ets research institute

APRIL
2025

TOEFL® RESEARCH SERIES

Testing Academic Language Proficiency

Comparing the TOEFL iBT® and the Duolingo Test of English

Sara T. Cushing

TOEFL–RR-104

ETS Research Report
No. RR–25-01

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Testing Academic Language Proficiency: Comparing the TOEFL iBT® Test and the Duolingo English Test

Sara T. Cushing

Department of Applied Linguistics & ESL, Georgia State University

Abstract

This report provides an in-depth comparison of TOEFL iBT® and the Duolingo English Test (DET) in terms of the degree to which both tests assess academic language proficiency in listening, reading, writing, and speaking. The analysis is based on publicly available documentation on both tests, including sample test questions available on the test websites. For each skill area, the construct as defined by the test developer is discussed, followed by a review of the test content and the cognitive processes involved in completing items related to that skill. The results are evaluated against three propositions from the TOEFL® validity argument (Chapelle et al., 2008). The analysis suggests that although both tests provide evidence of general language proficiency, the input to test takers and the test items for TOEFL iBT are more academic in nature, both in terms of the content and the cognitive demands of the task. Some of the aspects of TOEFL iBT that make the test relatively more academic include extensive reading and listening passages and integrating information from multiple texts in speaking and writing. The report also highlights some of the ways in which a commitment to automated item generation and scoring limits the ability of a test like DET to fully represent the construct of academic language proficiency.

Keywords: academic language proficiency, admissions tests, TOEFL iBT®, Duolingo English Test, construct validity

Corresponding author: Sara T. Cushing, Email: stcushing@gsu.edu

Since the late 1950s, U.S. universities wanting to admit international students have recognized the need to assess the English language proficiency of their prospective students. The TOEFL® program emerged out of these concerns in the 1960s (see Taylor & Angelis, 2008, for a historical overview) and was, for decades, the undisputed leader in large-scale proficiency tests

in the United States. The TOEFL iBT®, introduced in 2005, is still the most well-known and widely accepted test in the United States and has a very well-documented, robust record of validation research (see ETS, 2024, for a history of the TOEFL program). TOEFL iBT is widely accepted in all other major English-speaking academic destinations, including Canada, the United Kingdom, and Australia. In recent decades, however, other tests have entered the market, particularly as natural language processing (NLP) and machine learning (ML) techniques have rapidly advanced, automating processes such as item generation and scoring and opening up opportunities for entrepreneurs to enter the market with tests that promise results comparable to those of TOEFL iBT with a shorter administration time, faster score reporting, and lower prices.

One such test is the Duolingo English Test (DET), which was introduced in 2016. The DET technical manual (Cardwell et al., 2024b, p. 2) states that the DET's mission is "to lower barriers to education access for English language learners around the world . . . by leveraging technological advances in annual test updates to produce an accessible and affordable high-stakes language proficiency test that produces valid, fair, and reliable test scores." Despite early critiques from language testing scholars (Wagner & Kunnan, 2015; Wagner, 2020), DET gained widespread provisional acceptance during the COVID-19 pandemic in 2020 when test centers worldwide had to close and other available options for assessing English language proficiency were limited.

In the intervening years, DET has introduced several new item types and retired others, addressing some of the critiques leveled against earlier versions of the test. Specifically, Wagner & Kunnan (2015) criticized DET for a number of shortcomings, including the following: (a) a gap between the DET test tasks and the characteristics of the language use domain (academic English); (b) a very limited construct being tested due to the requirement that all test tasks be automatically scored; (c) the inability of the test to assess the ability of examinees to use the language interactively (i.e., with other human beings); and (d) the lack of productive tasks (speaking and writing). Some of the newer tasks, described in this report, are intended to address these issues within the DET framework of automated task generation and scoring. At the same time, TOEFL iBT has also revised some test tasks and reduced the length of the test twice, in 2019 and 2023 (ETS, 2024). As a result of these revisions, the overall test-taking time was reduced from 4 hours at launch in 2005 to 2 hours since 2023.

In light of these changes to both tests, it may be helpful to provide a comparison of the content of both tests so that institutions considering one or both tests for admissions may have a basis for comparison. The International Language Testing Association (ILTA; n.d.) includes these responsibilities, among others, for test users in the *ILTA Guidelines for Practice*:

- Use results from a test that is sufficiently reliable and valid to allow fair decisions to be made.
- Make certain that the test construct is relevant to the decision to be made.
- Clearly understand the limitations of the test results on which they will base their decision.

This report is an attempt to address this need. In the report, I provide an in-depth comparison of the TOEFL iBT and the DET, focusing primarily on test content and important aspects of test validity. My analysis is based on publicly available documentation on both tests, including sample test questions available on the test websites. DET allows unlimited practice tests, which I availed myself of a few times. Parts of the DET are computer-adaptive, so I was able to access items at different proficiency levels, depending on whether I provided correct or incorrect responses to items early on in the test. ETS provides a complete sample TOEFL iBT test by download.

The framework for the comparison is based on Kane's (2013) argument-based approach to validation, a widely accepted validation framework in language testing, which drives research supporting the interpretation and use of TOEFL iBT, described in depth in Chapelle (2008) and summarized for more general audiences in ETS (2020). Kane's argument-based approach has also been recently adopted by DET (Kostromitina, 2024). For this study, I focused on three propositions from ETS (2020, p. 5), which relate specifically to test content as it relates to academic language ability. These propositions are as follows:

- The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings.
- Tasks and scoring criteria are appropriate for obtaining evidence of test takers' academic language abilities.
- Academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks.

Similar propositions are found in Kostromitina (2024) for DET, specifically:

- Test developers have analyzed the target domain of language use (e.g., university study) to create the test tasks that elicit relevant performance.
- The test is designed to measure language skills through tasks that appropriately reflect target language use situations (e.g., university study).
- For tests used in university admissions specifically, the scores reflect the level of language performance that test takers are likely to display in their academic studies.

Common to both of these formulations is the idea that test tasks should be relevant to the target domain and elicit performances and scores that are indicative of language abilities in that domain. One notable difference is that DET uses academic study as an example rather than as the main purpose of the test, unlike TOEFL iBT, which is specifically designed as a test of academic language ability.

The report is organized as follows. First, I provide a brief overview of the two tests. Then after a brief consideration of integrated skills, I discuss the test content in terms of the four major language skills of reading, listening, writing, and speaking. In the final section, I revisit these propositions and discuss the extent to which publicly available information supports them. In this section, I also discuss relevant issues such as test security and test ethics.

Overview of the Two Tests

Overview of TOEFL iBT

A concise statement of the TOEFL construct is found in ETS (2020, p. 4): "TOEFL iBT test scores are interpreted as the ability of the test taker to use and understand English as it is spoken, written, read, and heard in college and university settings." In other words, TOEFL iBT is specifically designed to assess English language ability for postsecondary academic purposes.

TOEFL iBT contains four sections representing the four traditional language skills (reading, listening, writing, and speaking); however, unlike the previous paper-based TOEFL test, which primarily assessed skills separately, TOEFL iBT contains integrated speaking and writing tasks, which represent more accurately how language is used in academic settings. The test is administered on a computer in a testing center and takes approximately 2 hours. Scores are reported as total scores ranging from 0 to 120 and subscores in each of the four skill areas ranging from 0 to 30.

Table 1 (adapted from ETS, n.d.[d]) provides an overview of the test structure.

Table 1. TOEFL iBT Structure

Test section	Estimated timing	Questions/tasks	Description
Reading	35 minutes	20 questions • 2 passages • 10 questions per passage	Read passages and respond to questions

Listening	36 minutes	28 questions
		• 3 lectures
		  ○ 6 questions per lecture
		• 2 conversations
		  ○ 5 questions per conversation
			Answer questions about brief lectures or classroom discussions

Speaking	16 minutes	4 tasks • 1 independent • 3 integrated	Talk about a familiar topic (1) Read and/or listen to brief texts and then discuss (3)
Writing	29 minutes	2 tasks • 1 integrated • 1 writing for an academic discussion	Synthesize information in writing from reading and listening State and support an opinion in an online classroom discussion.

The specifics of each part of the test are discussed in more detail in the relevant sections below. Test items are created by experienced teams of item developers and undergo a rigorous review process (see, for example, Enright et al., 2008; ETS, 2024; and Huff et al., 2008, for the process of prototyping tasks and the whole test when TOEFL iBT was in development). Speaking and writing responses are scored by trained human raters in conjunction with proprietary automated scoring tools.

Overview of DET

The construct measured by DET, as expressed in various iterations of its technical manual, has changed over the years. In versions of the technical manual downloaded between 2020 and 2022, the following statement appeared: "[The DET] assesses test-taker ability to use the language skills required for literacy, conversation, comprehension, and production" (Cardwell et al., 2022, p. 3). In the manual dated May 2024 (Cardwell et al., 2024a, p. 4), this statement was modified somewhat, as follows: "The DET measures test-taker ability to use the language skills required for literacy, conversation, comprehension, and production, including the skills necessary for success in academic contexts." Five months later, in October 2024 (Cardwell et al., 2024b, p. 4), the statement underwent a major revision to read: "The DET measures test-taker ability to use the independent language skills of speaking, writing, reading, and listening

(SWRL skills). These subskills can also be combined into the integrated language skills required for literacy (reading and writing), conversation (speaking and listening), comprehension (reading and listening), and production (speaking and writing), including the skills necessary for success in academic contexts."

As reflected in the statements quoted above, DET originally reported language skills in combination rather than in isolation in recognition of the fact that language use nearly always involves more than one modality at a time. Before 2024, DET reported scores that included a total score as well as subscores for the combined skills described above. At some point in 2024, the DET developers added scores for single skills as well. All scores are reported on a scale from 10 to 160.

As noted above, DET has evolved somewhat to include fewer item types that appear to assess lower-order language knowledge and skills and more item types that measure higher-order skills (see, for example, Wagner & Kunnan, 2015; Wagner, 2020, for descriptions of earlier versions of the test). As of 2024, the test includes both computer-adaptive item types and nonadaptive item types. There are five computer-adaptive item types, described briefly below, that focus on what the manual refers to as "linguistic resources."

