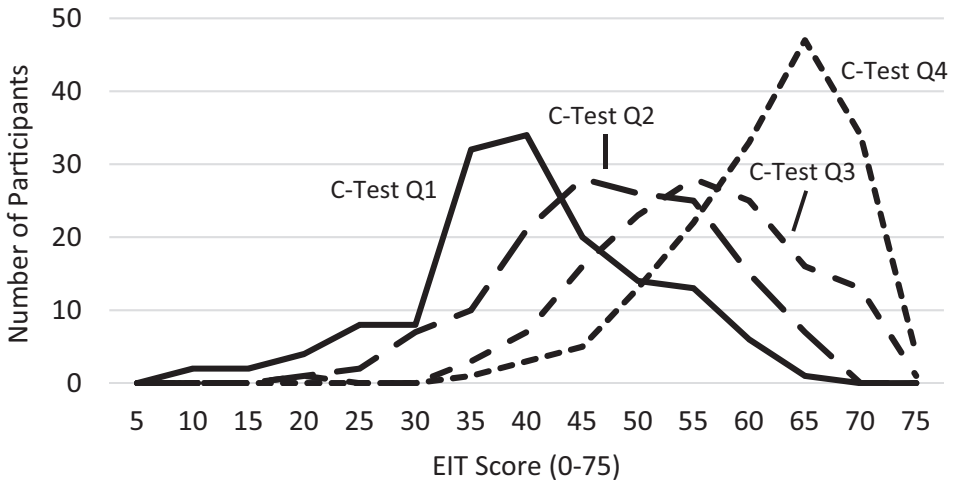


Figure 8 Distribution of elicited imitation task scores conditional on proficiency level (C-test quartiles).

four quartiles, providing a good indication of the extent to which the EIT is capable of distinguishing among learners at distinct English proficiency levels as measured by a criterion test (see further discussion in RQ10 results).

RQ7: What Is the Optimal Number and Composition of Items for the Elicited Imitation Task?

A key question in the design of the EIT was the number and length of input sentences needed to optimize the reliability of the EIT while minimizing administration time and the resources needed to score responses. To investigate this question, a simulation study was conducted using data from 491 pilot participants for whom a full set of 15 scored responses were available. Items were divided into "easy," "medium," and "difficult" categories based on the number of syllables in the input, and various combinations of five to eight items were used to compute a Cronbach's alpha value (Table 10). A total of 1,000 replications were conducted for each combination, with data from 90% of participants selected each time; separate analyses were conducted for each of the two sets used in the pilot study. Table 10 shows descriptive statistics computed from the 1,000 alpha values generated.

All combinations of five, six, seven, or eight items were highly reliable, with average reliabilities greater than .80. For the design eventually selected for the test (seven scored items: two easy, three medium, two difficult), average alpha values were .875 (Campus Tour set) and .879 (Presentations set) and ranged from .818 to .921. As expected, combinations with more items were more reliable; additionally, combinations with a greater proportion of medium-difficulty items were generally more reliable, but only by a small margin. The combination of two easy, three medium, and two difficult items was selected for the operational test on the basis of both high reliability and the potential to provide information across a broad range of language proficiency.

RQ8: Does Elicited Imitation Task Item Difficulty Increase as Stimulus Length Increases?

A key design assumption for the EIT is that the difficulty of individual items will increase as the number of syllables increases and the stimulus sentences become longer. In the current investigation, difficulty was operationalized as the average test taker score for a given item on the 6-point rating scale. Figure 9 shows the relationship between item length (x-axis) and average score on the item (y-axis). The strong negative relationship between number of syllables and average score is obvious, with scores decreasing steadily as syllable length increases. Interestingly, the difficulty of items seemed to level out once a length of approximately 19 syllables was attained, with subsequent items of increasing syllable length (up to 29) not showing additional decreases in average scores. Pearson correlations comparing syllable length with average score confirmed the very strong relationship across all items ($r = .88$) and in particular for items between 6 and 19 syllables in length ($r = .96$).

Turning to a more detailed examination of performance in relation to syllable length, Table 11 shows the average scores and percentages of ratings at each point on the rating scale for items of increasing stimulus length (from 6 to 29). It is

Table 10 Task-Level Reliability of Different Combinations of Items, by Set

	Item combination					Cronbach's alpha	
No. of items	Easy	Medium	Difficult	Mean	*SD*	Min.	Max.
Campus Tour set[a]							
5	1	2	2	0.838	0.031	0.741	0.904
6	2	2	2	0.847	0.025	0.763	0.901
	1	3	2	0.870	0.021	0.816	0.918
7	2	3	2	0.875	0.018	0.818	0.921
	1	3	3	0.889	0.016	0.836	0.922
	1	4	2	0.893	0.015	0.840	0.927
8	2	3	3	0.891	0.015	0.840	0.922
	2	4	2	0.895	0.014	0.850	0.926
	1	4	3	0.906	0.011	0.871	0.930
	1	5	2	0.908	0.011	0.874	0.932
Presentations set[b]							
5	1	2	2	0.849	0.025	0.768	0.902
6	2	2	2	0.849	0.024	0.761	0.908
	1	3	2	0.879	0.017	0.827	0.920
7	2	3	2	0.879	0.016	0.820	0.921
	1	3	3	0.899	0.013	0.856	0.928
	1	4	2	0.899	0.013	0.859	0.926
8	2	3	3	0.896	0.014	0.858	0.929
	2	4	2	0.898	0.012	0.861	0.930
	1	4	3	0.913	0.010	0.887	0.935
	1	5	2	0.914	0.010	0.887	0.937

