

Duolingo English Test: Technical Manual



Duolingo Research Report
December 15, 2023 (42 pages)
<https://englishtest.duolingo.com/research>

Ramsey Cardwell*, **Ben Naismith***, **Geoffrey T. LaFlair***, and **Steven Nydick***

Abstract

The Duolingo English Test Technical Manual provides an overview of the design, development, administration, and scoring of the Duolingo English Test. Furthermore, the Technical Manual reports validity, reliability, and fairness evidence, as well as test-taker demographics and the statistical characteristics of the test. This is a living document whose purpose is to provide up-to-date information about the Duolingo English Test, and it is updated on a regular basis.

Last Update: December 15, 2023

Contents

1	Introduction	3
2	Theoretical Basis	3
3	Test Constructs and Corresponding Task Types	4
3.1	Test Constructs	4
3.2	Test Task Types	5
4	Development, Delivery, and Scoring	15
4.1	Task and Item Development	15
4.2	Fairness and Bias Review	17
4.3	CAT Delivery	17
4.4	CAT Item Scoring	18
4.5	Extended Speaking and Writing Task Scoring	19
4.6	Subscores	20
5	Access and Accommodations	21
5.1	Access	21
5.2	Accommodations	23
6	Test Administration and Security	23
6.1	Test Administration	23
6.2	Onboarding	24
6.3	Administration Rules	24
6.4	Proctoring and Reporting	24
7	Test-Taker Demographics	25

Note: We would like to acknowledge the contributions of Burr Settles, the creator of the Duolingo English Test and author of the first Technical Manual.
*Duolingo, Inc.

Corresponding author:

Geoffrey T. LaFlair, PhD
Duolingo, Inc. 5900 Penn Ave, Pittsburgh, PA 15206, USA
Email: englishtest-research@duolingo.com

8	Test Performance Statistics	28
8.1	Score Distributions	28
8.2	Reliability Evidence	28
8.3	Relationship with Other Tests	30
9	Quality Control	32
9.1	Analytics for Quality Assurance in Assessment	32
9.2	Proctoring Quality Assurance	33
10	Conclusion	34
11	Appendix	35
	References	40

1 Introduction

The Duolingo English Test (DET) is a measure of English language proficiency for communication and use in English-medium settings. It assesses test-taker ability to use language skills that are required for literacy, conversation, comprehension, and production. The test is designed for maximum accessibility; it is delivered via the internet, without a testing center, and is available 24 hours a day, 365 days a year. In addition, as a computer-adaptive test (CAT), it is designed to be efficient; the test takes approximately one hour to complete, though as a CAT the exact time varies for each test taker. The test uses task types that provide maximal information about English language proficiency while being feasible to develop, administer, and score at scale. In all areas of the test, high standards of security and psychometric quality are maintained.

In adhering to the *Standards for Educational and Psychological Testing* (AERA et al., 2014) standards for test documentation (Chapter 7), this technical manual provides an overview of all aspects of the DET so that stakeholders can make informed decisions about how to interpret and use DET test scores. It contains a presentation of:

- the test’s tasks, the constructs they cover, and how they are created using human-in-the-loop generative AI;
- the adaptive delivery and scoring of test items using computational psychometrics;
- the test’s accessibility, delivery, and proctoring and security processes;
- demographic information of the test-taker population;
- and the statistical characteristics of the test.

Since its inception in 2016, the social mission of the DET has been to lower barriers to education access for English language learners around the world. The DET achieves this goal 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. These scores are intended to be interpreted as reflecting test-taker English language proficiency and to be used in a variety of settings, including for post-secondary admissions decisions. To date, the success of this mission is evidenced by the widespread adoption of the DET by more than 4,800 academic programs in 100 countries.

2 Theoretical Basis

The Duolingo English Test employs a novel assessment ecosystem (Burstein et al., 2022) composed of an integrated set of frameworks related to language assessment, design, psychometrics, test security, and test-taker experience (TTX). Furthermore, the processes and documentation of the DET—including test development, scoring, and documentation of validity, reliability and fairness evidence—have been **externally evaluated** against the AERA et al. (2014) standards and **internally evaluated** against the Responsible Artificial Intelligence (RAI) Standards (Burstein, 2023). These theoretical underpinnings motivate the research philosophy and values of the DET which aim to make the DET test-taker-centered by taking advantage of the latest developments in technology (including machine learning and artificial intelligence), applied linguistics, psychometrics, and assessment science.

On a more fundamental level, the Duolingo English Test subscribes to the interactionist definition of what a test can in fact assess (Chapelle, 1998; Messick, 1989, 1996; Young, 2011). In this conceptualization, test-taker performance reflects two elements and their interaction: 1) the underlying traits of the test taker (English proficiency), and 2) the context-specific behaviors of the test taker (task performance). For example, an individual may evidence a certain level of spoken proficiency during a face-to-face conversation but may struggle with the exact same conversation on the phone. It is therefore necessary to always consider the characteristics of the setting (including the task type and language modality) when drawing conclusions about a test-taker’s underlying traits. This theory of language aligns closely with the tenets of the communicative language ability (CLA) model which calls for language assessments to be informed by “language ability in its totality” (Bachman & Palmer, 1996, p. 67).

The result of the above considerations is a modern test that equally meets the assessment criteria and the needs of stakeholders, and which is continually being evaluated and iterated upon in all aspects of our assessment processes. Together, these ecosystem frameworks, testing standards, and research philosophy support a test validity argument built on a digitally informed chain of inferences, appropriate for a digital-first assessment of this nature and consistent with professional standards of practice. As a result, the adaptive DET can be seen to assess test takers’ proficiency in General English and English for Academic Purposes (EAP), both of which are essential for success in a range of academic or professional settings. The DET considers EAP to consist of the language knowledge and skills necessary to perform common communicative and pedagogical tasks in a range of educational contexts across academic disciplines. This definition highlights the importance of communicative competence in educational settings.

3 Test Constructs and Corresponding Task Types

In this section, we first describe the constructs being tested, that is, “the specific definition of an ability that provides the basis for a given assessment or assessment task and for interpreting scores derived from this task” (Bachman & Palmer, 2010, p. 43). We then describe the specific task types on the DET and how these relate to the intended constructs.

3.1 Test Constructs

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. These integrated skills areas correspond to the DET subscores, and each subscore can be interpreted as a combination of two of the more traditional language subskills of speaking, writing, reading, and listening (SWRL). (For white papers on how the DET assesses each SWRL skill, see Park et al. (2023) for speaking; Goodwin et al. (2023) for writing; Park et al. (2022) for reading, and Goodwin & Naismith (2023) for listening.) Figure 1 shows the relationship between traditional language subskills and DET subscores (see Section 4.6).

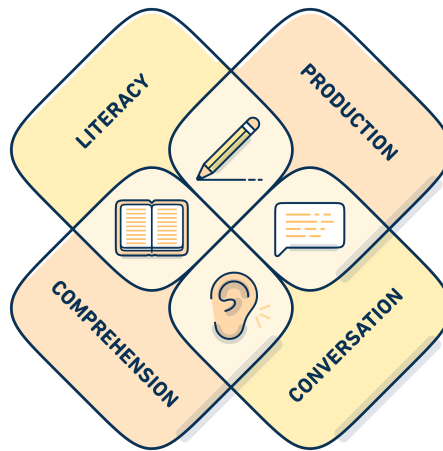


Figure 1. Relationship between SWRL language skills and DET subscores

The overall score and subscores reported by the DET are aligned with the Common European Framework of Reference (Council of Europe, 2001, 2020), commonly known as the CEFR, an international standard for describing language ability. The CEFR consists of a six-point ordinal scale: A1, A2 (Basic); B1, B2 (Independent); C1, and C2 (Proficient User). Because the CEFR is a competency-based framework, each proficiency level is operationalized as lists of tasks that a language user at that level is likely able to do, for example “Can read straightforward factual texts on subjects related to their field of interest with a satisfactory level of comprehension” (Overall reading comprehension, B1). Importantly, the CEFR provides a common framework for the basis of language syllabi, curricula, materials, and assessments around the world (Byram & Parmenter, 2012). The alignment of DET scores to CEFR levels is available on the scores page and is an important consideration throughout the test development process (e.g., Settles et al., 2020).

In total, the DET has thirteen different graded task types that collectively measure test-taker proficiency in the English-language constructs described above. These task types include both closed-ended task types (e.g., C-test and Yes/No Vocabulary) and open-ended task types (e.g., Picture Description and Writing Sample). The creation and selection of this specific combination of task types is guided by the DET Ecosystem (Burstein et al., 2022), especially the Language Assessment Design Framework. In this framework, task design and scoring target constructs relevant for General and Academic English language proficiency. In addition to test use validity, another consideration in test design is ensuring a delightful TTX. As a result of these considerations, DET tasks are intuitive, reducing the need for test-specific preparation (Carr, 2023). All DET task types are summarized in Tables 1 and 2 and are described individually in the subsequent sections.*

*See Section 4.6 for information on subscores.