- Yes/No vocabulary (15–18 items): Test takers are presented with either an English word or a pseudo-word and have to decide whether or not it is a word.
- Vocabulary in context (6–9 items): Test takers are presented with a sentence that includes a "damaged" word (only the first one to four letters of the word are presented) and must complete the word correctly. Example: "Maria closed her eyes tightly and wis___ for her interview to be successful. "
- C-test (4–6 tasks with 10–14 items in each): Test takers see a 3–5 sentence paragraph. The first and last sentences are intact but only the first half of every other word in the middle sentences is presented and the test taker has to fill in the blanks correctly.
- Dictation (6–9 items): Test takers hear a sentence and type what they hear.
- Read-aloud (4–6 items): Test takers record themselves reading a sentence aloud.

The rest of the tasks are categorized as "skills mastery" and consist of interactive reading, interactive listening, and several open-ended speaking and writing tasks to be discussed in the relevant sections below. The interactive reading and interactive listening tasks are computer adaptive in the sense that the algorithm selects tasks based on the estimated ability levels from

previous sections of the test. The other item types are not computer adaptive. These item types, which are discussed in more detail under the relevant sections below, are the following: interactive reading (2 sets), interactive speaking (2 sets), picture description (writing, 3 sets), interactive writing (1 set), extended speaking (text prompt, 1 set), extended speaking (audio prompt, 2 sets), writing sample (1 set), and speaking sample (1 set). The writing and speaking samples are shared with institutions when scores are reported.

A unique aspect of DET is that all items are all automatically generated and include human review at several stages in the process (see Cardwell et al., 2024b, p. 18, for a description). DET responses are also scored automatically by proprietary scoring models (see Nydick & Lockwood, 2024, for an overview of DET scoring).

Integrated Skills

The discussion below looks at the four traditional language skills separately, but because both TOEFL iBT and DET include tasks that integrate two or more of the skills, it may be useful to discuss how each test conceptualizes and operationalizes integrated skills. For several decades, scholars have noted that "academic writing is rarely done in isolation, but is virtually always done in response to source texts" (Weigle, 2004, p. 30) and that students are expected "to read, discuss, and think critically about" (Weigle, 2004, p. 30) a topic before they write about it. The same can be said for academic speaking in that class discussions are typically based on the expectation that students have read assigned readings and/or listened to a lecture. This integration of skills is pervasive in almost all real-world language use, especially situations that involve speaking or writing, but is particularly salient in academic speaking and writing.

The developers of both TOEFL iBT and DET recognize that responding to language test items frequently requires an integration of skills, even if it is as simple as reading a one-sentence writing prompt or answering written listening comprehension questions. Test designers, therefore, need to consider the extent to which responses to items intended to assess individual skills are impacted by other skills. For example, on a multiple-choice test of listening, the reading level of the test items should not be at a higher level of difficulty than the listening input. By the same token, in tasks intended to simulate academic speaking or writing by integrating aural or written source texts, care should be taken to ensure that difficulty to comprehend the source texts does not impact the test taker's ability to respond to the task, if inferences about speaking or writing ability are to be made from the performance. It is therefore incumbent upon

test developers to be explicit about the degree to which performance on any task type that contributes to a score for an individual language skill actually depends on that skill. This issue is considered throughout the following sections.

Listening

Defining the Listening Construct

Listening is one of the most important academic skills, yet it is notoriously difficult to assess. In every major study of how college students spend their time, listening emerges as the communication skill most used, comprising up to 55% of all time spent communicating (see Janusik & Wolvin, 2009, for a review). Students need listening skills for comprehending lectures, videos, and other nonparticipatory communication events; indeed, this one-way listening (Lynch, 2011) has traditionally been the focus of assessment in language proficiency tests.

In addition to lectures, students need listening skills as an essential part of interactional competence (i.e., two-way listening) where the listener is also a contributor to the communication. Students need listening skills for participating in many activities such as small group discussions, tutorials, and so on and for various academic navigation skills such as talking with an advisor or getting help with technology. When international students have problems with these interactions, some research suggests that comprehension problems may be at the root of these difficulties (Papageorgiou et al., 2021).

Thus, it can be argued that listening should have a prominent place in a proficiency test for academic admissions. However, listening is challenging to assess for numerous reasons. First, the listening process itself is invisible, and the degree to which a test taker has comprehended something requires a response that necessitates the use of other skills, such as reading and answering comprehension questions or providing spoken or written answers to questions. Second, creating a listening test involves numerous complex decisions, such as the choice of text; the characteristics of the speaker, including factors such as accent and gender; and various aspects of the presentation of a listening text, such as whether or not the input can be repeated or whether to include supporting visual information. Finally, the assessment of two-way listening in the context of real-time interactions is notoriously challenging (Lam, 2021; Wagner, 2022). This is particularly an issue with computer-delivered language tests that lack a human interlocutor.

In this section of the report, I first discuss the listening construct as defined by each test provider. Then I describe the content of the listening sections of each test, review the

characteristics of the input and responses, and discuss the cognitive processes involved in responding. Finally, I provide a comparison in light of propositions in the TOEFL iBT validity argument (as discussed above; Chapelle, 2008; ETS, 2020).

Listening Construct for TOEFL iBT

Papageorgiou et al. (2021) provided a construct definition for assessing listening comprehension in EAP settings as follows:

> The assessment of listening comprehension for general academic purposes measures test takers' abilities and capacities to comprehend realistic spoken language in the following subdomains of the English-speaking academic domain: social-interpersonal, academic-navigational, and academic-content. . . . To demonstrate these abilities and capacities, test takers are required to use linguistic resources effectively to comprehend aural input sufficiently in order to select, relate, compare, evaluate and synthesize key information from listening stimuli. (p. 87)

They further divided the construct into communication goals, which are to understand main ideas, supporting details, relationships among ideas, inferences, opinions, speaker purpose, and speaker attitude. The foundational and higher level abilities that are needed include the following:

- Processing extended spoken information in real time
- In order to comprehend meaning:
  - Making use of phonological information, including intonation, stress, and pauses
  - Making use of lexical and grammatical information
  - Making use of pragmatic information encoded in talk
- In order to understand connections between statements and between ideas
  - Processing organization devices (cohesive and discourse markers, exemplifications, etc.)

The TOEFL iBT listening items aim to assess the following subskills (Papageorgiou et al., 2021, p. 88):

- Understanding main ideas and important details
- Recognizing a speaker's attitude or function

- Understanding the organization of listening material
- Understanding the relationships between ideas presented
- Making inferences or connections between pieces of information

Listening Construct for DET

In their white paper on listening for Duolingo, Goodwin and Naismith (2023) emphasized the integration of listening with other language skills, stating that, since listening is "inherently integrated with the other skills of speaking, writing, and reading" (p. 3), for testing purposes "listening ability as part of integrated modalities such as speaking-listening ability should be included in construct definition" (p. 3). However, they do not provide a statement of the theoretical construct of listening the test is intended to measure. That is, unlike the TOEFL listening construct presented above, there is no statement that describes parameters for the listening input and how test takers demonstrate their listening ability, which then inform the specifications for listening texts and tasks. Instead, Goodwin and Naismith (2023, p. 13) provided an overview of how each existing listening item type on the test addresses specific listening subskills, processes, and attributes, based on a recent review of the literature on listening assessment (Aryadoust & Luo, 2023). The listening subskills that are included, which combine subskills from Aryadoust & Luo's (2023) taxonomy with descriptors from the CEFR, include the following:

- Listening for specific information
- Listening for detailed understanding
- Understanding local linguistic meaning
- Listening for gist
- Listening for implication or inference
- Communicative listening ability
- Integrated listening skills

A comparison of these two constructs reveals some overlap and some areas of difference. Both tests target the skills of listening for main ideas or gist, listening for specific information, and making inferences. TOEFL iBT focuses more on skills needed for listening to extended discourse and also makes reference to essential subdomains of academic listening. DET, on the other hand, includes communicative listening ability (i.e., two-way listening as described above)

and integrated listening skills but does not specifically mention the domains of language that are relevant to academic contexts.

Listening Test Content

In this section, I first describe the content of the listening sections and item types of both tests in terms of the overall structure and features of the test tasks. Sources for this information include descriptions on the respective websites, practice tests, and published articles; that is, information available to the general public. For TOEFL iBT, I downloaded the practice test available online (ETS, n.d.[a]) and listened to official practice materials (ETS, n.d.[b]). For DET, I looked at items in the official user's guide (Duolingo, 2024), took two practice tests, asked a research assistant to take another practice test, and recorded these tests.

TOEFL iBT has a dedicated listening section, which comprises five monologic or dialogic texts (lectures and conversations), followed by comprehension questions, which are primarily multiple-choice questions but also include drag and drop or grid items. The listening score contributes to the total test score, and a separate listening scaled score is provided as well. The listening section takes 36 minutes in total, of which approximately half consists of actual listening and the rest is responding to the questions.

The speaking and writing sections of the TOEFL iBT include some items that require students to listen and read, as well as write or speak, but as these items do not contribute to the listening subscore, I will not discuss them here.

The DET, on the other hand, does not have a dedicated listening section and, until 2024, did not report a separate score for listening. Instead, items that involve listening contribute to the scores for comprehension and conversation, as discussed in the introduction, and the total score.

There are two item types on the DET that directly involve listening. One is a dictation task, where test takers hear a single sentence and have to transcribe it. They are allowed to listen to the sentence up to three times. The dictation task is adaptive, and test takers encounter six to nine sentences during the test.

The second item type is interactive listening. In this task, test takers are given a scenario such as seeking advice from a professor or a fellow student. Over 8 to 10 turns of a conversation, the test taker hears the interlocutor's turn and then has to choose the most appropriate response. After each item, the correct response is provided in writing, along with the transcript of the previous turn, and then the student listens to the next turn. At the end, the test taker has 75

seconds to write a summary of the conversation. Test takers encounter two of these tasks on a test. The total testing time for these two item types (dictation and interactive listening) ranges from around 18 to 23 minutes, of which perhaps 5 to 6 minutes is dedicated to listening.