Note. Shaded rows show the combination used in the operational TOEFL Essentials test. [a]Easy = 9–11 syllables; medium = 13–15; difficult = 18–29. [b]Easy = 6–11 syllables; medium = 13–16; difficult = 18–22.

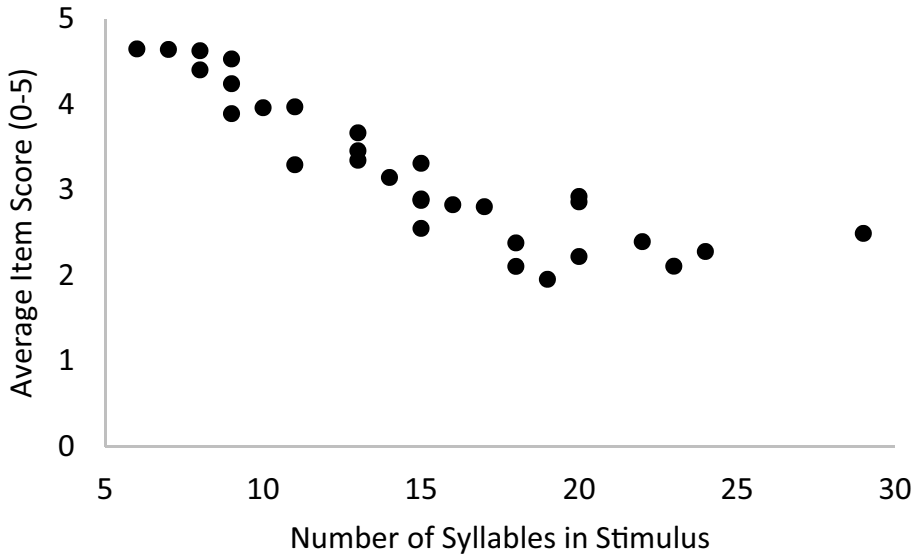


Figure 9 Relationship between length of stimulus and mean score for the pilot data. *Note*. $k = 30$ items; $n = 314–340$ for each item.

apparent that the shortest items were easiest, with most test takers scoring perfect (5) or nearly perfect (4) ratings on items up to nine syllables in length. From 10 syllables through 19, the percentage of scores at points 0, 1, 2, and 3 increases noticeably, while scores at 4 and 5 points systematically decrease. Overall, the average scores for items at each syllable length decrease linearly from 6 syllables to 19 syllables, reflecting the intended design of these items. It is also apparent that the set of items ranging from 18 to 29 syllables reflects, on average, similar levels of difficulty, with very few test takers

Table 11 Average Scores for Stimuli of Differing Lengths

No. of syllables	No. of items	No. of test takers	Final score		Raw score, %					
			M	*SD*	0	1	2	3	4	5
6	1	318	4.6	1.09	3.5	1.6	1.9	1.3	6.9	84.9
7	1	322	4.6	0.92	2.2	0.3	1.9	5.0	14.3	76.4
8	2	664	4.5	0.97	1.8	1.1	2.4	8.9	18.1	67.8
9	3	1,004	4.2	1.11	2.1	2.4	5.0	14.7	24.4	51.4
10	1	337	4.0	1.03	1.2	3.3	6.5	17.8	41.2	30.0
11	2	666	3.6	1.31	1.8	8.1	15.8	19.8	23.7	30.8
13	3	996	3.5	1.32	1.9	11.5	14.1	20.3	28.2	24.0
14	1	324	3.1	1.34	4.9	10.8	21.0	23.1	25.0	15.1
15	4	1,316	2.9	1.32	3.3	18.6	23.3	23.3	19.5	12.0
16	1	321	2.8	1.26	5.9	11.5	26.5	32.7	14.0	9.3
17	1	340	2.8	1.25	5.6	13.5	26.5	27.1	21.2	6.2
18	2	657	2.2	1.13	8.4	23.4	34.1	22.7	10.2	1.2
19	1	319	2.0	1.11	11.3	31.0	33.2	16.9	7.2	0.3
20	3	997	2.7	1.25	4.9	19.3	25.2	31.5	11.6	7.5
22	1	314	2.4	1.17	5.4	22.9	30.9	27.1	9.9	3.8
23	1	318	2.1	1.22	15.7	20.8	25.5	33.3	4.4	0.3
24	1	337	2.3	1.00	5.3	23.1	38.6	24.0	8.6	0.3
29	1	334	2.5	1.00	4.5	13.8	36.8	35.9	6.9	2.1

Note. Where one rater indicated that the response was unscorable, a third score awarded by an adjudicator was used if available. Otherwise, the score from the second rater was used, if the rater awarded a score for the response.