Table 1. Constructs and Task Types

Subscore	Skills	Description	Task Types
Literacy	Reading Writing	Reading and writing English from basic informational text to advanced expository/persuasive texts at CEFR levels A1–C2	<ul style="list-style-type: none"> • C-test • Extended Writing • Interactive Listening (S) • Interactive Reading • Picture Description (writing) • Writing Sample • Yes/No Vocabulary
Conversation	Listening Speaking	Listening and producing spoken English from basic discourse (e.g., informational) to advanced discourse (e.g., lectures) at CEFR levels A1–C2	<ul style="list-style-type: none"> • Dictation • Elicited Imitation • Extended Speaking (audio prompt) • Extended Speaking (text prompt) • Interactive Listening (DC) • Picture Description (speaking) • Speaking sample
Comprehension	Reading Listening	Understanding spoken and written English from basic informational discourse (e.g., mini-talks) to advanced discourse (e.g., extended monologues) at CEFR levels A1–C2	<ul style="list-style-type: none"> • C-test • Dictation • Elicited Imitation • Interactive Listening (DC) • Interactive Reading • Yes/No Vocabulary
Production	Speaking Writing	Producing spoken and written English from basic informational discourse (e.g., paragraphs) to advanced discourse (e.g., persuasive arguments) at CEFR levels A1–C2	<ul style="list-style-type: none"> • Elicited Imitation • Extended Speaking (audio prompt) • Extended Speaking (text prompt) • Extended Writing • Interactive Listening (S) • Picture Description (speaking) • Picture Description (writing) • Speaking sample • Writing Sample

Note: S = summarization, DC = dialogue completion.

3.2 Test Task Types

We now describe the 13 task types (and their sub-tasks) outlined in Table 2. These tasks vary in their authenticity, that is, “the degree of correspondence of characteristics of a given language test task to the features of a TLU [target language use] task” (Bachman & Palmer, 2010, p. 23). For example, highlighting relevant text in a reading passage as part of the Interactive Reading task is highly authentic as it is a skill that many test takers will employ in academic (or other) contexts. In contrast, the Yes/No Vocabulary task which targets vocabulary knowledge is less authentic, as deciding if a word is real or not is not an activity that test takers are likely to encounter outside of a language testing context. The DET deliberately includes tasks with different levels of authenticity in order to maximize the efficiency of the task for measurement accuracy while also fully covering the intended constructs.

Table 2. Task types and Administration Order

Task Type	Name for Test Takers	Adaptive	Freq
Focus area 1: Linguistic resources			
C-test	Read and Complete	Yes	4–6
Yes/No Vocabulary	Read and Select	Yes	4–6
Dictation	Listen and Write	Yes	4–6
Elicited Imitation	Read Aloud	Yes	4–6
Focus area 2: Skills mastery			
Interactive Reading	Interactive Reading	Yes	2
Interactive Listening	Interactive Listening	Yes	2
Picture Description (writing)	Write About the Photo	No	3
Extended Writing	Read, Then Write	No	1
Picture Description (speaking)	Speak About the Photo	No	1
Extended Speaking (text prompt)	Read, Then Speak	No	1
Extended Speaking (audio prompt)	Listen, Then Speak	No	2
Writing Sample	Writing Sample	No	1
Speaking sample	Speaking Sample	No	1

C-test

The C-test task type (see Figure 2) measures a test-taker’s global language proficiency in the written modality (Norris, 2018), capturing chiefly knowledge of vocabulary and grammar (Eckes & Grotjahn, 2006). In addition, C-test scores correlate moderately well with discrete language components including reading ability (Khodadady, 2014; Klein-Braley, 1997), spelling skills (Khodadady, 2014), and vocabulary (Karimi, 2011). It has been shown that scores from C-tests are significantly correlated with scores from many other major language proficiency tests (Daller et al., 2021; Khodadady, 2014).

In this task, the test taker is presented with a short text. The first and last sentences of the text are fully intact, while alternating words in the intervening sentences are “damaged” by deleting the second half of the word. Test takers respond to the C-test items by completing the damaged words in the paragraph. Test takers need to rely on context and discourse information to reconstruct the damaged words (which span multiple lexical and morphosyntactic categories).

The C-test passages themselves reflect a range of different text types including fiction (e.g., colloquial narratives), news articles, and textbook passages. The linguistic features of these passages have been carefully analyzed to ensure a variety of text types and difficulty levels. In total, more than 150 linguistic features are annotated and accounted for, including features related to parts of speech, verb types, and passage length (see McCarthy et al., 2021 for the complete list).

Yes/No Vocabulary

The “yes/no” vocabulary test (see Figure 3) measures breadth of receptive vocabulary knowledge (Beckmans et al., 2001). Such knowledge is a critical element of learning a second/additional language (L2) and impacts all language skills, including reading (Laufer, 1992; Roche & Harrington, 2014), listening (Bonk, 2000; Staehr, 2008), speaking (Milton et al., 2010; Milton, 2013), and writing (Kyle & Crossley, 2016; Ruegg et al., 2011). More specifically, the “yes/no” vocabulary test has been shown to predict listening, reading, and writing abilities (McLean et al., 2020; Milton et al., 2010; Staehr, 2008), and it has been used to assess vocabulary knowledge at various CEFR levels (Milton, 2010).

In this task type, test takers are presented with a set of written English words mixed with pseudo-words designed to appear English-like. We use an LSTM recurrent neural network trained on the English dictionary to create realistic pseudo-words, filtering out any real words, including any acceptable spellings, and pseudo-words that orthographically or phonetically resemble real English words too closely. Test takers respond by selecting the real English words. The proportion of real words varies across tasks, making it harder to guess correctly. Traditional yes/no vocabulary tests simultaneously present a large set of mixed-difficulty stimuli (e.g., 60 words and 40 pseudo-words). On the DET, a set of yes/no vocabulary items is presented adaptively, with multiple, smaller sets each containing a few stimuli of the same difficulty administered based on how the test taker performed on previous items (see Section 4.3 for more on the computer-adaptive administration).

2:55

Type the missing letters to complete the text below.

International Venue of the Year

The Phones 4u Arena seats over 21,000 and is the largest indoor arena Europe. It been
 International Venue of Year for years the
 popular in world. The sports grounds also host large pop concerts.

NEXT

Figure 2. Example C-test Task

0:52

Select the real English words in this list.

regret	banic	signature	unfalled	deal	corner
controde	fantally	prinanter	insurd	poid	unforgettable
strementy	friever	sharket	darious	eerses	parken

NEXT

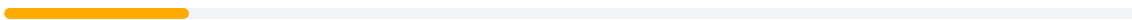
Figure 3. Example Yes/No Vocabulary Task

Dictation

Dictations are an integrated skills task (listening and writing) that assess test-taker ability to recognize individual words and to hold them in memory long enough to accurately reproduce them; both abilities are critical for spoken language understanding (Bradlow & Bent, 2002; Buck, 2001; Smith & Kosslyn, 2007). Dictation tasks have also been found to be associated with language-learner intelligibility in speech production (Bradlow & Bent, 2008).

For the DET Dictation task, test takers listen to a spoken sentence or short passage and then transcribe it using the computer keyboard (see Figure 4). Test takers have one minute to listen to the stimulus and transcribe what they heard. They can play the passage up to three times.

0:55



Type the statement that you hear.



Number of replays left: 2

NEXT

Figure 4. Example Dictation Task

Elicited Imitation

The read-aloud variation of the elicited imitation task (see Figure 5) is an integrated skills task measuring test-taker reading and speaking abilities (Jessop et al., 2007; Litman et al., 2018; Vinther, 2002). The goal of this task is to evaluate the intelligibility and fluency of speech production, which are affected by segmental/phonemic and suprasegmental properties like intonation, rhythm, and stress (Anderson-Hsieh et al., 1992; Derwing et al., 1998; Field, 2005; Hahn, 2004). Furthermore, intelligibility is correlated with overall spoken comprehensibility (Derwing et al., 1998; Derwing & Munro, 1997; Munro & Derwing, 1995), meaning that this task type can capture aspects of speaking proficiency.

This task type requires test takers to read, understand, and speak a sentence. After reading the target sentences, test takers respond by using the computer's microphone to record themselves reading the sentence exactly as they heard it. The DET uses state-of-the-art speech recognition technologies to extract features of spoken language, such as acoustic and fluency features that predict these properties (in addition to automatic speech recognition), thus evaluating the general intelligibility of speech.

0:14

Record yourself saying the statement below.



"The students have already been assigned their seats."