The listening tasks for the two tests are summarized in Table 2. As the table shows, perhaps the most striking difference between the two tests is the amount of time devoted to actual listening. TOEFL iBT listening texts range from 3 to 5 minutes in length, for a total of about 16 minutes of listening to both lectures and conversations. In contrast, the two DET listening item types only require listening to much shorter texts, either single sentences in the dictation section or one to three sentences in the interactive listening section.

Table 2. Description of TOEFL iBT and DET Listening Task Types

Characteristic	TOEFL iBT Listening	DET Dictation	DET Interactive Listening
Task description	Test taker listens to a text and answers written comprehension questions	Test taker listens to a single sentence and transcribes it	Test taker listens to turns from one side of a conversation and chooses the most likely response.[a]
Number of tasks	5 (2 conversations, 3 lectures)	6–9	2
Number of items per task	5–6	1	5–7
Item type	Selected response comprehension questions (multiple-choice)	Transcription	Selected response (choose most appropriate response to previous turn)
Length of listening passage	Conversations: 500–600 words; around 3 minutes Lectures:700–800 words; around 5 minutes	5–20 words (under 10 seconds)	8–50 words (under 30 seconds)
Total time for section	36 minutes	6–9 minutes (1 minute per item)	14–16 minutes
Total listening time	16 minutes	2–3 minutes	5–6 minutes

[a]At the end of the interactive listening task, test takers write a short summary of the entire conversation. However, since the written transcript of conversation is presented to the test taker, I am not considering this to be a listening task.

Turning now to the specific characteristics of the listening texts, summarized in Table 3, it can be seen that the test designers have made different decisions about how to present the listening input and what kinds of support to provide listeners. Of the many factors that distinguish academic listening from other listening domains, two relevant ones are the accent of

the speaker and the content domains. TOEFL iBT includes a somewhat wider range of accents (see Ockey & French, 2016) than DET, though neither test includes the range of accents that students at North American universities are likely to encounter. As for topics, TOEFL iBT specifically targets academic-content and academic-navigational domains. The interactive listening task on DET, which was a fairly recent addition to the test, does focus on the academic-navigational domain, with conversations between students and professors or students and other students, often regarding seeking advice on academic topics. The dictation items, on the other hand, tend to be much more general and seem to be generated to include (at the higher levels) sophisticated syntactic structures rather than to represent natural oral texts. For example, the three most complex sentences in the DET users' guide (Duolingo, 2024, p. 76) are the following:

- Finally, the results of this investigation were published in a scientific magazine.
- Even without seeing you, I would have recognized you by the sound of your voice.
- They will have tried to talk to you by the time the story has published.

Although the second and third sentences have some characteristics of oral discourse, such as first and second person pronouns, the first seems to be a sentence that would be more likely to be read than heard.

Table 3. Characteristics of the Listening Texts

Characteristic	TOEFL iBT Listening	DET Dictation	DET Interactive Listening
Visual support	Context visual; some content visuals	No visual support	Context visual (generic cartoon avatar)
Memory support	Notetaking allowed	May be repeated up to three times	Correct answer and transcript of spoken turn provided after each item
Accent	North American, British, Australian, New Zealand	American	American
Gender	Mix of male and female	Mix of male and female	Mix of male and female
Speaker	Voice actor	Voice actor	Computer generated voice
Topics	Academic content Academic navigational	General	Academic navigational

Cognitive Processes

One of the propositions of the TOEFL validity argument (ETS, 2020, p. 5) is that "Academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks." In this section, I provide a brief discussion of the cognitive processes in academic listening, primarily relying on Field's (2013) model of academic listening. This model includes five levels of processing: input decoding (identifying word boundaries in the input), lexical search, parsing (recognizing the syntactic structure of phrases and sentences), constructing meaning (i.e., of individual propositions in the input), and constructing discourse (making sense of the whole). According to Field, the process of discourse construction consists of four aspects: choose, connect, compare, and construct. The listener has to *choose* what to attend to in the listening input by considering whether the segment is relevant and important to the listener's goals, *connect* it to the previous utterance, *compare* it to determine whether it is consistent with what has been said so far, and *construct* an overall sense of the speaker's main points. This process involves making inferences beyond the propositions of the text, including inferences about the speaker's purpose or the pragmatic intent of a given utterance.

From this list of listening skills, it can be argued that the lower level skills of decoding the input, parsing, and constructing meaning from individual utterances are covered by the listening tasks from both TOEFL iBT and DET. However, the skill of discourse construction can only be assessed through longer listening texts, which are only found on TOEFL iBT. Some examples from each test illustrate this point.

In the TOEFL iBT listening test, items are designed to assess the degree to which a listener can understand the speaker's intention (see Figure 1). To answer this question correctly, the listener needs to understand both how this utterance fits into the overall conversation (in which the student has offered assistance for an event but the professor has turned him down) and also interpret the implied meaning of the words and phrases contained in the utterance. It is important to note here that all four options are plausible to someone who has not listened to or understood the conversation up to this point.

Figure 1. TOEFL iBT Listening Item

Narrator Listen again to part of the conversation. Then answer the question.

Female Professor There's not much glory in it, but we're looking for someone with some knowledge of anthropology who can enter the articles…I hesitate to mention it, but I don't suppose this is something you would...

Why does the professor say this:

Female Professor I hesitate to mention it, but I don't suppose this is something you would...

A. To express doubt about the man's qualifications for the project
B. To ask the man if he would be willing to work on the project
C. To ask the man to recommend someone for the project
D. To apologize for not being able to offer the project to the man

Compare this to an interactive listening passage encountered in a practice test on the DET. The first item involves reading a scenario, then listening to the first turn and selecting an appropriate response to the speaker from written options. In the scenario, the test taker is told to take on the role of student asking a friend for help deciding whether or not to study abroad. The first turn (spoken by an avatar) is a greeting from the friend. The possible responses include one that essentially repeats the scenario ("I'm trying to decide whether to study abroad"), one that implies that the friend is thinking about studying abroad, and one that refers to having recently studied abroad "I just got back from my semester abroad.") In this example, test takers do not need to hear the opening turn to choose the correct answer, as long as they can read the scenario. Some understanding of conventional conversational openers is involved, but there is only one response that is relevant to the scenario.

The transcript of the first turn and the correct option are then presented to the test taker, who listens to the second turn, which is followed by five written options. Of these, two can be eliminated right away because they are not on the topic of study abroad (one is about getting into graduate school and one is about a thesis topic).

Similar issues occurred on most of the interactive listening tasks I was presented with during several practice tests. I should note furthermore that the test taker has control over when to play the speaker's next turn. It is thus difficult to argue that this task assesses listening per se, since many of the answers can be derived from reading without actually comprehending much of the aural input.

Discussion

As discussed in the introduction, this section of the paper addresses the following three propositions in TOEFL validity argument (ETS, 2020, p. 5):

- The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings.
- Tasks and scoring criteria are appropriate for obtaining evidence of test takers' academic language abilities.
- Academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks.

Based on these propositions, it can be concluded that, to the extent possible, a test of academic listening should have the following qualities:

- Listening input that is representative of the kinds of oral texts that students will encounter in academic settings.
- Listening tasks (what students will do with the listening input; i.e., how students respond to the listening) will be relevant to academic listening tasks.
- The cognitive processes and strategies involved in processing the aural input will be similar to those needed for essential academic listening encounters.
- Scoring of the responses will provide evidence of student listening abilities that will be useful in making decisions (e.g., for admission and/or English language support).

I now compare both tests in terms of these qualities.

Listening Input

As noted above, academic listening involves both one-way and two-way listening. The input texts for TOEFL iBT include both monologues and dialogues, representing both lectures and academic conversations. Furthermore, the texts comprise relatively lengthy stretches of discourse (up to 5 minutes) and sample from the subdomains most important to academic listening, particularly academic content and academic navigational. The texts are long enough to address test takers' ability to process extended discourse and include characteristics of both monologic and dialogic discourse, including a fair amount of academic language.

The texts are scripted and spoken by professional voice actors, so while they include such features of oral language as false starts and hesitations, concessions have been made to the testing situation so that they may be somewhat slower and more clearly articulated than genuine

oral communication. For similar reasons, the TOEFL iBT listening texts do not represent the range of accents that students in North American countries are likely to encounter. Finally, on TOEFL iBT, there are no listening texts that require demonstration of listening in interaction, perhaps resulting in what Wagner (2022) terms an "impoverished construct" (p. 222). However, some of the listening items in the conversation do require understanding of conversation structure and implicatures, which addresses interactional competence at least indirectly.

The input texts for the DET dictation task do not appear to be sampled from relevant academic domains and, in fact, are not always representative of oral discourse, as discussed previously. Furthermore, they consist of single sentences with no surrounding context. The interactive listening texts, on the other hand, are intended primarily to represent the domain of academic-navigational texts and do require test takers to select appropriate contingent responses, which is an attempt to measure "interactional competence" (Cardwell et al., 2024b, p. 10) within the constraints of a formal test. However, as also discussed previously, many of the items do not necessarily require understanding the spoken text, as test takers can often infer the correct answer from the provided scenario or from eliminating clearly irrelevant distractors. Furthermore, the voices are computer generated and represent American accents only. Thus, the DET construct is equally "impoverished," to use Wagner's (2022) terminology, and is arguably more so given its complete lack of extended discourse.

Listening Tasks and Cognitive Processes

The listening items on TOEFL iBT are primarily multiple-choice comprehension questions targeting a variety of listening skills as noted previously. While these questions necessarily involve reading and thus cannot be said to assess listening skills only, they do target higher order discourse processing skills such as inferencing and making connections between ideas. Furthermore, the questions cannot easily be answered without having understood the listening. In addition, answering multiple-choice questions is an academic skill that most students, particularly in U.S. universities, must master.

The dictation task on the DET is, on its surface, not an academic task, as students are not generally called upon to transcribe oral texts word for word unless they are conducting certain kinds of qualitative research. However, the task can be said to measure lower level listening skills such as decoding and parsing. The interactive listening task requires selection of an appropriate contingent response that is relevant to the previous turn in a conversation and to the

discourse as a whole. As such, the task does draw on test takers' knowledge of pragmatics and interactional competence. However, as discussed previously, many of the items can be answered without actually listening to or comprehending the utterance. Thus, as a measure of listening per se, this task is somewhat questionable.