Table 12 Pearson Correlations Between Elicited Imitation Task Total Score and Other Measures

Measure (scale, no. of items)	Pearson correlation with EIT (0–75)	*N*
Speaking		
Virtual interview (0–30, $k = 6$)	.84	555
Read aloud (0–24, $k = 6$)	.78	560
Writing total score (0–10, $k = 2$)	.73	581
Listening total score (0–13, $k = 13$)	.68	583
Reading total score (0–16, $k = 16$)	.57	583

Note. EIT = elicited imitation task.

scoring at 4 or 5 points and a majority scoring at 2 or 3 points. Interestingly, additional length beyond 18 or 19 syllables does not imply additional difficulty, with some scoring variability apparent in this set of items.

RQ9: Are Elicited Imitation Task Scores Associated With Measures of Related Abilities?

We also investigated whether EIT scores were predictably related to measures of other aspects of English language proficiency, with the assumption that EIT scores would be more strongly related to measures of speaking ability and less strongly related to measures of other skills. Accordingly, EIT scores were compared with scores on other language measures in the pilot test, including measures of speaking, writing, listening, and reading. Table 12 shows the Pearson correlations between the EIT scores and tests of each of these abilities. Not surprisingly, the strongest correlations were found with the two speaking tasks, Virtual Interview and Read Aloud, thereby providing considerable convergent validity evidence supporting the interpretation of the EIT as an effective measure of English speaking ability. Interestingly, writing scores also demonstrated a strong relationship with EIT scores, suggesting the possibility of overlapping variance due to a shared constructed-response construct (i.e., items that required substantial language production), though this observation bears additional investigation. Correlations between EIT scores and listening scores were slightly lower though still strong, reflecting the role played by listening in performing well on the EIT but also the fact that, in this reduced-redundancy type of task, the emphasis is not on listening comprehension per se but rather on processing each input sentence for language form and meaning to reproduce it exactly. Finally, reading showed the lowest (though still moderately strong)

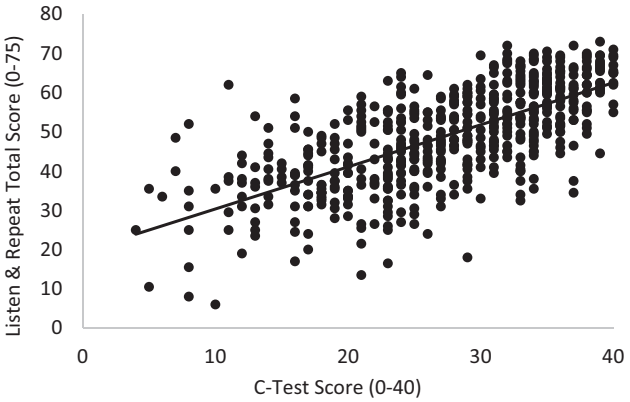


Figure 10 Total elicited imitation task score versus total score on a C-test for the pilot data. *Note*. Data are shown for $N = 581$ test takers.

relationship with EIT scores, reflecting that these tests measure somewhat distinct, if related, aspects of language ability. On the whole, these strong to very strong relationships provide important initial support for the interpretation of the EIT score as both a trustworthy indication of speaking ability and an indication of holistic English proficiency. They also point to the likelihood of EIT scores contributing helpful degrees of discrimination among learners at distinct and broadly differing proficiency levels.

RQ10: Are Elicited Imitation Task Scores Associated With a Global Measure of Proficiency (C-Test)?

To examine the capacity of the EIT to discriminate consistently among learners at broadly differing English proficiency levels, a comparison was made with an independent criterion measure chosen to represent holistic English proficiency. As a reminder, the C-test consisted of two passages with 20 blanks each, for a total of 40 possible points, and it was intended to capture underlying English proficiency differences among test takers. Figure 10 shows the relationship between total scores on the 15-item EIT and scores on the 40-item C-test for each pilot test taker. The linear positive relationship between the two sets of scores is apparent, and the corresponding Pearson correlation ($r = .69$) indicates a relatively strong relationship, with C-test scores (i.e., holistic English proficiency) accounting for nearly 50% of the variance in EIT scores ($r^2 = 47.3\%$). Note, however, that there is a good deal of spread for the EIT scores at any given C-test score, and the disparities are more extensive toward the lower end of the C-test score scale. These patterns are not surprising, given the distinct modalities for each item type (i.e., writing vs. speaking) as well as the fact that both measures were derived from pilot test designs rather than from finalized operational assessments. In general, then, the moderate to strong relationship exhibited between the two measures suggests that the EIT was capable of discriminating across the broadly differing proficiency levels represented by participants in the pilot study as indicated on an independent C-test measure (i.e., not part of the TOEFL Essentials test battery).