RECORD NOW

Figure 5. Example Elicited Imitation Task

Interactive Reading

The Interactive Reading task type complements the other test task types that assess reading with a focus on reading processes (C-test and Elicited Imitation) by focusing on reading comprehension (Park et al., 2022). This task type requires test takers to engage with a text by sequentially performing a series of sub-tasks tapping different subconstructs of reading (reading to find information, reading for comprehension, and reading to learn) and all using the same text as the stimulus. This task type is interactive in the sense that as the test progresses, different portions of the text are presented to the test taker, ensuring a comprehensive evaluation of their reading skills. By incorporating interactivity, the DET interactive writing tasks provide a dynamic and immersive testing experience that better reflects real-world language usage.

The first sub-task shows the test taker the first half of the text with 5–10 words missing (see Figure 6); test takers must select the word that best fits each blank. Next, test takers are shown the remainder of the text with one sentence missing (see Figure 7); test takers must select the sentence that best completes the passage from among several options. With the text now complete, test takers are shown sequentially two questions and asked to highlight the part of the text that contains the answer (see Figure 8). Test takers are then asked to select an idea that appears in the passage from among several options, only one of which is correct (see Figure 9). Finally, test takers are asked to choose the best title for the text from among several options (see Figure 10).

Each Interactive Reading passage is classified by genre as either narrative or expository; each test taker receives one narrative passage and one expository passage. Additionally, the number of complete-the-sentence blanks across the two tasks is controlled such that each test taker receives approximately the same number. Test takers receive an additional minute to complete the longer Interactive Reading task.

Interactive Listening

The Interactive Listening task type contributes to measurement of the constructs of L2 listening, reading, and writing (LaFlair et al., 2023). It complements the Dictation task type, which focuses more on listening processes, by also measuring aspects of interactional competence. This listening task type is interactive in that it requires test takers to participate in a situationally driven conversation in a university setting. The Interactive Listening task demonstrates correspondence to the TLU domain of English-medium postsecondary

6:53 for the next 6 questions

PASSAGE

Biophysics is the study of the physical properties of living things. This field refers to physics, which **1** the science of matter and energy, and also to biology, the science of living **2**.

Biophysicists study the physical **3** of organisms and the **4** of physical processes on **5** things. For example, biophysicists might study the effect certain chemicals **6** on living cells, determine how tiny structures within cells work, or explain how injuries and diseases **7** the structure of skin. Some biophysicists also **8** the interaction of radiation with **9** systems.

Select the best option for each missing word.

1 Select a word

2 Select a word

3 Select a word

4 Select a word

5 Select a word

6 Select a word

7 Select a word

8 Select a word

9 Select a word

NEXT

Figure 6. Example Interactive Reading “Complete the Sentences” Sub-task

studies through the conversation topics, interlocutors (students and professors), and communicative functions, which include asking for clarification about lecture content, making requests, gathering information, asking for advice, planning study sessions, and participating in other university-oriented conversations (Biber & Conrad, 2019).

An Interactive Listening task starts with a scenario that describes who the test taker is talking with and for what purpose. This scenario is followed by the dialogue completion sub-task. Some dialogue completions require the test taker to select the first turn in the conversation, while others start with the interlocutor. After each interlocutor turn (which is presented in audio format only), the test taker must select the best response (among multiple options presented in writing) to continue the conversation (see Figure 11). The test taker receives visual feedback after each response; if the response is correct, the box around the text turns green; otherwise, the box turns red, and the correct response is shown. In this way, test takers can respond to the remaining turns based on the intended input. Once the conversation ends, the test taker may use any remaining time to review the conversation before proceeding to the summary sub-task (see Figure 12). In the summarization sub-task, the test taker has 75 seconds to compose a written summary of the conversation.

Each Interactive Listening task exhibits one of three types of conversations: student–student conversations that focus on requests, advice seeking, and other university-oriented purposes; student–professor conversations that focus on similar purposes; and student–professor conversations that focus on information seeking where the student needs to get information about a specific topic from their professor. Each test session includes two Interactive Listening tasks: one student–student conversation and one student–professor conversation.

Extended Writing and Writing Sample

Each test session includes five extended independent writing tasks, which measure test takers’ English writing abilities: three Picture Description (writing) tasks and two independent writing tasks (Extended Writing and Writing Sample) based on written prompts (see Figures 13 and 14). The Picture Description tasks provide opportunities for test takers to use descriptive language, whereas the

5:38 for the next 5 questions

PASSAGE

Biophysics is the study of the physical properties of living things. This field refers to physics, which is the science of matter and energy, and also to biology, the science of living things. Biophysicists study the physical properties of organisms and the effects of physical processes on living things. For example, biophysicists might study the effect certain chemicals have on living cells, determine how tiny structures within cells work, or explain how injuries and diseases affect the structure of skin. Some biophysicists also study the interaction of radiation with biological systems.

Biophysics is an interdisciplinary field; it is not limited to physics or biology. Biophysicists might also work on projects involving chemistry, geology, and other fields.

Select the best sentence to fill in the blank in the passage.

- They have even studied the physical properties of the cells in the human body.
- Biophysics is an interdisciplinary field; it is not limited to physics or biology.
- Forensic science is the application of the techniques of the physical sciences to analyze evidence.
- The discovery of quantum mechanics in 1925 ushered in a new world of physics.

NEXT

Figure 7. Example Interactive Reading “Complete the Passage” Sub-task

4:21 for the next 4 questions

PASSAGE

Biophysics is the study of the physical properties of living things. This field refers to physics, which is the science of matter and energy, and also to biology, the science of living things. Biophysicists study the physical properties of organisms and the effects of physical processes on living things. For example, biophysicists might study the effect certain chemicals have on living cells, determine how tiny structures within cells work, or explain how injuries and diseases affect the structure of skin. Some biophysicists also study the interaction of radiation with biological systems. Biophysics is an interdisciplinary field; it is not limited to physics or biology. Biophysicists might also work on projects involving chemistry, geology, and other fields.

Click and drag to highlight the answer to the question below.

How does biophysics relate to physics and biology?

Highlight text in the passage to set an answer

NEXT

Figure 8. Example Interactive Reading “Highlight the Answer” Sub-task

independent writing tasks require test takers to demonstrate more discursive knowledge of writing in addition to language knowledge (Weigle, 2002). Both task types 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).

3:46 for the next 3 questions

PASSAGE

Biophysics is the study of the physical properties of living things. This field refers to physics, which is the science of matter and energy, and also to biology, the science of living things. Biophysicists study the physical properties of organisms and the effects of physical processes on living things. For example, biophysicists might study the effect certain chemicals have on living cells, determine how tiny structures within cells work, or explain how injuries and diseases affect the structure of skin. Some biophysicists also study the interaction of radiation with biological systems. Biophysics is an interdisciplinary field; it is not limited to physics or biology. Biophysicists might also work on projects involving chemistry, geology, and other fields.

Select the idea that is expressed in the passage.

- Biophysicists study the physical properties of organisms and how they interact with their environments.
- Electric charges can cause molecular reactions by changing their shape, size, or position.
- Living things are always in motion and they use this motion to perform many functions.
- Cells and tissues are the basic building blocks of living things, such as humans and animals.

NEXT

Figure 9. Example Interactive Reading “Identify the Idea” Sub-task

2:52 for this question

PASSAGE

Biophysics is the study of the physical properties of living things. This field refers to physics, which is the science of matter and energy, and also to biology, the science of living things. Biophysicists study the physical properties of organisms and the effects of physical processes on living things. For example, biophysicists might study the effect certain chemicals have on living cells, determine how tiny structures within cells work, or explain how injuries and diseases affect the structure of skin. Some biophysicists also study the interaction of radiation with biological systems. Biophysics is an interdisciplinary field; it is not limited to physics or biology. Biophysicists might also work on projects involving chemistry, geology, and other fields.

Select the best title for the passage.

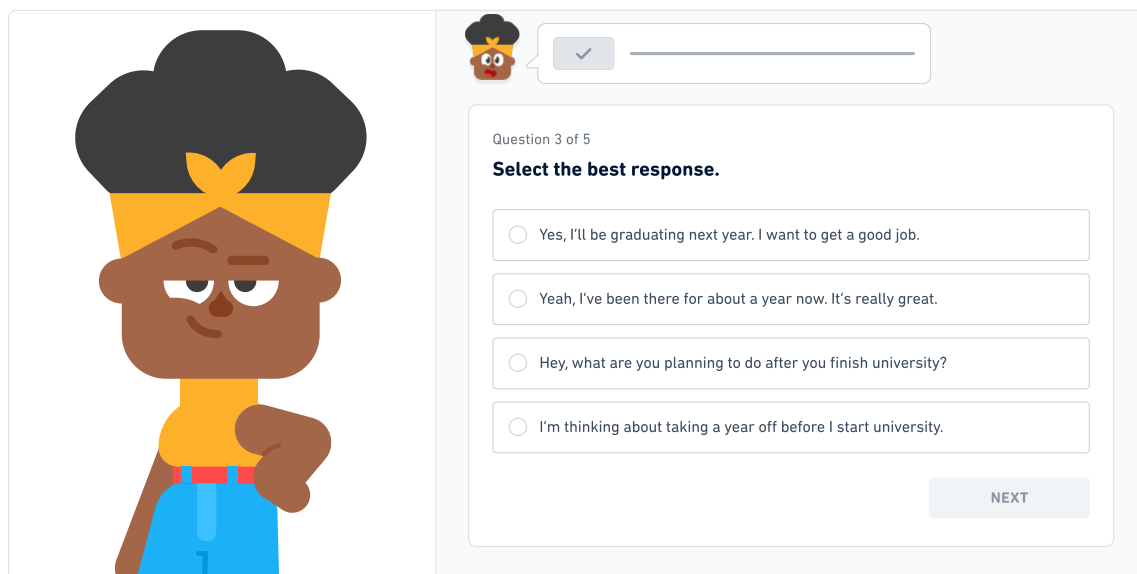
- Computer Simulation of Living Systems
- The Nature of Motion
- The Processes of Life
- An Introduction to Biophysics