Scoring

The TOEFL iBT score report includes both a total score based on all four subtests and a separate score for listening based on the listening section. This listening score can be useful for decision making; for example, Wagner (2016) found that the listening subtest score was a better predictor of teaching competence for international teaching assistants than the speaking subscore. Before July 1, 2024, DET did not report a separate listening score and only reported integrated scores; that is, scores on the two listening tasks contributed to the conversation and comprehension scores. As noted previously, the DET now reports a separate listening score; according to Cardwell et al. (2024b), this score is based on both the dialogue completion and summarization tasks within the interactive listening task, the extended speaking task (audio prompt), and the dictation task.

If the DET listening score is, in fact, based solely on the scores for the six to nine dictation tasks encountered by a test taker and their correct answers on the integrated listening tasks (many of which do not in fact require listening), then the score represents at best the ability to understand single sentences. Given the importance of listening to academic success, as discussed at the beginning of this section, these listening scores should be interpreted with caution.

Both TOEFL iBT and DET claim to assess listening, but the construct is defined differently in the two tests, and this is reflected in how listening is assessed in each test. TOEFL iBT hews to a more robust construct of academic listening, providing samples of extended listening texts on academic topics with multiple-choice and other selected response formats for assessing comprehension. While the focus is on one-way listening, the test includes both lectures and conversations, thus covering the construct of one-way listening fairly completely within the limitations that are necessitated by the constraints of large-scale tests. The test items address aspects of listening that go beyond simple sentences and require test takers to comprehend fairly lengthy stretches of discourse.

DET, on the other hand, assesses listening in a very limited way. The digital-first orientation of the test creates challenges for designing listening tasks, perhaps more than any other skill area. The dictation task does provide baseline information on low-level listening skills, and the interactive listening task taps into aspects of interactional competence that are arguably important for academic success but are not always tied to listening itself. Furthermore, the ability to comprehend extended spoken discourse is an important aspect of academic listening, and this ability is not assessed on DET. Thus, at this stage of its development, DET is not adequate for reliably measuring the listening skills required for academic success.

Reading

Defining the Reading Construct

Reading academic texts is an essential skill for success in postsecondary education and has been a central focus of testing language for academic purposes for decades. Furthermore, reading is the main mode through which discrete language skills such as vocabulary and grammar are most easily assessed. It is not surprising, then, that reading plays a large role in both TOEFL iBT and DET and has traditionally been one of the main foci of tests for academic purposes.

Hermida (2009) describes successful academic reading as a "deep approach" in which a reader

> uses higher order cognitive skills such as the ability to analyze, synthesize, solve problems, and think meta-cognitively in order to negotiate meanings with the author and to construct new meaning from the text. The deep reader focuses on the author's message, on the ideas she is trying to convey, the line of argument, and the structure of the argument. The reader makes connections to already known concepts and principles and uses this understanding for problem solving in new contexts. (p. 2)

This view of reading for academic purposes is consistent with the views expressed by Liu and Read (2023) in a recent literature review. Fluent reading of individual texts involves the lower level skills of word decoding/recognition, parsing, and constructing sentence meaning, along with higher level skills such as constructing a text model (comprehending the main ideas and specific details explicitly found in the text) and a situation model (creating a mental

representation of the text meaning that includes information implied in the text or drawn from the reader's world knowledge; see Kintsch, 2013). Other cognitive processes involved in reading include inferring the writer's pragmatic communication (i.e., goals, intended audience, and stance toward the information in the text) and rhetorical organization such as genre (Graesser & Forsyth, 2013).

While tests of academic reading have traditionally focused on close reading of individual texts, Liu and Read (2023) maintain that a fuller construct of academic reading should also include synthesizing information across texts, critically evaluating source materials, and reading strategically (i.e., shifting reading speed according to reading purpose).

Reading Construct for TOEFL iBT

The TOEFL 2000 Reading Framework (Enright et al., 2000), which formed the basis for the current TOEFL iBT reading section, takes a reading purpose perspective on the assessment of reading: reading to find information (search reading), reading for basic comprehension, reading to learn, and reading to integrate information across texts. Each of these purposes entails a different set of skills that build on each other as the purpose gets more complex. Reading to find information involves word recognition, working memory, and a fluent reading rate. Reading for comprehension involves being able to construct a text model and an appropriate situation model, which may involve cycling through the text several times to integrate information from different parts of the text. Reading to learn involves creating a more elaborated text model, which takes into account the rhetorical structure and author's purpose, and involves deeper processing of the text. Finally, reading to integrate information requires all of the above across multiple texts and the development of an organizational frame that is not explicitly stated in any text.

The authors of the framework paper go on to map a variety of reader tasks onto these purposes, as follows:

- Reading to find information and reading for basic comprehension: identify/interpret
- Reading to learn: summarize, define/describe/elaborate/illustrate
- Reading to integrate: compare/contrast/classify, problem/solution, explain/justify, persuade, narrate

In an update of the original framework paper, Schedl et al. (2021) provide a model of academic reading comprehension assessment that consists of three components: reader purpose/goals, characteristics of texts, and reader linguistic and processing abilities. Reader

purpose/goals include finding information, general comprehension, learning from texts, evaluating information, and integrating information across texts. The characteristics of text components includes text type, rhetorical structure, and text features. Reader linguistic and processing abilities include awareness of text structure, background knowledge, engagement of comprehension strategies, morphological and phonological knowledge, semantic knowledge, syntactic knowledge, text processing abilities, word and sentence recognition/vocabulary knowledge, and working memory (efficiency)/pattern recognition.

Reading Construct for DET

Prior to 2022, DET did not include any tasks that assessed reading other than the modified C-test as described previously. As of 2024, the main reading task on DET is the interactive reading task, introduced in Park et al. (2022) and described below. According to Park et al. (2022, p. 3), DET "envisions the construct of reading both in terms of the purposes with which the test takers read and in terms of the cognitive processes employed while reading (Chapelle, 1999), all in a way that is relevant in academic contexts." In the latest version of the DET manual, the reading construct for DET is defined as "comprehending written English from basic informational texts to advanced expository/persuasive texts at CEFR levels A1–C2" (Cardwell et al., 2024b, p. 5).

Park et al. (2022, p. 4) outlined the intended construct for the DET interactive reading task and mapped the item types onto reading purposes and activated cognitive skills. Reader purposes include reading to search for information, for quick understanding, and for main ideas; to learn; to integrate; and to use information. Activated cognitive skills include search processes, strategic processing abilities, fluency and reading speed, main ideas comprehension, text structure awareness, discourse organization, summarization abilities, synthesis skills, evaluation and critical reading, and inferences about text information. Comparing the information here with the discussion of TOEFL iBT Reading, it can be seen that both models include reading purposes and cognitive skills in reading, with a fair amount of overlap. However, the TOEFL iBT Reading model includes a consideration of text characteristics, which is not present in the DET model.

Reading Test Content

The reading section of the TOEFL iBT consists of two reading passages with 10 questions each. Reading passages are approximately 700 words long, and the question types for each passage come from the following (see ETS, n.d.[c] for details):

- Factual information questions (recognize information explicitly stated in the text)
- Inference questions (identify or understand information not explicitly stated in the text)
- Vocabulary (identify the meaning of words or phrases as they are used in a specific reading passage)
- Sentence simplification (choose a sentence that means the same as a sentence from the reading passage)
- Insert sentence (demonstrate understanding of passage organization by inserting a sentence somewhere in a paragraph of the passage)
- Prose summary (choose three statements that express the most important ideas in a passage)

Beyond the reading section itself, additional reading texts are found in other sections of the test, specifically the integrated speaking and writing tasks. In the sample test provided by ETS, two of the speaking tasks included readings between 80 and 90 words, the integrated writing task included a reading passage of 329 words, and the academic discussion task involved reading three discussion posts ranging from 50 to 75 words. Thus, apart from the reading section, TOEFL iBT test takers encounter an additional 600 to 700 words of extended text, for a total of at least 2,000 words, not counting instructions and test items.

The DET includes two tasks that involve reading texts longer than a single sentence: the read and complete task, which is a modified C-test, and the interactive reading task. Like most C-tests, in the read and complete task the test taker is presented with a short reading passage in which the first and last sentence are complete. In the other sentences, the second half of every other word is deleted (for an example of the task, see Cardwell et al., 2024b, p. 9). Test takers have 3 minutes to complete each passage and encounter between three and six of these items during the test.

The interactive reading task consists of two stages. In the first stage, test takers complete a short multiple-choice, gap-filling task with around 10 words deleted from a text of about 150 words. This text is the introduction to a slightly longer text, which is displayed in the second stage with the blanks filled in correctly but one of the additional sentences gapped. The test taker completes four additional tasks with this text, as follows:

- Select the best sentence to complete the passage (multiple choice).

- Highlight a section of the test that contains the answer to a detail question (2 items).
- Select an idea that is contained in the passage (multiple choice).
- Select the best title for the passage (multiple choice).

Table 4 below summarizes the reading tasks for both tests.