Discussion

In this section, we briefly recap the main findings for each RQ, and we interpret these findings in light of the TOEFL Essentials validity argument. We also identify several new questions about the EIT on the basis of interesting patterns observed in the study results. We then consider some limitations of the current study and what we feel should be priorities for future research to support the use of the EIT in the TOEFL Essentials test. Finally, we consider the extent to which initial evidence generally supports the suitability of the EIT for use in the TOEFL Essentials context, in terms of both the empirical backing for score interpretation and the role the EIT is intended to play within the larger TOEFL Essentials assessment design.

Summary of Main Findings

RQ1: Test Takers' Perceptions of Validity

A majority of pilot test takers (60%) indicated that the EIT was a "very good" or "good" assessment of their speaking ability, lending some support to the use of this task type to represent English speaking proficiency in academic and general life domains. However, this perception contrasted to an important degree with a Virtual Interview task that was piloted at the same time, where the approval rate ("very good" or "good") was 92%. This difference is not unexpected for a reduced-redundancy task, and it supports the hybrid approach taken in designing the TOEFL Essentials test, where measures of foundational language abilities, such as the EIT, are combined with measures that more directly incorporate communicative demands. The extent to which scores on the EIT — on their own and in combination with other speaking tasks — are able to predict ability to use English in diverse academic and general life domains is a key question prioritized for future research.

RQ2: Impact of Animation on Performance

We found little difference in average EIT scores for versions with elaborated animated graphics compared with parallel versions with only a simplified graphic. Mean scores across versions were statistically significantly different, but these mean differences were 2.4 and 1.9 scale points, a relatively small portion of the 0–40 scale of the prototype version of the EIT. Patterns in performance were very similar across proficiency levels, suggesting that lower proficiency test takers were not disadvantaged by this feature. Given the apparent lack of negative effects on performance, animated visual input was included in the operational TOEFL Essentials test to provide an additional indication of language-use context for the EIT. The provision of context was deemed to be an important test design feature that emphasized the interrelation between language knowledge and/or skill and the goal of being able to use language successfully in actual English communication contexts. Whether the provision of context in this EIT and other sections of the test triggers this association among test takers and others (e.g., language teachers) is an important question for subsequent investigations related to the consequences of assessment design and use.

RQ3: Impact of Production Features on Performance

The production features investigated in the TOEFL Essentials pilot had no discernible effect on mean scores for the EIT overall and across learners at distinct proficiency levels. Although preliminary, this finding suggests that within the range of feature variations explored in the pilot, EIT items that vary in topic, speaker accent (United Kingdom vs. United States), and type of visual (map vs. graphic organizer) can likely be considered parallel for measurement purposes. While varying the features does not seem to introduce any effects on test taker performance and scores, the inclusion of topic, accent, and graphic variety is deemed an important design quality of the TOEFL Essentials test to increase test taker interest and reflect a range of language varieties and contexts of language use.

RQ4: Appropriateness of Allowed Response Time

Test takers seemed to have sufficient time to respond to items in the EIT, as indicated by the observation that a majority of participants went on to the next screen before the response time had expired, as well as anecdotal observations that test takers were unlikely to be cut off while speaking. There was also no consistent relationship between the length of stimulus sentence and the proportion of test takers choosing to go on, and low-proficiency test takers were as likely to use the "next" button as test takers as a whole. Moreover, there was little difference in average scores in instances when the test taker chose to go on versus using the full response time, although scores for several items were slightly higher when the "next" button was used. These results suggest that the response times allowed in the pilot study were adequate for repeating the stimulus and that provision of a "next" button did not appear to negatively impact task achievement. These response times were therefore adopted for the operational test. For technical reasons, the operational test does not currently have a "next" button to allow the test taker to advance, but such a capability is under investigation. Provision of a "next" button, and similar affordances, may give the test taker an important sense of control over the test taker's own test-taking performance, and this dimension of the EIT design deserves future scrutiny.

RQ5: Rater Scoring Consistency

The TOEFL iBT raters who participated in the pilot study were able to apply the scoring rubric with reasonable levels of consistency, with exact agreement being achieved for 67.5% of responses overall and adjacent agreement seen for an additional 30.8% of responses (over 98% exact or adjacent agreement). Interrater correlations were also high, with Spearman rho values above .80 for responses 9–23 syllables in length. The scoring rubric used in the pilot went through additional minor revisions to wording prior to being finalized for operational use, but the pilot established that trained raters could successfully apply the criteria developed for scoring the EIT. The observation that EIT responses from a broad proficiency range of learners could be scored consistently by human raters suggested the possibility that reliable automated scoring might also be achievable, indicating an important trajectory for future research.