NEXT

Figure 10. Example Interactive Reading “Title the Passage” Sub-task

In the Picture Description (writing) tasks, the stimuli (i.e., the photos) were selected by people with graduate degrees in applied linguistics. These images are designed to give test takers the opportunity to display their full range of written language abilities as they contain stimulating depictions of people, animals, and objects in a wide range of contexts. The written prompts for the independent writing tasks ask test takers to describe something, recount an experience, or argue a point of view. The final independent writing task in a test administration is the Writing Sample; a test taker’s written response to this task is provided to institutions with which the test taker shares their results.

2:37 for the next 3 questions



Question 3 of 5

Select the best response.

- Yes, I'll be graduating next year. I want to get a good job.
- Yeah, I've been there for about a year now. It's really great.
- Hey, what are you planning to do after you finish university?
- I'm thinking about taking a year off before I start university.

NEXT

Figure 11. Example Interactive Listening “Dialogue Completion” Sub-task

1:12

Summarize the conversation you just had in 75 seconds.

Your response

NEXT

Figure 12. Example Interactive Listening “Summarization” Sub-task

0:54

Write a description of the image below for 1 minute.



Your response

NEXT

Figure 13. Example Picture Description (writing) Task

4:55

Write about the topic below for 5 minutes.

Discuss the advantages and disadvantages of being the leader of a group. Provide reasons to support your opinion.

Your response

CONTINUE AFTER 3 MINUTES

Figure 14. Example Extended Writing / Writing Sample Task

Extended Speaking & Speaking Sample

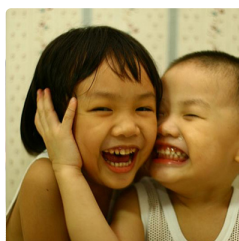
Each test session includes five extended independent speaking tasks measuring test takers' English speaking abilities: one Picture Description (speaking) task and four prompt-based independent speaking tasks (two Extended Speaking [audio prompt], one Extended Speaking [text prompt], and one Speaking Sample which is shared with institutions). All these task types require test takers to speak for

an extended time period and to leverage different aspects of their organizational knowledge (e.g., grammar, vocabulary, and discourse coherence) and functional elements of their pragmatic language knowledge (e.g., ideational knowledge; [Bachman & Palmer, 1996](#)). All extended speaking task types elicit samples that evidence speaking proficiency in terms of the speaking subconstructs of Content, Discourse coherence, Grammar, Vocabulary, Fluency, and Pronunciation. As with the Extended Writing and Writing Sample tasks, test takers must demonstrate proficiency in discussing topics in the different domains described in the CEFR (Personal, Public, Educational, and Professional).

The extended speaking tasks are administered after the computer-adaptive portion of the test. As with the Picture Description (writing) task, the stimuli (i.e., the photos) for Picture Description (speaking) were selected by people with graduate degrees in applied linguistics. These images contain stimulating depictions of people, animals, and objects in a wide range of contexts. For the independent speaking prompts, three are presented as written prompts and one as an aural prompt (see [Figures 15–17](#)). These prompts ask test takers to describe something, recount an experience, or argue a point of view. A recording of a test taker’s spoken response to the Speaking Sample task is provided to institutions with which the test taker shares their results.

1:27

Speak about the image below for 90 seconds.



● RECORDING...



CONTINUE AFTER 30 SECONDS

Figure 15. Example Picture Description (speaking) Task

4 Development, Delivery, and Scoring

This section explains how the computer-adaptive items of the test were developed, how the computer-adaptive portion works, and how the items are scored. Additionally, it provides information about the automated scoring systems for the speaking and writing tasks and how they were evaluated.

4.1 Task and Item Development

All Duolingo English Test task types are designed and approved by language testing experts, who utilize the Expanded Evidence-Centered Design (e-ECD; [Arieli-Attali et al., 2019](#)) approach within the framework of the DET’s theoretical assessment ecosystem ([Burstein et al., 2022](#)). Many expert-designed items are based on authentic English-language content sources. These prompts become input for automatic item generation. In order to create enough items for each task type at varying levels of difficulty, the DET item pool is automatically generated using unique methods for each item type. For example, the reading passages and accompanying items for the Interactive Reading item type are automatically generated by Generative Pre-trained Transformer 3 (GPT-3; [Attali et al., 2022](#); [Park et al., 2022](#)).

1:24



Speak about the topic below for 90 seconds.

Talk about studying science in school.

- What kind of science do students study?
- Is it important to study science?
- Why or why not?


● RECORDING...  CONTINUE AFTER 30 SECONDS

Figure 16. Example Extended Speaking (text prompt) / Speaking Sample Task

1:26



Speak about the topic for 90 seconds.



Number of replays left: 2


● RECORDING...  CONTINUE AFTER 30 SECONDS

Figure 17. Example Extended Speaking (audio prompt) Task

As a result of the large item pool, each test taker only sees a minuscule proportion of existing items, and any two test sessions are unlikely to share a single item. After the items are generated, they go through an extensive fairness and bias (FAB) review process. See Figure 18 for a simplified heuristic of the item development process.