Table 4. Description of TOEFL iBT and DET Reading Task Types

Characteristic	TOEFL iBT Reading	DET Read and Complete	DET Interactive Reading
Task description	Test taker reads academic passages and responds to comprehension and vocabulary questions.	Test taker reads and completes a short C-test passage.	Test taker responds to six item types based a single text.
Number of tasks	Two passages	3–6	2
Number of items per task	6–10	10–15	6
Item type	Selected response comprehension questions (multiple-choice) • Recognize factual information • Recognize implied information • Identify the meaning of words in the text Select a shorter sentence that has the same meaning as a sentence in the text • Insert a sentence into the text • Identify three main ideas from the text	Modified C-test (fill in second half of every other word)	• Complete sentence with gapped words • Complete paragraphs with gapped sentence • Locate the answer to a comprehension question • Choose the idea that is present in the text • Choose the best title for a text
Length of each reading passage	Approximately 700 words	Between 100 and 250 words	Approximately 225 words
Timing	36 minutes for 2 passages with 10 multiple choice questions and 1 selected response summary item (38 words/minute excluding items)	3 minutes per text	8 minutes for each passage with 6 questions (28 words/minute excluding items)
Total time for section	36 minutes	9–27 minutes	16 minutes

Text Characteristics

An important consideration in a test of English for academic purposes is the nature of the texts to be read. For this section, I compared the samples of the test that are available online as

sample tests (https://englishtest.duolingo.com/practice) and only included the interactive reading task from DET. Texts are compared in terms of passage length, topic/genre, readability statistics (Flesch reading ease and Flesch-Kincaid Grade Level calculated by Microsoft Word) and the degree of support given (glosses). These features are summarized in Table 5. As the table shows, TOEFL iBT texts are up to three times longer than DET texts (close to 700 words as opposed to 220) and consist of academic/informational texts, whereas some DET texts are more narrative/journalistic in nature. The TOEFL iBT texts are more linguistically complex as well, with lower reading ease scores and more passive sentences. One caveat to this observation is that the DET texts provided in the online practice test may not be representative of the most challenging texts presented to actual test takers, since the test algorithm provides tasks of different difficulty levels to test takers based on their scores on previous sections of the test.

However, the versions of the task I encountered in practice DET tests were similar or easier than the ones described in the table below. Finally, some technical terms in TOEFL iBT reading texts are glossed for the reader.

Table 5. Characteristics of the Reading Texts

Characteristic	TOEFL iBT (sample test downloaded)	DET Interactive Reading (practice test online)
Length of passage	Passage 1: 669 words (excluding title and glosses) Passage 2: 692 words	Passage 1: 221 words Passage 2: 224 words
Topic/genre	Passage 1: Academic/informational (environmental problem/solution) Passage 2: Academic/informational (ancient history)	Passage 1: Academic/informational (environmental problem/solution) Passage 2: Narrative/journalistic (inspirational story about a tutor's effect on her student)
Text characteristics:	Passage 1: Average sentence length: 24.8 Flesch reading ease: 30 Flesch-Kincaid Grade Level: 15.2 Percent passive sentences: 34.6% Passage 2 Average sentence length: 18.7 words Flesch reading ease: 44.7 Flesch-Kincaid Grade Level: 10.8 Percent passive sentences: 21.6%	Passage 1: Average sentence length: 17.0 Flesch reading ease: 58.6 Flesch-Kincaid Grade Level: 9.3 Percent passive sentences: 15.3% Passage 2: Average sentence length: 20.3 Flesch reading ease: 33.2 Flesch-Kincaid Grade Level: 13.6 Percent passive sentences: 0%
Text support given	Important vocabulary glossed	None

Cognitive Processes

One important consideration for reading is the cognitive processes involved in the reading task. For the purpose of this analysis, I have chosen to use the academic reading framework from Liu and Read (2023). This framework was chosen for several reasons. First, it was based on an analysis of reading needs for academic purposes that included both importance and perceived difficulty of skills, as well as the feasibility of including skills in a test. Second, it was not designed with either TOEFL iBT or DET in mind, so it is, in that sense, neutral. The results are summarized in Table 6.

Table 6. Cognitive Processes in TOEFL iBT and DET Reading

Cognitive Process	TOEFL iBT Reading	DET C-test	DET Interactive Reading
Core academic language knowledge			
Understand general academic vocabulary	X	X	X
Understanding single sentences with complex structure	X	X	X
Careful reading for intra-textual model building			
Integrating textual information across sentences	X		X
Inferring the situation (environment, event and relationship) implied in a text	X		
Understanding author's point of view (such as attitudes, beliefs, and opinion)	X		
Inferring the contextual meaning of figurative language	X		
Careful reading for intertextual model building			
Understanding the relationships between multiple texts	(X)		
Drawing implications/conclusions based on multiple texts	(X)		
Expeditious reading			
Searching for specific meaning	X		X
Skimming for general idea	X		X

Note. The integrated speaking and writing tasks draw on multiple texts. An X inside parentheses [(X)] notes that this skill is not assessed in the reading section of TOEFL iBT.

As the table shows, both tests assess core academic and expeditious reading strategies. Both tests also assess integrating textual information across sentences. However, only TOEFL iBT includes items that assess other important skills involved in careful reading for intratextual model building (making inferences and understanding the author's point of view).

An important limitation for any test of reading is the effect that the test items themselves have on the reading process. In particular, some research suggests that test takers tend to limit

their close reading to the portions of a text that are likely to contain the answer, rather than attempting to create a model for the whole text (Rupp et al., 2006). Furthermore, success on multiple-choice reading items depends not only on the text itself but on the effectiveness of the test items, including the distractors. A recent systematic review of the literature on constructing and diagnosing distractors in multiple-choice items (Gierl et al., 2017) provides a list of the most common recommendations for writing effective distractors for multiple-choice items. The most frequent recommendation is to base distractors on "identifying common misconceptions related to thinking, reasoning, and solving the problem" (Gierl et al., 2017, p. 1102).

According to Park et al. (2022), distractors for the DET Interactive Reading task are not created in this manner. Rather, distractors are created by automatically generating several related passages and then using sentences/titles from these other passages as distractors for the passage completion, main idea, and title. This development process can result in several implausible distractors, making the items themselves easier than they might otherwise be. For example, the DET official guide (Duolingo, 2024, p. 23) shows a "select the idea" item from a narrative passage about John blowing a fuse while rewiring his house. Only one of the options mentions a fuse; the others refer to the fax machine, the remote control, or extension cords, none of which appear in the text. It is not clear exactly what this item is testing beyond the ability to scan the passage and pick the option that includes words from the passage.

Compare this DET reading item to a TOEFL iBT reading item found in the official sample test (Figure 2). The figure includes the first paragraph of the passage and one of the items.

Figure 2. TOEFL iBT Reading Item

READING:

A topic of increasing relevance to the conservation of marine life is bycatch—fish and other animals that are unintentionally caught in the process of fishing for a targeted population of fish. Bycatch is a common occurrence in longline fishing, which utilizes a long heavy fishing line with baited hooks placed at intervals, and in trawling, which utilizes a fishing net (trawl) that is dragged along the ocean floor or through the mid-ocean waters. Few fisheries employ gear that can catch one species to the exclusion of all others. Dolphins, whales, and turtles are frequently captured in nets set for tunas and billfishes, and seabirds and turtles are caught in longline sets. Because bycatch often goes unreported, it is difficult to accurately estimate its extent. Available data indicate that discarded biomass (organic matter from living things) amounts to 25–30 percent of official catch, or about 30 million metric tons.

1. According to paragraph 1, which of the following is true about the impact of various methods of fishing on the problem of bycatch?

(A) Almost all commercial fishing methods capture fish and animals that the fishers do not want.
(B) Switching from trawling to longline fishing would save seabirds and turtles from being unintentionally caught.
(C) Longline fishing is particularly dangerous for dolphins and whales.
(D) Trawling on the ocean floor produces less bycatch than does trawling through mid- ocean waters.

This item in Figure 2 requires a fairly sophisticated understanding of the passage as a whole, and all the distractors contain words from the passage. Hence, this item appears to assess comprehension at a deeper level than the DET items, which include automatically generated distractors.

Discussion

The discussion of reading parallels the discussion of listening, in that, based on propositions from the TOEFL validity argument, a test of academic reading should have the following qualities:

- Reading input will be representative of the kinds of written texts that students will encounter in academic settings.
- Reading tasks (what students will do with the reading input (i.e., how students respond to the reading) will be relevant to academic reading tasks.
- The cognitive processes and strategies involved in processing the written input will be similar to those needed for academic reading.
- Scoring of the responses will provide evidence of student reading abilities that will be useful in making decisions (e.g., for admission and/or English language support).

Reading Input

It is clear from Table 5 that the reading input for TOEFL iBT is more extensive and more complex than the reading input in DET, with two to three times the amount of text that needs to be processed. Furthermore, the reading texts in TOEFL iBT frequently contain technical terms that are either defined or glossed in the reading passages, which is how technical terms are frequently handled in academic texts.

Reading Tasks

As discussed, the reading tasks on TOEFL iBT also address a wider range of reading skills that are important for academic reading, particularly in terms of inferencing and addressing the author's purpose. Furthermore, the item types seem to target relevant subskills more precisely in TOEFL iBT, in part because the distractors for DET items are automatically generated from similar passages rather than being based on potential misunderstandings of the text itself.

Scoring

As noted above, before late 2024 a separate score for reading was not reported by DET; rather, scores on the C-test and interactive reading tasks (along with the yes/no vocabulary and vocabulary in context tasks) contributed to the scores for comprehension and literacy. The interpretation of this score must be informed by the nature of the tasks that are included, which are heavily weighted toward individual words and sentences and not comprehension of extended texts.

Both TOEFL iBT and DET assess reading, but there is a marked contrast between the approaches to reading assessment taken by the two test developers. The passages in TOEFL iBT are much longer and more academic in nature than the DET interactive reading passages, and the total amount of reading in TOEFL iBT is much greater than DET. TOEFL iBT also includes more items that measure inferencing, discerning the author's purpose or stance, and other reading skills that are essential for academic reading. The recent addition of the interactive reading task in DET is a step in the right direction, but the test still emphasizes lower level skills and vocabulary over higher level, complex reading skills and, thus, remains limited in its ability to assess academic reading.

Writing

Defining the Writing Construct

An essential feature of writing for assessment purposes is that test takers must "produce coherent, comprehensible texts" (Cumming et al., 2021, p. 108); that is, the ability to write cannot be assessed indirectly through the assessment of the many subcomponents of writing without the production of an actual text. Among several useful frameworks for discussing writing for assessment purposes is that of Shaw and Weir (2007), who posit six main cognitive processes for writing:

- Macro-planning, or gathering ideas and identifying the constraints of the writing task such as genre, purpose, and audience
- Organization, or identifying relationships among ideas, putting them in order, and prioritizing them in terms of how important they are to the main idea of the writing
- Micro-planning, or planning out language output at both the sentence and paragraph level
- Translation from abstract ideas into linguistic form
- Monitoring, or evaluating the text for mechanical accuracy and adherence to the writer's intention and intended argument structure
- Revising, or making adjustments or corrections to the ongoing text as a result of monitoring

The test construct is also reflected in the rating criteria used to score writing, as the characteristics of writing that are valued are those that are scored.