RQ6: Usefulness of Scores for Distinguishing Proficiency Levels

The EIT versions investigated in the pilot test spread out participants at different proficiency levels very effectively, as reflected by a distribution of scores extending from 6 to 73 on a 75-point scale. The overall mean total score was 49.17, with room for more than 2 standard deviations of scores on either side of the mean, suggesting that the EIT can provide information across a broad range of ability. The reliability of the EIT used in the pilot was also very high (Cronbach's $\alpha = .93$, $k = 15$). These findings offer strong fundamental evidence in support of using the EIT to distinguish consistently among test takers across a broad range of English ability.

RQ7: Optimal Number and Composition of Items

A simulation study suggested that various combinations of five, six, seven, or eight items, drawn from categories of "easy" (6–11 syllables), "medium" (13–16 syllables), and "difficult" (18–29 syllables) items, were highly reliable, with average reliabilities consistently greater than .80. For the TOEFL Essentials test, with an eye toward maximizing efficiency of test administration time, a combination of two easy, three medium, and two difficult items was selected on the basis of both high reliability and coverage across a broad range of language proficiency. The extent to which any of the individual items (of different lengths) contributes to the overall reliable variance elicited among test takers (at different proficiency levels) presents an interesting question for future investigation.

RQ8: Impact of Stimulus Length on Item Difficulty

The difficulty of individual items increased with stimulus length, as expected. The relationship between stimulus length and item difficulty (as measured by average score) was especially strong for items of 6–19 syllables in length, where a Pearson correlation of .96 was observed. This finding provides critical evidence in support of the key design principle underlying EIT item development, which enables the systematic creation of items that are capable of challenging learners at different proficiency levels to predictable degrees. Interestingly, item difficulty appeared to plateau beyond an item length of 19 syllables, raising the question of optimal minimum and maximum lengths of item stimuli for future investigation. In addition, other characteristics of the input sentences, such as syntactic structure, were not tightly controlled. Slight discrepancies were observed in the difficulty of items with the same or similar numbers of syllables, indicating the need to investigate other factors in stimulus design (e.g., lexical and grammatical complexity) as well as the possibility that construct-irrelevant factors (e.g., working memory of test takers) may influence performance.

RQ9: Relationship to Related Measures of Ability

In the pilot investigation, EIT scores were most highly correlated with other measures of speaking ability, as expected (Pearson $r = .84$ for a Virtual Interview task, .78 for a Read Aloud task), providing a key source of criterion-related validity evidence in support of interpreting the EIT as a holistic measure of speaking proficiency. Correlation with the writing tasks was also high (.73); this result bears further investigation but may reflect a shared requirement to construct language. Correlation with the listening section (.68) was also high, as might be expected given the importance of listening comprehension for the EIT, while correlation with the reading section (.57) was somewhat lower but still moderately

strong. Overall, the results are consistent with the expected relationship between language abilities measured by TOEFL Essentials, where a general level of language proficiency is anticipated to have an effect on all measures, while measures of specific language skills should be most closely related with each other.

RQ10: Association With a Global Proficiency Measure

There was a moderately strong relationship between scores on the pilot EIT and scores on a C-test, which served as an independent criterion measure of general language proficiency (Pearson $r = .69$, $r^2 = 47.3\%$). This observed relationship is in accordance with expectations given that the EIT and C-test similarly measure general language proficiency but assess abilities in different modalities (i.e., oral vs. written language). Future investigations of the relationship of EIT scores with other criterion measures will provide additional evidence for interpreting the extent to which the EIT is able to assess general English proficiency and the more specific speaking skill.

Limitations of the Current Study

This report describes the design and piloting of a new EIT, reflecting the steps taken to finalize design features and provide basic evidence to support the validity of the EIT for use in the TOEFL Essentials test. As an initial effort, the study naturally has a number of limitations; several specific limitations were considered in the previous section, and others are addressed in the following discussion of future validity research. More generally, one limitation was the sample of test takers on which the findings are based. Recruitment of participants for both the prototyping and pilot phases of the study focused on obtaining a sample of individuals across a wide range of English language proficiency, consistent with the requirement that the TOEFL Essentials test be capable of measuring across a broad spectrum of ability. During the pilot study, considerable effort was also made to recruit individuals representative of the language backgrounds expected of the TOEFL Essentials candidature (i.e., from 10 different regions of the world). However, at the time of writing, the extent to which this sample of individuals will be representative of actual test takers is unknown, and so validity claims based on the current evidence are provisional.