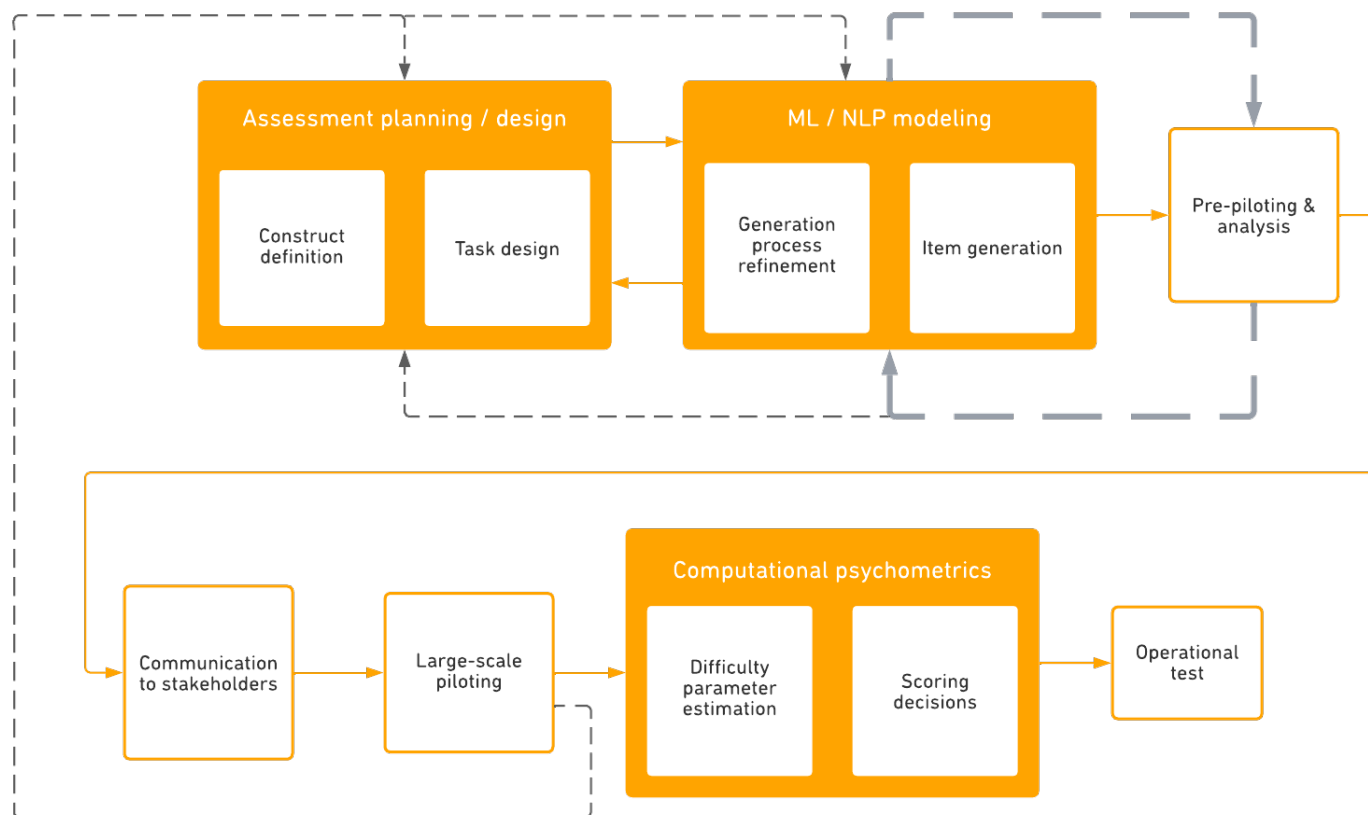


Figure 18. DET Item Generation Process

4.2 Fairness and Bias Review

DET items undergo FAB review by human raters to ensure items are fair towards test takers of diverse identities and backgrounds (e.g., cultural, socioeconomic, and gender). FAB raters are selected to represent diverse identities and perspectives, and all raters have demonstrated experience and interest in promoting equity and diversity. Raters are trained to identify potential sources of construct-irrelevant variance due to either specialized knowledge (e.g., highly technical or culture-specific information) or potentially offensive or distracting content (e.g., cultural stereotypes or descriptions of violence). Items flagged for FAB issues are removed from the item bank. FAB rating data is also used to improve automatic flagging of potentially problematic items. In addition, differential item functioning (DIF) analyses after the test administrations are conducted regularly (Belzak, 2023; Belzak et al., 2023).

4.3 CAT Delivery

Once items are generated, calibrated (\hat{b}_i estimates are made), and placed in the item pool, the DET uses CAT approaches to administer and score tests (Segall, 2005; Wainer, 2000). Because computer-adaptive administration gives items to test takers conditional on their estimated ability, CATs have been shown to be shorter (Thissen & Mislevy, 2000) and provide uniformly precise scores for most test takers when compared to fixed-form tests (Weiss & Kingsbury, 1984).

The primary advantage of a CAT is that it can estimate test-taker ability (θ) more precisely with fewer test items. The precision of the θ estimate depends on the item sequence: test takers of higher ability are best assessed by items with higher difficulty b_i (and likewise for lower values of θ and b_i). The true value of a test taker's ability is unknown before test administration. As a result, an iterative, adaptive algorithm is required.

At the beginning of a test session, a test taker receives a set of items from pre-determined item difficulty ranges in order of increasing difficulty. The CAT algorithm makes a provisional estimate of $\hat{\theta}_t$ based on responses to this item set to time point t . Then the difficulty of the next item is selected as a function of the current estimate: $b_{t+1} = f(\hat{\theta}_t)$. The provisional estimate $\hat{\theta}_t$ is updated after each administered item. Essentially, $\hat{\theta}_t$ is the expected *a posteriori* estimate based on all the administered items up to time point t . This process repeats until a stopping criterion is satisfied.

The CAT approach, combined with concise and predictive task types, helps to minimize test administration time significantly. DET sessions are variable-length, meaning that exam duration and number of items vary across administrations. The iterative, adaptive procedure continues until the test exceeds a maximum length in terms of minutes or items, as long as a minimum number of items has been administered. Most tests are less than an hour long (including speaking and writing; excluding onboarding and uploading) while collecting over 200 item measurements.

Once the stopping criterion is satisfied, an ability estimate is calculated on responses to each CAT task type separately. These score estimates of each CAT task type are then used with the scores of the Interactive Listening, speaking, and writing tasks to compute an overall score and the four subscores.

4.4 CAT Item Scoring

All test items are graded automatically via statistical procedures appropriate for the task type. For two CAT task types—C-test and Yes/No Vocabulary—each task comprises multiple discrete items to which responses are deemed correct or incorrect. In the case of C-test tasks, completing each damaged word is a distinct item. For Yes/No Vocabulary, deciding whether an individual stimulus is a real English word is an item. Such task types are scored with a 2PL (two-parameter logistic) item response theory (IRT) model, for which the parameters were estimated on response data from all valid test sessions of a one-year period. Regression calibration (Carroll et al., 2006) was used in item calibration to control for the differing ability of test takers responding to each item, since items are administered based on a test taker's responses to previous items, and thus more difficult items are seen by more able test takers. Each item of the aforementioned task types has a unique set of 2PL item parameters, which are then used to estimate an expected ability based on a test taker's responses to all administered items of the same type.

The next CAT task type administered—Interactive Reading—comprises both selected-response sub-tasks with a clearly defined number of response options and a highlight sub-task with a large, undefined number of possible responses. Items for selected-response sub-tasks are graded dichotomously (correct/incorrect) and scores estimated via 2PL IRT models. Responses to the highlighting sub-task are compared to a single correct answer (a particular part of the text). For grading purposes, a text selection is defined as a point in the two-dimensional space for the location of the start and end indices of the selection. A continuous grade between 0 and 1 is then calculated based on the discrepancy (geometric distance) between the point representations of the response and the correct answer.

The final CAT task type—Interactive Listening—also comprises a mixture of sub-tasks. Items for the dialogue completion sub-task are graded dichotomously and scores estimated via 2PL IRT models. The written summary sub-task is scored using a sub-task-specific automated scoring model, as described in the next section.

For Dictation items, responses are graded on a $[0, 1]$ scale as a function of the edit distance^{*}. The maximum grade value is 1, occurring when the provided response is identical to the expected response. Values less than one indicate various degrees of accuracy. Item grades are then used, in combination with item difficulty parameters, to estimate test-taker ability. Because a substantial proportion of Dictation responses receive a perfect grade, item difficulty parameters are estimated with a custom model that combines elements of models for both discrete and continuous data, similar to the model of Molenaar et al. (2022).

The responses to the Elicited Imitation tasks are aligned against an expected reference text, and similarities and differences in the alignment are evaluated to produce a grade on a $[0, 1]$ scale. In addition to the alignment of responses to the target text, Elicited Imitation grades also incorporate data on speaking features like pronunciation and rate of speech.

^{*}“Edit distance” is a concept from natural language processing referring to the number of single-letter modifications necessary to transform one character string into another. It is used as a measure of similarity between two text strings.

4.5 Extended Speaking and Writing Task Scoring

The speaking and writing tasks are scored by automated scoring systems developed by experts at Duolingo in the fields of machine learning (ML), natural language processing (NLP), and applied linguistics: the *Duo Speaking Scorer* and *Duo Writing Scorer*, with separate scoring models for the different task types. These models evaluate each item response based on a number of theoretical speaking and writing subconstructs (i.e., factors contributing to speaking and writing quality). These subconstructs are reflected in [human scoring rubrics*](#) and are operationalized for automated scoring through the measurement of numerous research-supported linguistic features. Table 3 presents these subconstructs for speaking and writing and provides examples of how these subconstructs are described in both human and automated scoring.

Table 3. Extended speaking and writing scoring subconstructs

Subconstruct	Example dimensions	Example automated feature
Content	Task achievement, Relevance, Effect on the reader/listener, Appropriacy of style, Development	the cosine similarity between the prompt's embedding and the response's embedding (relevance feature)
Discourse coherence	Clarity, Cohesion, Progression of ideas, Appropriacy of format, Structure (writing only)	binary overlap between sentence pairs: overlap of arguments, nouns, or word stems between two sentences (cohesion feature)
Lexis	Lexical diversity, Lexical sophistication, Word choice, Word formation, Error severity, Spelling (writing only)	the proportion of lemmatized words from the response that are level CEFR C1 and above (lexical sophistication feature)
Grammar	Range of structures, Grammatical complexity, Error frequency, Error severity, Appropriacy	the mean tree depth among the dependency trees of each sentence in the response (grammatical complexity feature)
Fluency (speaking only)	Speed, Chunking, Breakdowns, Repairs	number of words per second (speed feature)
Pronunciation (speaking only)	Intelligibility, Comprehensibility, Individual sounds, Word stress, Rhythm, Sentence stress, Intonation, Connected speech	the acoustic model's confidence in the transcription (intelligibility feature)

Duo Writing Scorer was evaluated on 2,460 test sessions,[†] and the *Duo Speaking Scorer* was evaluated on 1,922 test sessions. Each session had 1-2 responses rated by a human rater using the rubrics previously described, and the human ratings for each session were averaged. These raters are experts in the field of English language teaching and assessment and possess the following minimum requirements: 5 or more years of TESOL experience with adults; TESOL certification (e.g., Cambridge DELTA); relevant Bachelor's degree or equivalent; expert English proficiency (C2 on CEFR); and experience in high-stakes speaking/writing assessment. Raters undergo initial training, must pass a certification test, and are regularly monitored. Table 4 shows the correlation (Pearson r) between the average human rating and the automated score (i.e., Human-Machine) and between human ratings for single responses (i.e., Human-Human) for extended speaking and writing tasks. Pearson r correlations over 0.50 can be interpreted as large (Cohen, 1988). That human-machine and human-human correlations are so similar indicates that the automated scorers are attending to and equally weighting the same subconstructs of speaking and writing.