One of the main features that distinguishes academic writing is the ability to write from sources. Cumming et al. (2021) argued that "most students' writing for academic purposes involves them displaying (and ideally, also showing evidence of them transforming) their knowledge in direct relation to the content and contexts they have been studying, reading, hearing about, and discussing in academic courses" (p. 113). Furthermore, Cumming et al. (2016) stated that "educators around the world would agree that learning to write effectively from sources is a fundamental academic literacy skill" (p. 47).

From these principles it can be implied that a test of academic writing should, at minimum, be long enough to start with drafting, provide opportunities for planning and revising, and include at least one task that requires accurate reporting from sources, either through reading or listening. The writing prompt should also elicit sufficient writing of the type that can be evaluated using the criteria important to the test developers.

Writing Construct for TOEFL iBT

Pearlman (2008, p. 253) states the following claim about writing in TOEFL iBT: "Test taker can communicate effectively in writing in English-language academic environments" along with two subclaims:

- Can formulate and communicate ideas in writing on a variety of general topics, producing extended, organized written text expressing and supporting his/her own

opinions based on knowledge and experience, taking into account the knowledge of the intended audience

- Can coherently organize and accurately express in writing the content and structure of academic discourse, demonstrating an understanding of key ideas on an academic topic as presented in reading and lecture formats and the rhetorical relationships such as claim/rebuttal, problem/solution, and proposal/counter proposal that link the information in these texts

Writing Construct for DET

The construct for DET writing is expressed most concisely in Cardwell et al. (2024b, p. 11): The open-ended writing tasks on the DET are intended to "elicit written responses that evidence writing proficiency in terms of the writing subconstructs of content, discourse coherence, grammar, and vocabulary and proficiency in discussing topics in the different domains described in the CEFR (personal, public, educational, and professional)."

Writing Test Content

The TOEFL iBT writing section consists of two tasks: an integrated writing task and an academic discussion task. For the integrated writing task, the test taker reads a short passage on an academic topic, then listens to a brief lecture or conversation about the same topic and must respond to a written prompt that requires use of information from both sources for an effective response. Test takers are given 20 minutes to complete their response.

The TOEFL iBT Writing section was revised in 2023 to eliminate the traditional independent task and replace it with an academic discussion task. Although this new task is substantially shorter than the independent task (10 minutes as opposed to 30), it is intended to measure essentially the same construct: "the test-taker's ability to create a short piece of writing in English that expresses their ideas in a clear and coherent way" (Davis & Norris, 2023, p. 9).

TOEFL iBT writing tasks are scored by a combination of human raters and automated scoring using rubrics that have been well publicized. Writing scores are transformed to a standard score between 0 and 30. The TOEFL website provides samples of writing that have achieved high scores so that test takers can have a model to follow.

DET writing tasks include a sentence dictation task and several open-ended writing tasks. The open-ended tasks include a picture description task, an interactive writing task, and a task simply labeled "Writing Sample," which appears to be a 5-minute prompt-based writing task.

The DET interactive writing task, first used on the test in 2024, consists of two stages. The first stage is a prompt-based writing task (presumably identical to the so-called Writing Sample task), and in the second phase, there is an automatically generated prompt to add additional information. The second prompt is generated after an automated analysis of the themes included in the original piece of writing using a list of predetermined themes associated with the topic. As Goodwin et al. (2024) explain:

> If the writing prompt were*, Describe the last time you did something that challenged you. What did you do? What did you learn from the experience?,* example themes could include Navigating Failure, Developing Problem-Solving Skills, or Applying Lessons to Future Challenges. Each of these themes is further associated with a follow-up prompt that asks the test taker to discuss the theme in relation to the topic of the initial prompt, e.g., *Discuss how this challenging experience required you to develop or use problem-solving skills. Describe the strategies you used and how they helped you approach this task.* (p. 10)

Writing on the two tests is summarized in Table 7.

Table 7. Comparison of Writing on TOEFL iBT and DET

Characteristics	TOEFL iBT	DET
Number of tasks	2	5
Task description and timing	Integrated writing: writing based on reading and listening short texts (20 minutes) Academic discussion; state and support an opinion in an online classroom discussion (10 minutes)	Picture description: write about a photo (3; 1 minute each) Interactive writing: write to a short prompt (5 minutes) and a follow-up based on suggested related theme (3 minutes) Writing sample: write to a short prompt (5 minutes)
Expected genre	Integrated writing: informational Academic discussion: opinion	Picture description: Description Interactive writing/writing sample: narrative, opinion
Domain	Academic	Public Personal Academic Professional
Total writing time	30 minutes	16 minutes
Expected length	Integrated writing: 150–225 words Academic discussion: At least 100 words	No guidance given
Scoring	Human and automated scores; published rubric	Automated scores only

Cognitive Processes

Returning to Shaw and Weir's (2007) framework for writing discussed previously, we can evaluate the two tests in terms of the cognitive processes involved in completing the writing tasks (see Table 8).

Table 8. Comparison of Cognitive Processes in TOEFL iBT and DET Writing

Cognitive process	TOEFL iBT	DET
Macroplanning	X	
Organization	X	
Micro-planning	X	X
Translation	X	X
Monitoring	X	X
Revising	X	

It can be argued that the short time allowed for writing on DET (maximum 5 minutes per task) does not allow for processes beyond micro-planning, translation, and monitoring, whereas the longer time allowed on TOEFL iBT, particularly in the integrated writing task, permits all six of these processes. Furthermore, additional cognitive processes involved in source-based writing were identified by Plakans (2009), including selecting relevant parts of source texts to include in writing, paraphrasing, and connecting ideas from source texts with the writer's own experience (see Cumming et al., 2021, for a further discussion). The discussion task also provides a realistic academic audience and purpose for writing, so that the writer needs to take into account what the assumed reader already knows about the topic and what the expected tone of the writing should be (Davis & Norris, 2023; Papageorgiou et al., 2021). These skills are essential for academic writing and are not addressed in DET.

All DET writing tasks are scored automatically using a proprietary automated scoring system. According to the DET technical manual, the scoring model evaluates each response based on relevant writing (and speaking) subconstructs, which are "reflected in human scoring rubrics" (Cardwell et al., 2024b, p. 20) and operationalized through multiple linguistic features. DET publishes rubrics for the Photo Description task, the Interactive Writing/Writing Sample, and the Interactive Listening Summarization task that are based on CEFR descriptors, but these rubrics do not appear to be used in actual scoring. However, the manual reports a high correlation (.85) between the automated scores and human raters using the rubrics.

The subconstructs for DET writing include content, discourse coherence, lexis, and grammar. The manual provides a table that lists four to six dimensions for each subconstruct and one automated measure that addresses one of these dimensions. For example, the dimensions for content (for both writing and speaking) include task achievement, relevance, effect on the reader/listener, appropriacy of style, and development, and the example automated feature (similarity between the prompt and the response) addresses relevance only. It is therefore not clear what features measured by the scoring tool address other aspects of content. Test takers are not given additional information about how to achieve high scores, except to "Vary your sentence structure and word choice as much as possible" (Duolingo, 2024, p. 12), which suggests that these factors weigh heavily in the scoring algorithm.

Discussion

As with the other sections of the test, I now turn to a consideration of a validity argument for academic tests, weighing the writing in TOEFL iBT and DET against these propositions:

- Writing tasks will be relevant to academic writing tasks.
- The cognitive processes and strategies involved in writing will be similar to those needed for essential academic writing tasks.
- Scoring of the responses will provide evidence of student writing abilities that will be useful in making decisions (e.g., for admission and/or English language support).

Relevance to Academic Writing

Both tests provide opportunities to write open-ended responses. Recent additions to DET have enhanced the writing section so that there is a brief summary writing task in the interactive listening task (a summary of a conversation) and a two-part writing task, in which writers are prompted to add to what they have already written on a topic. However, the picture description task and brief writing sample have limited relevance to academic writing. In contrast, TOEFL iBT provides more opportunities to write in genres that are important in academic contexts, to write more extended prose, and to summarize and synthesize academic texts accurately.

Cognitive Processes

As noted previously, many of the same cognitive processes are involved in any writing activity, including microplanning and the translation of ideas into words and sentences. However, the DET writing tasks are so short that it is doubtful much macroplanning or revision is likely to take place. Furthermore, compared to DET writing, TOEFL iBT writing elicits more

cognitive processes that are relevant to academic writing, including selecting information from texts, integrating new information with the writer's own thoughts, and tailoring writing with a specific audience and purpose in mind.

Scoring Processes

In terms of scoring, there is a direct correspondence between the scoring rubric used by TOEFL iBT and the reported scores, whereas the relationship between the writing scores on DET and the published rubrics is more opaque. TOEFL iBT also provides model written responses in their published materials so test takers have more guidance on how responses should be structured.

To summarize, although both tests require test takers to produce original writing samples, the writing topics and tasks on TOEFL iBT are more relevant to academic writing than are those of DET, particularly as they require test takers to summarize and integrate academic content into their writing, which is an essential academic language skill.

Speaking

Defining the Speaking Construct

Speaking can be one of the most challenging skills to assess in a large-scale assessment, primarily because natural conversation ideally requires a live interlocutor, which may introduce unwanted variation because of the great variability in conversational styles of different examiners. Both TOEFL iBT and DET assess speaking on the computer rather than with a live examiner.