Another limitation of the current study is the extent to which potentially interacting phenomena could be controlled. This report documents development work on the EIT, and as a development project, it was not always possible to conduct the types of tightly controlled experiments that would make it possible to isolate the effects of specific variables. One obvious example of this issue is the evaluation of item production features on test taker performance (RQ3), where the variables of task topic, speaker accent, and visual content were confounded. This situation reflects the fact that the goal of the pilot study was to confirm that the EIT worked as designed, rather than to investigate the relative impacts of different features on performance. Nonetheless, the current data do not allow for detailed investigation of variables that may impact performance, and better-controlled studies are needed in future research on the EIT. Another important limitation is that the current study focused on selected aspects of a validity warrant for the interpretation and use of the EIT, namely, establishing the extent to which the EIT could elicit performances in predictable ways, raters could score them consistently, and the resulting scores could distinguish among examinees systematically. Additional research will be called for to more thoroughly inquire into these aspects of the validity argument and, in particular, to further explore the uses and consequences of the EIT.

A final limitation has to do with the possibility that distinct EIT task sets (e.g., Campus Tour vs. Presentation) represent distinct levels of difficulty to test takers. While efforts were made during item development to construct EIT task sets of comparable difficulty, it is an empirical question whether each task set elicits equivalent performances. Of course, score equating procedures can be used to adjust for slight discrepancies in difficulty such that resulting total scores on different EIT versions are comparable, but it will be important in future research to investigate the relationship between performances on EIT task sets that are intended to present similar difficulty to test takers.

Future Research to Support the TOEFL Essentials Validity Argument

The results described in this report provide initial backing for multiple inferences in the TOEFL Essentials validity argument as outlined by Papageorgiou et al. (2021). However, in keeping with the types of investigations conducted at the pilot stage, the bulk of the evidence supports the evaluation inference, that is, "observations of performance on the TOEFL

Essentials test tasks are evaluated to produce scores reflective of targeted language abilities" (Papageorgiou et al., 2021, p. 40). Most immediately, these initial findings will need to be confirmed with similar analyses of data from the operational test, and we believe that analyses of rater performance and impact of task design features should be afforded high priority for follow-up investigations. Additionally, the potential for subgroup differences has yet to be evaluated and is a high priority for research to support the evaluation inference.

Evidence to support the higher levels of the validity argument, including the generalization, explanation, extrapolation, and utilization inferences, will typically entail analyses of the TOEFL Essentials test as a whole, but work specific to the EIT remains to be done here as well. For the generalization inference (observed scores are consistent over parallel forms), ongoing monitoring of EIT difficulty across test forms will provide key backing for validity claims, but additional research is needed to understand the relationship of item difficulty to linguistic features other than sentence length, primarily to confirm item-writing specifications. For the explanation inference (expected scores can be attributed to the relevant construct), it will be necessary to confirm the relationship of EIT scores to other measures of language ability within the operational TOEFL Essentials test, given that the intercorrelations reported here (RQ9) are only preliminary evidence from a pilot version of the test, which did not contain the full suite of TOEFL Essentials items. It will similarly be of considerable interest to compare results of the EIT to real-world measures of communicative language ability, to support the extrapolation inference (scores reflect language performance in intended real-life contexts). Overall, interpretation of the construct measured by the EIT is complex and will be informed in part by relationships to other measures of ability.

Finally, the findings of this report do not address the utilization inference, which is unavoidable given that the research was conducted prior to the launch of the operational test. However, one critical utilization-related concern specific to the EIT is the potential for negative washback; as a reduced-redundancy measure of holistic L2 proficiency, the EIT is generally not intended to form a basis for language-learning instructional activities. Concerns regarding washback played a role in the overall design of the TOEFL Essentials test, where relatively artificial measures like the EIT are combined with other tasks that more directly reflect the types of language use found in academic and daily life contexts. Note, though, that the creation of meaningful sets of EIT items that represent an actual language-use scenario, coupled with the provision of communication context, provides at least some degree of enhanced reality and correspondence with real-life language use. However, to avoid any possible unintended impacts on teaching and learning, there will be a need for periodic evaluation of test preparation materials and related instructional practices to identify instances of misuse, combined with ongoing communication regarding appropriate test preparation practices.

Conclusion

On the basis of initial evidence from prototyping and pilot investigations, the innovative EIT design explored here is capable of fulfilling its primary purpose of efficiently separating test takers according to their general English speaking proficiency. The EIT was found to elicit speaking performances consistently, with little noticeable influence from item delivery features. Performances were ratable with a high degree of reliability by trained raters, and resulting scores systematically spread test takers across a range of abilities. The basic item design parameter of increasing difficulty by increasing stimulus length was confirmed through the very strong negative correlation between stimulus length and performance rating. The EIT scores were also strongly correlated with criterion-related measures, more so with other measures of speaking but also with an independent measure of holistic English proficiency. The evidence accumulated thus far, then, provides substantial initial backing for the use of the EIT as one key component of the overall speaking ability construct measured in the TOEFL Essentials test.

Importantly, the EIT is not the only measure of speaking ability in the TOEFL Essentials test. Several other measures (dialogic Read Aloud, Virtual Interview) were also designed to present test takers with a variety of speaking tasks that emphasize distinct dimensions of their speaking ability. The EIT plays a critical role in providing a highly reliable and quick estimate of the overall speaking proficiency level of the test taker, and it seems to do so effectively at the full range of English proficiency covered by the TOEFL Essentials test, including beginning English learners as well as very advanced speakers of the language. Coupled with other measures that probe distinct aspects of speaking ability (e.g., intelligibility and fluency, communicative effectiveness), the EIT supports an overall assessment of the test taker's English speaking proficiency.