*http://go.duolingo.com/DET_speaking_and_writing_rubrics

†The dataset used for evaluating the *Duo Writing Scorer* pre-dated the introduction of the Interactive Listening summarization task.

Table 4. Correlation of extended productive skills tasks by scoring method

	Human–Machine (<i>r</i>)	Human–Human (<i>r</i>)
Speaking tasks	0.84	0.85
Writing tasks	0.85	0.87

4.6 Subscores

In addition to the overall score, the DET reports four subscores* that are also on a scale of 10–160 and assess four integrated skill areas: Literacy (reading and writing tasks), Conversation (speaking and listening tasks), Comprehension (reading and listening tasks), and Production (speaking and writing tasks). LaFlair (2020) provides multivariate analyses of DET response data that support this skill configuration and shows that subscores estimating these skills have reliability and added value (beyond an overall score) that meet professional standards for subscore validity (Sinharay & Haberman, 2008). Each subscore can be interpreted as a combination of two of the more traditional language subskills: speaking, writing, reading, and listening (SWRL). Figure 19 shows the relationship between the DET task types, the subscores, and SWRL subskills.

		Literacy	Conversation	Comprehension	Production
Reading	C-Test	X		X	
	Interactive Reading	X		X	
	Yes/No Vocabulary	X		X	
Listening	Dictation		X	X	
	Interactive Listening (dialogue completion)		X	X	
Writing	Extended Writing	X			X
	Interactive Listening (summarization)	X			X
	Picture Description (writing)	X			X
	Writing Sample	X			X
Speaking	Elicited Imitation		X	X	
	Extended Speaking (audio prompt)		X		X
	Extended Speaking (text prompt)		X		X
	Picture Description (speaking)		X		X
	Speaking Sample		X		X

Figure 19. Contribution of Task Types to DET Subscores

*Due to the way the subscores are computed, there may be cases where test takers with the same overall score have different subscore profiles.

5 Access and Accommodations

Given the Duolingo English Test’s mission to lower barriers and increase opportunities for English learners, broad accessibility is one of the central motivations for the test’s existence and a primary consideration in any changes to the test. A combination of universally accessible features and accommodations for test takers with disabilities ensures that all test takers have an equal opportunity to demonstrate their English proficiency.

5.1 Access

The DET reflects principles of Universal Design (UD), a framework for designing products and spaces with the goal of maximum accessibility from the start; the concept originated in the field of architecture but has also been adapted to assessment design (Thompson et al., 2002). Maximizing test accessibility through intentional design benefits all test takers, both those with and without disabilities, while simultaneously reducing the need for selective accommodations. The ethos of UD is evident in the origin of the DET and the DET’s assessment ecosystem (Burstein et al., 2022), in which all aspects of test design and administration are infused with consideration of the test-taker experience (TTX). The DET’s at-home on-demand approach, intuitive user interface, and asynchronous proctoring collectively are designed to reduce physical, socioeconomic, and psychological barriers to test access and optimal test performance.

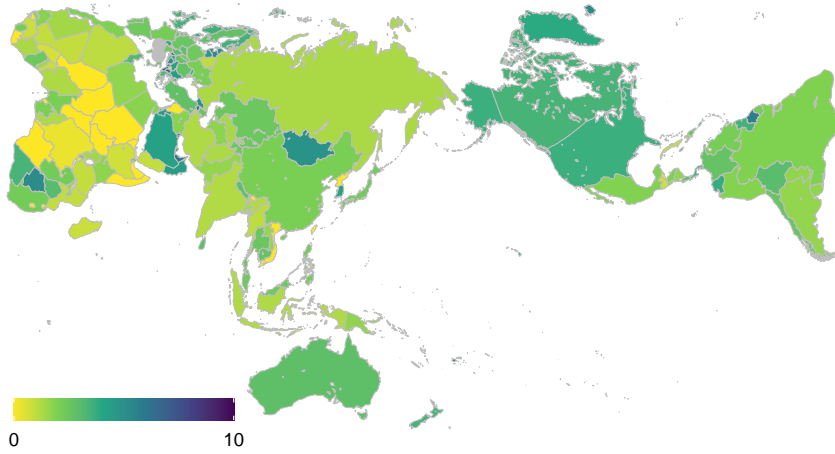
Perhaps the most salient accessibility benefit of the DET is that at-home testing obviates the need to travel to a physical test center. Traveling to a test center can be burdensome for both socioeconomic and disability-related reasons. Test centers are necessarily concentrated in relatively large urban areas, and some countries do not have any test centers that administer high-stakes ELP tests. It is also not guaranteed that a prospective test taker can obtain a test seat at their closest test center at a time that meets their needs. Many test takers therefore must spend time and money to travel significant distances, even internationally, in order to take a test. This burden is compounded for test takers with disabilities, who might require special transportation or assistance. For such test takers, even local travel can pose a non-trivial barrier. The DET allows most individuals to have their English proficiency evaluated from the most accessible location—their own home.

The AuthaGraph maps (Rudis & Kunimune, 2020) in Figure 20 visualize the issue of physical test access by showing the concentration of test centers in the world (top panel) compared to internet penetration in the world (middle panel), and the concentration of DET test takers (bottom panel; for all tests administered since August 1, 2017). The top two panels of Figure 20 show how much more easily an internet-based test can be accessed than a test center (although Central Africa is admittedly underserved by both models). While the ratio of population to internet access and to test center access is a somewhat limited metric, the potential audience for the DET is clearly orders of magnitude larger than those with access to traditional test centers. The map in the bottom panel shows that the DET is beginning to realize this potential, with people taking the DET from places with relatively low concentrations of test centers (e.g., Colombia, Kazakhstan, and Zimbabwe). By delivering assessments on-demand, 24 hours a day, on any of the world’s estimated two billion internet-connected computers, the DET is at the forefront of maximizing test access while maintaining test use validity and test security.

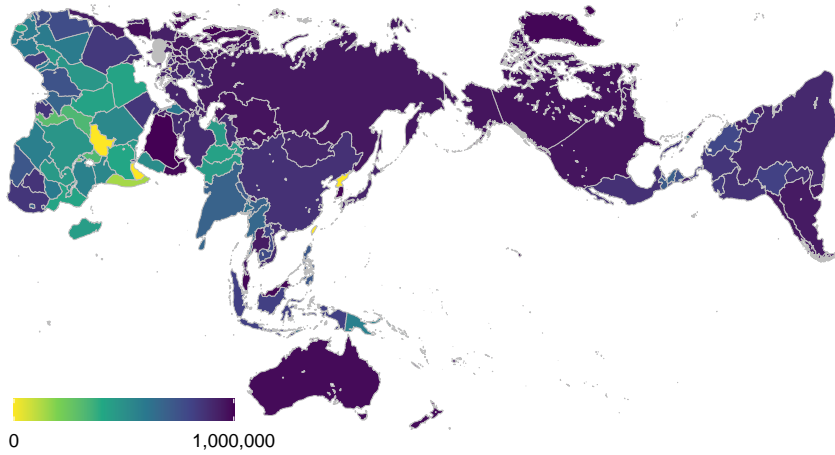
In addition to lowering physical barriers to test access, the DET also embodies accessibility in the economic sense, most obviously through its registration fee, which is a fraction of alternative tests’ fees. Furthermore, the DET does not charge extra fees for sharing scores with institutions or appealing proctoring decisions. The DET’s at-home on-demand nature removes the need to travel to a test center, potentially representing a cost saving several times greater than the test fee itself. These factors collectively reduce a potentially insurmountable barrier to taking an English language proficiency test, and also make it more feasible for test takers to reattempt the test if needed. The DET’s [Access Program](#) further reduces socioeconomic barriers for test takers with the greatest need by routinely providing fee waivers to institutions, providing fee waivers to organizations working with populations affected by natural disasters and armed conflicts, and partnering with the UNHCR to provide college counseling to refugee students.

Once test takers have gained access to the DET, the test’s design also reduces construct-irrelevant barriers to optimal test performance that could arise during the testing experience. Testing at home gives test takers control over the setup of their testing environment, including the furniture, lighting, and equipment, allowing them to take the test comfortably. This feature is particularly beneficial for test takers with disabilities who may require medical devices or special computer equipment such as screen magnification or a special keyboard. The ability to test in a comfortable and familiar environment can also reduce test anxiety (Stowell & Bennett, 2010). The relatively short duration of the test, facilitated by the DET’s adaptive nature, may be beneficial for test takers who cannot sit and/or concentrate continuously for long periods due to physical and/or psychological disabilities. The DET’s user interface complies with W3C Web Content Accessibility Guidelines (WCAG) 2.1 Level AA. Furthermore, the DET’s use of asynchronous proctoring (see Section 6.4) likely has a positive impact on TTX, as it does not require interaction with a human proctor and the accompanying concerns about privacy and potential interruptions during testing.