Speaking Construct for TOEFL iBT

Xi et al. (2021) define the speaking construct for TOEFL iBT as follows:

> The TOEFL iBT Speaking section measures test takers' abilities to communicate effectively in three subdomains of the English-speaking academic domain, including social interpersonal, academic navigational, and academic content. These include the abilities and capacities to use linguistic resources effectively to accomplish the following communication goals: a) to describe events and experiences and support or disagree with a personal preference or opinion about familiar topics in casual or routine social contexts drawing on personal experience; b) to select, relate, summarize, explain, compare, evaluate, and

synthesize key information from reading and listening materials on a typical campus life scenario on an academic topic typical of the college introductory course level. (p. 172)

Foundational and higher order abilities to accomplish these goals include the following:

- to pronounce words clearly and intelligibly;
- to use linguistic resources such as intonation, stress, and pauses to pace speech and to understand and express meaning precisely;
- to use linguistic resources such as vocabulary and grammar;
- to understand and express meaning precisely to use organizational devices (cohesive and discourse markers, exemplifications, etc.); and
- to connect and develop ideas effectively and to convey content accurately and completely.

Speaking Construct for DET

Duolingo does not provide a construct definition in its published materials as precise as TOEFL iBT's, although the 2024 manual includes this description of the speaking construct: "Producing spoken English from basic discourse to advanced discourse at CEFR levels A1–C2" (Cardwell et al., 2024b, p. 5). Park et al. (2023, p. 5), in a white paper on DET speaking, claim that DET speaking tasks map to oral production activities outlined in the CEFR, including three that involve sustained monologues (describing experience, giving information, putting a case) and two that are more relevant to dialogue (public announcements and addressing audiences).

Speaking Test Content

The TOEFL iBT Speaking section consists of one independent task and three integrated speaking tasks. Some of the integrated speaking tasks involve listening and speaking only, while others involve reading, listening, and speaking.

In the reading/listening/speaking integrated task, test takers are given a short time (less than 1 minute) to read a brief passage (less than 100 words) on an academic or campus-related topic and then listen to a brief excerpt from a lecture or a conversation on the same topic. Following the listening, the test taker must provide an accurate summary of what was heard and/or read. In the listening/speaking task, the test taker listens to a brief lecture and summarizes it orally.

All tests are graded both by human raters and by an automated scoring tool, using a 4-point rubric with four rating categories: general description, delivery, language use, and topic development. The rubrics for the two item types are slightly different, with the integrated rubric focusing more on accurate inclusion of information from the prompt sources. Both rubrics are published on the ETS website (https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf).

The DET contains several short speaking tasks. One of the tasks included in the production and conversations scores is a sentence-reading task, which was referred to in previous technical manuals (prior to October 2024) as a "read-aloud version of the elicited imitation task" (Cardwell et al., 2024a, p. 15), a somewhat misleading statement, although this usage no longer appears in the latest version of the manual.

There are also several short speaking tasks that require open-ended responses. According to the DET manual (Cardwell et al., 2024b, p. 15) the tasks include the following: "prompt-based speaking tasks (Extended Speaking [audio prompt], Extended Speaking [text prompt], and Speaking Sample, which is shared with institutions) and an image-based speaking task, Picture Description (speaking)." The Speaking Sample is not further described in the manual but appears to be a second instance of a text-based extended speaking prompt. For the prompt-based speaking tasks, test takers are asked to speak for 90 seconds after up to 30 seconds of preparation time. The instructions for the picture prompt are to "speak about the image" for 90 seconds, while the text and audio prompts ask test takers to "recount an experience, give examples and recommendations, or argue a point of view" and are selected to represent the four CEFR domains of social life (personal, public, educational, and professional). For example, one audio prompt included in the DET test-taker guide (Duolingo, 2024, p. 69, #5) asks listeners to describe a place that they like, where it is, how they get there, and what they see there.

Similar to writing, DET spoken tasks are scored automatically. The same subconstructs are assessed in speaking as described in the writing section, with the addition of fluency and pronunciation. As with the writing, there are published scoring rubrics that are based on CEFR descriptors, but the relationship between the rubric descriptors and the actual scoring algorithm is not made clear.

Speaking on the two tests is compared in Table 9.

Table 9. Comparison of TOEFL iBT and DET Speaking

Characteristic	TOEFL iBT Speaking	DET Speaking
Task types	Independent/personal opinion (1) Integrated/based on listening and/or reading (3)	Read a sentence aloud Picture description Open ended response (reading prompt) Open-ended response (listening prompt)
Number of tasks	4	4–5
Speaking time	45–60 second depending on task	30–90 seconds depending on task
Total duration	About 16 minutes	About 10 minutes
Preparation time	15 seconds (independent) 30 seconds (integrated)	30–90 seconds depending on task
Response format	Structured response expected (clear introduction/support/ conclusion	No set structure
Focus	Academic focus	Topics come from personal, professional, educational and public sources
Scoring	Human and automated scores; rubric	Automated scoring with human verification

Cognitive Processes

Field (2011) provided a model of speaking that lists six stages, from conceptualization of an idea that the speaker intends to express through grammatical and morpho-phonetical encoding, which converts the ideas into words and phrases. Phonetic encoding transforms the string into neural instructions for the speech articulator (e.g., the lips, tongue, and vocal folds), which then produce the utterance itself. Finally, self-monitoring can occur when a speaker evaluates the utterance and provides self-repair when relevant. Based on the description of the tasks, it can be argued that both TOEFL iBT and DET allow test takers to demonstrate all six stages in Field's model (see Table 10), although the need for self-monitoring may be reduced due to the lack of a live interlocutor on both tests. However, the TOEFL iBT's integrated speaking tasks are more cognitively demanding, as they require test takers to recall and/or select information from aural or written sources and use this information in their responses. This additional demand on cognitive resources is indicated by a double X (XX) in the Conceptualization row in the table.

Table 10. Comparison of Cognitive Processes in TOEFL iBT and DET

Characteristic	TOEFL iBT	DET
Conceptualization	XX	X
Grammatical encoding	X	X
Morpho-phonological encoding	X	X
Phonetic encoding	X	X
Articulation	X	X
Self-monitoring	(X)	(X)

Note. XX = additional demand on cognitive sources; (X) = may or may not be required.

Discussion

As with the other skills, the propositions to be considered are the following:

- Speaking tasks will be relevant to academic speaking tasks.
- The cognitive processes and strategies involved in speaking will be similar to those needed for essential academic speaking tasks.
- Scoring of the responses will provide evidence of student speaking abilities that will be useful in making decisions (e.g., for admission and/or English language support).

Relevance to academic speaking tasks: Of all four skills, speaking is the skill that is arguably most similar on TOEFL iBT and DET, as both tests require monologic responses to computer inputs with similar amounts of time. Neither test assesses speaking in interaction, which is a limitation of both tests. However, the topics from TOEFL iBT are all drawn from the academic domain and the speaking tasks themselves are academic in nature. In particular, the integrated tasks require test takers to talk about academic content they have read or listened to, which reflects the environment in which prospective students will find themselves in. Overall, the content of the TOEFL iBT is more academic than that of DET.

Cognitive Processes and Strategies

As noted above, the cognitive processes and strategies are similar in both tests, in that test takers are given similar preparation times and speak for similar lengths of time. However, the requirement to summarize academic content in speaking is found in TOEFL iBT only, and this requirement adds to the cognitive complexity of the speaking tasks, in addition to being more authentic.

Scoring

As with the writing section, the scoring rubric used by TOEFL iBT is published and the reported scores are directly related to this rubric, while it is not as clear what automated indices contribute to the DET scores or how they are related to the DET rubric

To summarize, among the four skills, speaking is perhaps the most similar in terms of format, number, and length of tasks across TOEFL iBT and DET. However, TOEFL iBT more directly addresses specific aspects of speaking that are important in academic contexts, both in terms of topics and the demands of the task.

Summary of Findings

To summarize, the analysis presented in this report is framed in terms of three propositions from the TOEFL validity argument. I will discuss each proposition in turn.

The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings.

The analysis of the test content reveals that TOEFL iBT, grounded as it is in a thorough domain analysis and explicitly intended as a test of language in academic settings, contains more academic tasks in the areas of reading, listening, speaking and writing than does DET, which does not claim to be a test of academic English and draws content from four domains listed in the CEFR. Although recent revisions to DET have addressed some of the shortcomings of earlier versions of the test in terms of academic content, much of the test content comes from other domains such as personal or public. There is almost no listening on DET that can be considered academic, even though the interactive listening task includes academic-navigational content. The reading passages on the DET are quite short and not necessarily academic in nature.

Tasks and scoring criteria are appropriate for obtaining evidence of test takers' academic language abilities.

In the areas of listening and reading, TOEFL iBT includes listening tasks that require test takers to process extended texts in order to extract main ideas and details, make inferences, and discern the speaker or writer's intentions. The integrated speaking and writing tasks, furthermore, require test takers to convey information from spoken and written texts accurately in speaking and writing.

Many of the DET tasks focus on lower level language skills, especially vocabulary, and the types of reading and listening questions do not appear to assess the ability to read or listen to

extended texts, distinguish main ideas from details, or make the kinds of inferences that are needed for successful apprehension of academic texts. The speaking and writing tasks, especially picture description tasks, do not seem designed to elicit academic language forms or functions, and the writing tasks in particular are of such limited length that they cannot elicit evidence of a test taker's ability to produce extended discourse with a unified thesis and adequate support.

The addition of reported individual skill scores for DET in 2024, while presumably an attractive feature for test users, calls into question the issues raised earlier in this report about the integration of skills. According to the DET manual (Cardwell et al., 2024b), the overall score is an average of the four individual skill scores, yet it is questionable whether each of the individual skill scores truly represents performance in that skill. This issue is particularly problematic for listening; as noted above, many if not most of the interactive listening task items can be guessed by a test taker who is sufficiently proficient in reading.

Academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks.

In reading and listening, examination of a number of DET items suggests that many of them can be answered without engaging in the text in any depth. In the interactive listening tasks, as discussed above, many items can be answered simply by reading the scenario and eliminating implausible distractors without understanding much if any of the spoken texts, and the summary task at the end can also be accomplished by reading through the conversation transcript. In the interactive reading task, many of the automatically generated distractors can be easily discounted without reading the passage. It would therefore be difficult to infer academic language proficiency based on these items.

The multiple-choice items on the TOEFL iBT listening and reading sections, on the other hand, follow best practices in terms of designing distractors, so that apart from guessing, which is always a possibility, the correct answer can only be obtained if one has truly understood the text. Thus, the evidence for this proposition is stronger for TOEFL iBT in these two sections than for DET.