References

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, *16*(1), 61–70. https://doi.org/10.2307/3586563

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second language competence. In E. Tarone, S. M. Gass, & A. D. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 245–261). Lawrence Erlbaum.

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what. *International Journal of Language Studies*, *7*(1), 1–32. http://www.ijls.net/sample/71-1.pdf

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653–675. https://doi.org/10.2307/3587999

Caulfield, J., & Smith, W. C. (1981). The reduced redundancy test and the cloze procedure as measures of global language proficiency. *The Modern Language Journal*, *65*(1), 54–58. https://doi.org/10.2307/326207

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

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism Toronto*, *19*, 197–202. ERIC. https://files.eric.ed.gov/fulltext/ED184334.pdf

Deygers, B. (2020). Elicited imitation: A test for all learners? Examining the EI performance of learners with diverging educational backgrounds. *Studies in Second Language Acquisition*, *42*, 933–957. https://doi.org/10.1017/S027226312000008X

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*(3), 464–491. https://doi.org/10.1093/applin/aml001

ETS. (2021). *TOEFL Essentials Test speaking scoring guide*. https://www.ets.org/s/toefl-essentials/rsc/pdf/speaking-rubric.pdf

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, *66*(2), 419–447. https://doi.org/10.1111/lang.12157

Gradman, H. L., & Spolsky, B. (1975). Reduced redundancy testing: A progress report. In R. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 59–70). Center for Applied Linguistics. ERIC. https://files.eric.ed.gov/fulltext/ED107161.pdf

Grotjahn, R., Klein-Braley, C., & Raatz, U. (1992). C-Tests in der praktischen Anwendung: Erfahrungen beim Bundeswettbewerb Fremdsprachen. In *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Vol. 1, pp. 263–296). Brockmeyer.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research* (Vol. 41). John Benjamins. https://doi.org/10.1075/lllt.41

Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*(1), 115–122. https://doi.org/10.1111/jedm.12007

Kenyon, D. M., & MacGregor, D. (2012). Pre-operational testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 295–306). Routledge.

Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an elicited imitation test for SLA research. *The Modern Language Journal*, *100*(3), 655–673. https://doi.org/10.1111/modl.12346

Klein-Braley, C. (1985). A cloze-up on the C-test: A study in the construct validation of authentic tests. *Language Testing*, *2*, 76–104. https://doi.org/10.1177/026553228500200108

Kostromitina, M., & Plonsky, L. (2021). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263121000395

Larsen-Freeman, D. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly*, *9*, 409–419. https://doi.org/10.2307/3585625

Lever, J. F., & Lonsdale, D. W. (2015). Elicited imitation for Brazilian Portuguese. *Brazilian English Language Teaching Journal*, *6*(2), 142–161. https://doi.org/10.15448/2178-3640.2015.2.21251

Long, M. H., & Richards, J. C. (1990). *The development of second language proficiency*. Cambridge University Press.

Markman, B. R., Spilka, I. V., & Tucker, G. R. (1975). The use of elicited imitation in search of an interim French grammar. *Language Learning*, *25*(1), 31–41. https://doi.org/10.1111/j.1467-1770.1975.tb00107.x

Mozgalina, A. (2015). *Applying an argument-based approach for validating language proficiency assessments in second language acquisition research: The elicited imitation test for Russian* [Unpublished doctoral dissertation]. Georgetown University.

Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism, 2*, 1–37.

Nissan, S., & Schedl, M. (2012). Prototyping new item types. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 281–294). Routledge.

Norris, J. M. (2008). *Validity evaluation in language assessment*. Peter Lang. https://doi.org/10.3726/978-3-653-01171-5

Norris, J. M. (Ed.). (2018). *Developing C-tests for estimating proficiency in foreign language research*. Peter Lang. https://doi.org/10.3726/b13235

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 716–761). Blackwell. https://doi.org/10.1002/9780470756492.ch21

Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573–589). Routledge.

Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, *37*(5), 693–715. https://doi.org/10.1093/applin/amu060

Ockey, G. J., Papageorgiou, S., & French, R. (2016). Effects of strength of accent on an L2 interactive lecture listening comprehension test. *International Journal of Listening*, *30*(1–2), 84–98. https://doi.org/10.1080/10904018.2015.1056877

Oller, J. W., Jr. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning*, *23*(1), 105–118. https://doi.org/10.1111/j.1467-1770.1973.tb00100.x

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners* [Unpublished doctoral dissertation]. University of Hawai'i at Manoa.