Number of Test Centers per Million People



Number of Internet Users per Million People



Number of Duolingo Test Takers per Million People

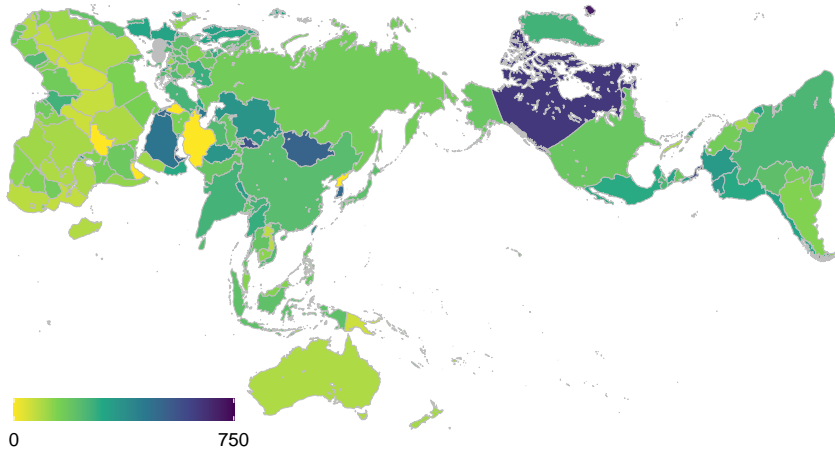


Figure 20. Heatmaps of Test Center Accessibility as of 2023 (top), Internet Accessibility as of 2022 (middle), and Concentration of DET Test Takers as of 2023 (bottom)

5.2 Accommodations

The DET’s inherently accessible design features reduce the need for certain testing accommodations (e.g., extended breaks between test sections). Nevertheless, the DET provides accommodations for both physical (e.g., visual or hearing impairment) and psychological (e.g., autism spectrum disorder) conditions that could constitute construct-irrelevant barriers to optimal test performance. To receive an accommodation, test takers must submit a request at <https://englishtest.duolingo.com/accommodations> describing both their reason for requesting an accommodation (with supporting documentation, if applicable) and the accommodation requested. The available accommodation options are

- 50% extra time per question
- Accessibility devices (alternate keyboard, etc.)
- Hearing aids
- Headphones
- Listening device
- Screen magnifier/reader
- Other accommodation (to be described by the test taker)

All requests for documented needs are accommodated to the extent reasonable. To ensure accessibility, we have significantly streamlined the process for requesting accommodations compared to the industry standard. The DET requests similar documentation to other English proficiency tests but only requires test takers to fill out a single online form. All inquiries receive a response within three days.

6 Test Administration and Security

The Duolingo English Test is administered to test takers via the internet. The security of DET scores is ensured through a robust and secure onboarding process, automated security measures, rules that test takers must adhere to during the test administration, and a strict proctoring process. All test sessions are proctored after the test has been administered and prior to score reporting. Additional security is also provided by the DET’s large item bank, CAT format, and active monitoring of item exposure rates, which collectively minimize the probability that test takers can gain any advantage through item pre-knowledge (i.e., exposure to test content before encountering it during an operational test session). Item pre-knowledge is further minimized by preventing repeat test takers (i.e., individuals who take the test more than once) from seeing the same item within a certain period.

Overall, the test security framework is an essential dimension of the larger assessment ecosystem (Burstein et al., 2022), used to protect the integrity of test scores at all stages of the assessment process (LaFlair et al., 2022). The remainder of this section presents a summary of the information found in the [Security, Proctoring, and Accommodations](#) whitepaper (Duolingo English Test, 2021).

6.1 Test Administration

Test takers are required to take the test alone in a quiet environment on a laptop or desktop computer running Windows or macOS and equipped with a front-facing camera, a microphone, and speakers (headphones are not permitted). An internet connection with at least 2 Mbps download speed and 1 Mbps upload speed is recommended for test sessions. Test takers are required to take the test through the DET desktop app, which provides a more stable and secure test-taking experience. Test takers are prompted to download and install the desktop app after clicking “Start Test” on the DET website. The desktop app automatically prevents navigation away from the test and blocks tools such as spelling and grammar checkers and automatic word completion. For test sessions that take place in a browser, the browsers are locked down after onboarding, meaning that any navigation away from the browser invalidates the test session. Additionally, browser plugins are automatically detected, and test takers are required to disable them before beginning the test.

6.2 Onboarding

Before the test is administered, test takers complete an onboarding process. This process checks that the computer's microphone and speaker work. It is also at this time that test takers are asked to show identification and are informed of the test's administration rules, which they must agree to follow before proceeding. In order to ensure test-taker identity, an identity document (ID) must be presented to the webcam during onboarding. An image of the ID is captured*. IDs must meet certain criteria, such as being government-issued, currently valid, and including a clear picture of the test taker.

6.3 Administration Rules

The behaviors that are prohibited during an administration of the DET are listed below. These rules require test takers to remain visible to their cameras at all times and to keep their camera and microphone enabled throughout the test administration. The rules are displayed in the test taker's chosen interface language† to ensure comprehension. Test takers are required to acknowledge understanding and agree to these rules before proceeding with the test. If the test session is automatically terminated for reasons such as moving the mouse off-screen or a technical error, a test taker may attempt the test again for free, up to a total of three times. Test takers may contact customer support to obtain additional test attempts in the case of recurring technical errors. Other reasons for test cancellation include:

- Leaving the camera preview
- Looking away from the screen
- Covering ears
- Leaving the web browser
 - Leaving the window with the cursor
 - Exiting full-screen mode
- Speaking when not instructed to do so
- Communicating with another person at any point
- Allowing others in the room
- Using any outside reference material
- Using a phone or other device
- Writing or reading notes
- Disabling the microphone or camera

6.4 Proctoring and Reporting

After the test has been completed and uploaded, all DET sessions undergo a thorough proctoring review by trained human proctors with TESOL/applied linguistics expertise. This review is supplemented by artificial intelligence to call proctors' attention to suspicious behavior. Proctors have access to both audio and video recordings of the entire test session, including both a view of the test taker and a recording of the computer screen. Each test session is reviewed independently by at least two proctors. When necessary, the test session is sent to a third level of review, to be evaluated by a senior proctor or operations manager. This process takes no more than 48 hours after the test has been uploaded. After the process has been completed, score reports are sent electronically to the test taker and any institutions with which they have elected to share their scores. Test takers can share their scores with an unlimited number of institutions. While AI provides assistance at every stage of proctoring, the proctors make the final decision on whether to certify a test. Certain invalid results are eligible to be appealed within 72 hours by submitting a form from the test taker's homepage describing the reason for the appeal. Once the form has been submitted, the test taker will receive an emailed response within four business days informing them of the appeal ruling.

*ID images are stored temporarily in a highly secure digital repository in compliance with all applicable data privacy regulations and best practices.

†Currently available user interface languages: Chinese, English, French, German, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Thai, Turkish, Vietnamese

7 Test-Taker Demographics

This section summarizes test-taker demographics based on all certified Duolingo English Test sessions between December 15, 2022 and December 14, 2023. During the onboarding and offboarding process of each test administration, test takers are asked to report their first language (L1), date of birth, reason for taking the test, and their gender identity. The issuing country/region of test takers' identity documents is logged when they show government-issued identification during the onboarding process.

Reporting gender identity during the onboarding process is optional, but reporting date of birth is required. Table 5 shows an approximately even distribution of male and female gender identities. However, the gender distribution of test takers varies considerably across countries. Figure 21 depicts the proportion of reported gender identities for all countries with more than 300 test takers, ranging from a maximum of 83% male to a maximum of 67% female.

Table 5. Percentages of Test-Taker Gender (December 15, 2022 — December 14, 2023)

Gender	Percentage
Female	47.04%
Male	52.83%
Other	0.13%
Total	100.00%

The median test-taker age is 22. Table 6 shows that 81% of DET test takers are between 16 and 30 years of age at the time of test administration.

Table 6. Percentages of Test-Taker Age (December 15, 2022 — December 14, 2023)

Age	Percentage
< 16	3.21%
[16, 21)	31.14%
[21, 26)	34.93%
[26, 31)	15.32%
[31, 41)	11.49%
≥ 41	3.90%
Total	100.00%

Test takers are asked to report their L1s during the onboarding process. The most common first languages of DET test takers include Mandarin, Spanish, Arabic, English*, French, and Portuguese (see Table 7). There are 147 unique L1s represented by test takers of the DET, and the test has been administered to test takers from 214 countries and dependent territories. The full tables of all test-taker L1s and places of origin can be found in the Appendix (Section 11).