By the same token, the tasks and topics for writing and speaking on TOEFL iBT are designed specifically to elicit evidence of academic language ability, particularly in terms of selecting relevant information from input texts and integrating it with the test taker's own ideas in a spoken or written response. None of the DET speaking or writing tasks are integrated in this

same sense; while they may be based on either an aural or a written prompt, the prompts elicit the test takers own thoughts or ideas and do not require the sort of integration of content that a truly academic test task does.

Conclusion

Admissions decisions should never be based on test scores alone. Test users should also consider additional factors, including the reading, writing, listening, and speaking demands of the academic program, any other available evidence of English language ability (e.g., interviews, writing samples), and the amount of English language support available to those students who arrive on campus with lower-than-expected English language skills.

At the same time, returning to the ILTA (n.d.) guidelines referenced at the beginning of this report, test users need to ensure that the tests they accept are valid and reliable and based on a construct relevant to the decision being made. As I have documented here, based on a consideration of the content and other published materials from both TOEFL iBT and DET, the construct of academic language ability is more clearly operationalized and assessed in TOEFL iBT than in the current iteration of DET.

References

Aryadoust, V., & Luo, L. (2023). The typology of second language listening constructs: A systematic review. *Language Testing*, *40*(2), 375–409. https://doi.org/10.1177/02655322221126604

Cardwell, R., LaFlair, G. T., & Settles, B. (2022). *Duolingo English Test: Technical manual* (Duolingo Research Report). Duolingo. Retrieved May 12, 2022, from https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf

Cardwell, R., Naismith, B., LaFlair, G. T., & Nydick, S. (2024a, May). *Duolingo English Test: Technical manual* (Duolingo Research Report). Duolingo. Retrieved August 1, 2024, from https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf

Cardwell, R., Naismith, B., LaFlair, G. T., & Nydick, S. (2024b, October). *Duolingo English Test: Technical manual* (Duolingo Research Report). Duolingo. Retrieved December 1, 2024, from https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf

Chapelle, C. (1999). From reading theory to testing practice. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 150–166). Cambridge University Press.

Chapelle, C. A. (2008). The TOEFL® validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). Routledge. https://doi.org/10.4324/9780203937891

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

Cumming, A., Lai, C., & Cho, H. (2016). Students' writing from sources for academic purposes: A synthesis of recent research. *Journal of English for Academic Purposes*, *23*, 47–58. https://doi.org/10.1016/j.jeap.2016.06.002

Davis, L., & Norris, J. M. (2023). *A comparison of two TOEFL® writing tasks* (Research Memorandum No. RM-23-06). ETS. https://www.ets.org/Media/Research/pdf/RM-23-06.pdf

Duolingo, (2024). Duolingo English Test: *Official guide for test takers.* https://englishtest.duolingo.com/prepare/guide

Enright, M. K., Bridgeman, B., Eignor, D., Lee, Y.-W., & Powers, D. E. (2008). Prototyping measures of listening, reading, speaking, and writing. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 145–225). Routledge. https://doi.org/10.4324/9780203937891

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (RM-00-04). ETS. https://www.ets.org/Media/Research/pdf/RM-00-04-Enright.pdf

ETS. (n.d.[a]). *Sample the TOEFL iBT test.* TOEFL. https://www.ets.org/toefl/test-takers/ibt/prepare/sample-test.html

ETS. (n.d.[b]). *TOEFL iBT listening section.* TOEFL. https://www.ets.org/toefl/test-takers/ibt/about/content/listening.html

ETS. (n.d.[c]). *TOEFL IBT reading section.* TOEFL. https://www.ets.org/toefl/test-takers/ibt/about/content/reading.html

ETS. (n.d.[d]). *TOEFL iBT test content.* TOEFL. https://www.ets.org/toefl/test-takers/ibt/about/content.html

ETS. (2020). *TOEFL® research insight series: Vol. 4. Validity evidence supporting the interpretation and use of TOEFL iBT® scores.* https://www.ets.org:/pdfs/toefl/toefl-ibt-insight-s1v4.pdf

ETS (2024). *TOEFL® research insight series: Vol. 1. TOEFL iBT test framework and test development.* https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v1.pdf

Field, J. (2011). Cognitive validity in speaking tests. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 65–111). Cambridge University Press

Field, J. (2013). Cognitive validity. In E. Taylor & C. Weir (Eds.), *Examining listening* (pp. 267–290). Cambridge University Press.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*(6), 1082–1111. https://doi.org/10.3102/0034654317726529

Goodwin, S., & Naismith, B. (2023). *Assessing listening on the Duolingo English Test* (Research Report DRR-23-02). Duolingo. https://doi.org/10.46999/CORJ9896

Goodwin, S., Poe, M., Cardwell, R., Runge, A., Attali, Y., Mulcaire, P., Lo, K.-L., & LaFlair, G. T. (2024). *Facilitating the writing process on the DET: The interactive writing task* (Research Report DRR-24-02). Duolingo.

Graesser, A. C., & Forsyth. (2013). Discourse comprehension. In Daniel Reisberg (Ed.), *The Oxford handbook of cognitive psychology*. Oxford Library of Psychology. Retrieved January 23, 2025, from https://doi.org/10.1093/oxfordhb/9780195376746.013.0030.

Hermida, J. (2009). *The importance of teaching academic reading skills in first-year university courses.* SSRN. https://doi.org/10.2139/ssrn.1419247

Huff, K., Powers, D. E., Kantor, R. N., Mollaun, P., Nissan, S., & Schedl, M. (2008). Prototyping a new test. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 187–186). Routledge. https://doi.org/10.4324/9780203937891

ILTA. (n.d.). *ILTA Guidelines for practice in English.* https://www.iltaonline.com/general/custom.asp?page=ILTAGuidelinesforPractice

Janusik, L. A., & Wolvin, A. D. (2009). 24 hours in a day: A listening update to the time studies. *International Journal of Listening*, *23*(2), 104–120. https://doi.org/10.1080/10904010903014442

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kintsch, W. (2013). Revisiting the construction-integration model of text comprehension and its implications for instruction. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 807–839). International Reading Association

Kostromitina, M. (2024, September 12). Is the Duolingo English Test valid? *Test Center.* https://blog.englishtest.duolingo.com/is-the-duolingo-english-test-valid/

Lam, D. M. K. (2021). Don't turn a deaf ear: A case for assessing interactive listening. *Applied Linguistics*, *42*(4), 740–764. https://doi.org/10.1093/applin/amaa064

Liu, X., & Read, J. (2023). Designing a new diagnostic reading assessment for a local post-admission assessment program: A needs-driven approach. In X. Yan, S. Dimova, and A. Ginther (Eds.), *Local language testing: Practice across contexts* (pp. 83–102). Springer Cham. https://doi.org/10.1007/978-3-031-33541-9_5

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*(2), 79–88. https://doi.org/10.1016/j.jeap.2011.03.001

Nydick, S. W., & Lockwood, J. R. (2024). An overview of Duolingo English Test administration and scoring. *Duolingo Research Report*, DRR-24-03. Duolingo. Retrieved from https://duolingo-papers.s3.amazonaws.com/reports/Duolingo_whitepaper_test_scoring_2024_v1.pdf

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

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/pdfs/toefl/RM-21-03.pdf

Papageorgiou, S., Schmidgall, J., Harding, L., Nissan, S., & French, R. (2021). Assessing academic listening. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 61–106). Routledge. https://doi.org/10.4324/9781351142403-3

Park, Y., Cardwell, R., Goodwin, S., Naismith, B., LaFlair, G. T., Lo, K.-L., & Yancey, K. (2023). *Assessing speaking on the Duolingo English Test* (Research Report No. DRR-23-03). Duolingo. https://doi.org/10.46999/DJIY3654

Park, Y., LaFlair, G. T., Attali, Y., Runge, A., & Goodwin, S. (2022). *Interactive Reading—The Duolingo English Test (*Research Report DRR-22-02). Duolingo.

Pearlman, M. (2008). Finalizing the test blueprint. In In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 227–258). Routledge. https://doi.org/10.4324/9780203937891

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26*(4), 561–587. https://doi.org/10.1177/0265532209340192

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441–474. https://doi.org/10.1191/0265532206lt337oa

Schedl, M., O'Reilly, T., Grabe, W., & Schoonen, R. (2021). Assessing academic reading. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 22-60). Routledge. https://doi.org/10.4324/9781351142403-2

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

Taylor, C. A., & Angelis, P. (2008). The evolution of the TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 28–41). Routledge.

Wagner, E. (2016). *A study of the use of the TOEFL iBT® test speaking and listening scores for international teaching assistant screening* (TOEFL iBT Research Report No. 27). ETS. https://doi.org/10.1002/ets2.12104

Wagner, E. (2020). Duolingo English test, revised version July 2019. *Language Assessment Quarterly*, *17*(3), 300–315. https://doi.org/10.1080/15434303.2020.1771343

Wagner, E. (2022). L2 listening comprehension: Theory and research. In E. H. Jeon & Y. In'nami (Eds.), *Bilingual Processing and Acquisition* (pp. 213–233). John Benjamins.

Wagner, E., & Kunnan, A. J. (2015). The Duolingo English test. *Language Assessment Quarterly*, *12*(3), 320–331. https://doi.org/10.1080/15434303.2015.1061530

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*, 27–55. https://doi.org/10.1016/j.asw.2004.01.002

Xi, X., Norris, J. M., Ockey, G. J., Fulcher, G., & Purpura, J. E. (2021). Assessing academic speaking. In X. Xi and J. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 152–199). Routledge. https://doi.org/10.4324/9781351142403-5

Suggested Citation:

Cushing, S. T. (2025). *Testing academic language proficiency: Comparing the TOEFL iBT® test with the Duolingo English Test* (TOEFL Research Report No. RR-104). ETS.

Action Editor: Larry Davis

Reviewers: Spiros Papageorgiou and Jonathan Schmidgall

Find other ETS-published reports by searching the ETS ReSEARCHER database.