Ortega, L., Iwashita, N., Rabie, S., & Norris, J. M. (1999). *A multilanguage comparison of measures of syntactic complexity*. University of Hawai'i, National Foreign Language Resource Center.

Papadopoulou, D., & Clahsen, H. (2003). Parsing strategies in L1 and L2 sentence processing: A study of relative clause attachment in Greek. *Studies in Second Language Acquisition*, *25*(4), 501–528. https://doi.org/10.1017/S0272263103000214

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the TOEFL Essentials test 2021* (Research Memorandum No. RM-21-03). ETS. https://www.ets.org/Media/Research/pdf/RM-21-03.pdf

Park, H. I., Solon, M., Henderson, C., & Dehghan-Chaleshtori, M. (2020). The roles of working memory and oral language abilities in elicited imitation performance. *The Modern Language Journal*, *104*(1), 133–151. https://doi.org/10.1111/modl.12618

Sarandi, H. (2015). Reexamining elicited imitation as a measure of implicit grammatical knowledge and beyond … ? *Language Testing*, *32*(4), 485–501. https://doi.org/10.1177/0265532214564504

Spada, N., Shiu, J. L. J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*(3), 723–751. https://doi.org/10.1111/lang.12129

Trace, J. (2020). Clozing the gap: How far do cloze items measure? *Language Testing*, *37*(2), 235–253. https://doi.org/10.1177/0265532219888617

Tracy-Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (2014). "Repeat as much as you can": Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143–166). Multilingual Matters. https://doi.org/10.21832/9781783092291-011

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73. https://doi.org/10.1111/1473-4192.00024

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*(4), 680–704. https://doi.org/10.1111/flan.12063

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

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497–528. https://doi.org/10.1177/0265532215594643

Zhou, Y. (2012). *Willingness to communicate in learning Mandarin as a foreign and heritage language* [Unpublished doctoral dissertation]. University of Hawai'i at Manoa.

Appendix A

Stimuli Sentences Used in the Pilot Study

Stimuli sentences were used for research purposes only. The number of syllables is in parentheses.

Campus Tour

Welcome to our university. (9)
It's very nice to meet you all. (8)
Living on campus is really fun. (9)
The café is a great place to meet friends. (10)

In the game room we watch movies and play pool. (11)
Physical education courses are required. (13)
A pharmacy is located in the health center. (13)
This is where researchers are creating new technology. (15)
Some students gain work experience here as lab assistants. (15)
With a student ID card you can open a free savings account. (17)
There are more than 2,000 works of art on display across the campus. (18)
Many of our courses, taught by excellent professors, can be taken online. (20)
Paying tuition and other school fees can also be taken care of right here. (20)
One great aspect about this place is that, when classes are in session, we have concerts on Fridays. (24)
I hope to see you next year, when you get admitted and decide to enroll in our fantastic university. (29)

Presentations

Welcome, everybody. (6)
I'm happy to see you all. (7)
Thank you for coming to my talk. (8)
Let me begin at the beginning. (9)
If you have any questions, feel free to ask. (11)
My topic for today is important to us all. (13)
To be clear, that is the biggest challenge we have to solve. (14)
Let's pause here to see if everybody understands so far. (15)
I'd like to expand further on that particular issue. (15)
Here's a quick summary of the main points before we continue. (16)
In the remaining time, I will outline a few key things we can all do. (18)
This brings up a critical and confusing problem which we will turn to next. (19)
In conclusion, I have made a variety of recommendations today. (20)
Hopefully I have been successful in convincing you of the importance of this topic. (23)
So now I would be interested to hear whatever comments, questions, or thoughts you might have. (22)

Appendix B

Background Information Questions

Please answer a few questions about yourself.
Please answer all of the questions.
When you have answered all of the questions you can click the button to go on.

1. What is your first language?
2. What is your gender?

   Female
   Male
   Prefer not to respond
   Other

3. What is your age?

   <18 years
   18 – 22 years
   23 – 30 years
   31 – 40 years
   >40 years

4. What is your level of education?

   I finished secondary school
   I am an undergraduate student (e.g., BA BS)
   I finished an undergraduate degree (e.g., BA, BS)
   I am a graduate student (e.g., MA, MS, PhD)
   I finished a graduate degree (e.g., MA, MS, PhD)
   None of the above

5. How many years have you learned English?

   2 years or less
   3 – 5 years
   6 – 8 years
   9 – 11 years
   12 years or more

6. What is your English speaking ability like? (*Used in prototyping study only*)

   I can easily say even complex ideas in English
   I can say complex ideas in English, but I have to work hard to do it
   I can say simple ideas fairly easily, but I can't say complex ideas
   I have to work hard to say even simple ideas in English

Suggested citation:

Davis, L., & Norris, J. (2021). *Developing an innovative elicited imitation task for efficient English proficiency assessment* (TOEFL Research Report No. 96). ETS. https://doi.org/10.1002/ets2.12338

Action Editor: Brent Bridgeman

Reviewers: Yeonsuk Cho and Cynthia Newman