Table 7. Most Frequent Test-Taker L1s (December 15, 2022 — December 14, 2023)

First Language
Chinese - Mandarin
English
Spanish
Telugu
Arabic
Portuguese
Korean
French
Indonesian
Hindi

* 63% of English-L1 test takers come from India and Canada

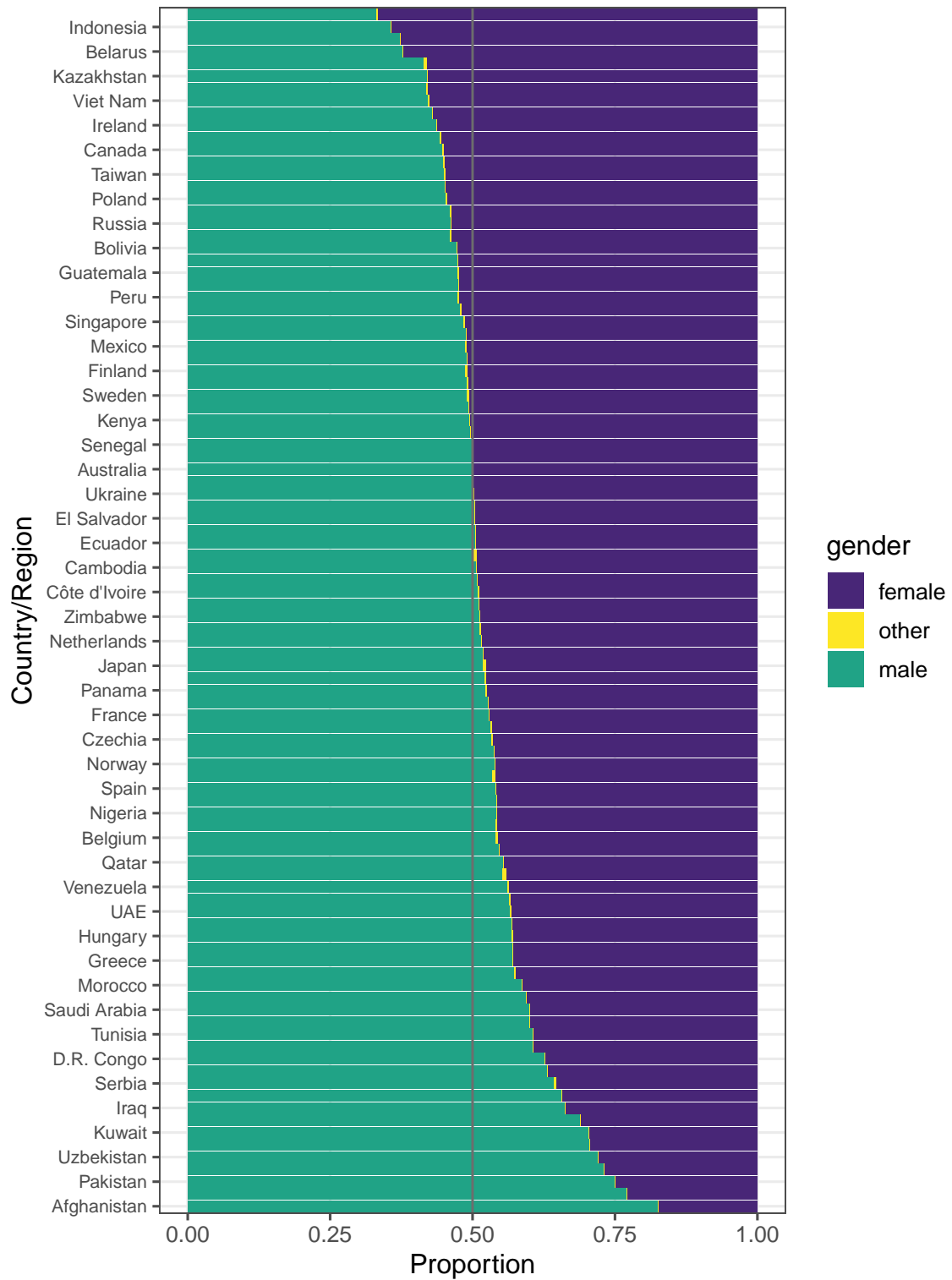


Figure 21. Proportion of Reported Gender Identities for all Countries and Territories with >300 Test Takers in Past 12 Months (only every other country labeled)

For each test session, the issuing country of the test taker’s identity document is recorded, as well as the country in which they are taking the test. For 83% of test takers, the ID issuing country and the country in which they take the test are the same. The other 17% represent test takers who are presumably residing outside of their country of origin when they take the DET. Tables 8 and 9 display, for such test takers, the top ten testing locations and the top ten ID issuing countries, respectively.

Table 8. Most Frequent Testing Locations for Test Takers Residing Outside Their Country of Origin (December 15, 2022 — December 14, 2023)

Top testing locations
USA
Canada
UK
Ireland
China
Germany
UAE
Hong Kong
Singapore
Saudi Arabia

Table 9. Most Frequent ID Issuing Countries for Test Takers Residing Outside Their Country of Origin (December 15, 2022 — December 14, 2023)

Top ID origins
China
India
South Korea
Ukraine
Brazil
Mexico
USA
Colombia
Russia
Japan

Test takers are also asked to optionally indicate their intention for taking the DET, with the choice of applying to a school (secondary, undergraduate, or graduate) and job-related purposes. Table 10 presents the distribution of test-taker intentions.

Table 10. Percentages of Test-Taker Intention (December 15, 2022 — December 14, 2023)

Intention	Percentage
Undergrad	43.02%
Grad	37.42%
Secondary School	5.73%
Work	1.66%
None of the Above	6.65%
(No Response)	5.53%

8 Test Performance Statistics

This section provides an overview of the statistical characteristics of the Duolingo English Test, including information about the score distributions and reliability of the overall score and subscores. The analyses of the subscores were conducted on data from tests that were administered between May 01, 2023 and December 14, 2023.

8.1 Score Distributions

Figure 22 shows the distribution of scores for the overall score and subscores (on the x-axis of each plot). From top to bottom, the panels show the distribution of test scores for the four subscores and the overall score using three different visualization techniques. The left panels show a boxplot of the test scores. The center panels show the density function of the test scores, and the right panels show the empirical cumulative density function (ECDF) of the test scores. (The value of the ECDF at a given test score is the proportion of scores at or below that point.) The density plot (center panel) provides evidence that the test is of appropriate difficulty (neither too easy nor too hard) for the test-taker population, given that there is no apparent ceiling effect on the overall score.

The plots in Figure 22 show some negative skew, which is reflected in the descriptive statistics in Table 11. The overall score mean and median are 108.65 and 110 respectively, and the interquartile range is 25. Tables 17–19 in the Appendix show the percentage and cumulative percentage of the total test scores and subscores. These are numerical, tabled representations of the plots in Figure 22.

Table 11. Descriptive Statistics for Total and Subscores (n = 99,415) (May 01, 2023 — December 14, 2023)

Score	Mean	SD	25th Percentile	Median	75th Percentile
Comprehension	115.99	20.19	105	115	130
Conversation	99.48	21.64	85	100	115
Literacy	109.41	21.76	95	110	125
Production	87.69	23.34	75	90	105
Total	108.65	20.30	95	110	120

8.2 Reliability Evidence

The reliability of the DET is evaluated by examining the relationship between multiple scores from repeat test takers (test–retest reliability) and the standard error of measurement (SEM). The data used to estimate each of these measures come from a subset of the 161,335 certified tests administered between March 29, 2022 and August 05, 2022.

There are two main challenges with using repeaters to estimate test reliabilities for the full test-taking population. The first is that repeaters are a self-selected, non-random subset of the full testing population. People who choose to repeat tend to represent a more homogeneous, lower-ability subpopulation than the full testing population. Unless addressed, this reduction in ability heterogeneity will tend to artificially reduce estimated reliabilities based on repeaters. The second challenge is that repeaters not only self-select *to* repeat the test, but also self-select *when* to repeat the test. Some repeaters take the test twice in a short period, while other repeaters may wait a year or more to retest. The more time that passes between repeat test takers’ sessions, the more opportunity there is for change in test takers’ true proficiency. Change in proficiency over time, which varies across individuals, must also be accounted for to avoid artificially reducing reliability estimates.

In order to address the challenges inherent to test–retest reliability, the analysis was conducted on a sample of repeaters who took the DET twice within two days. The restriction to such repeaters is intended to reduce the impact of heterogeneous proficiency changes on estimated test–retest reliability. To address the fact that repeaters are different from the full population of first-time test takers, DET assessment scientists used Minimum Discriminant Information Adjustment (MDIA; [Haberman, 1984](#)). Specifically, MDIA was used to compute weights so that the weighted repeater sample matches all first-time test takers with respect to country, first language, age, gender, computer operating system (Windows vs MacOS), TOEFL overall scores, IELTS overall scores, and the means and variances of the DET scores on the first attempt. Weighting in this manner mitigates the potential biasing effects of repeater self-selection on test–retest reliability estimates ([Haberman & Yao, 2015](#)). A weighted test–retest correlation was calculated for the overall score and all four subscores. Bootstrapping was used to calculate normal 95% confidence intervals for each reliability estimate.

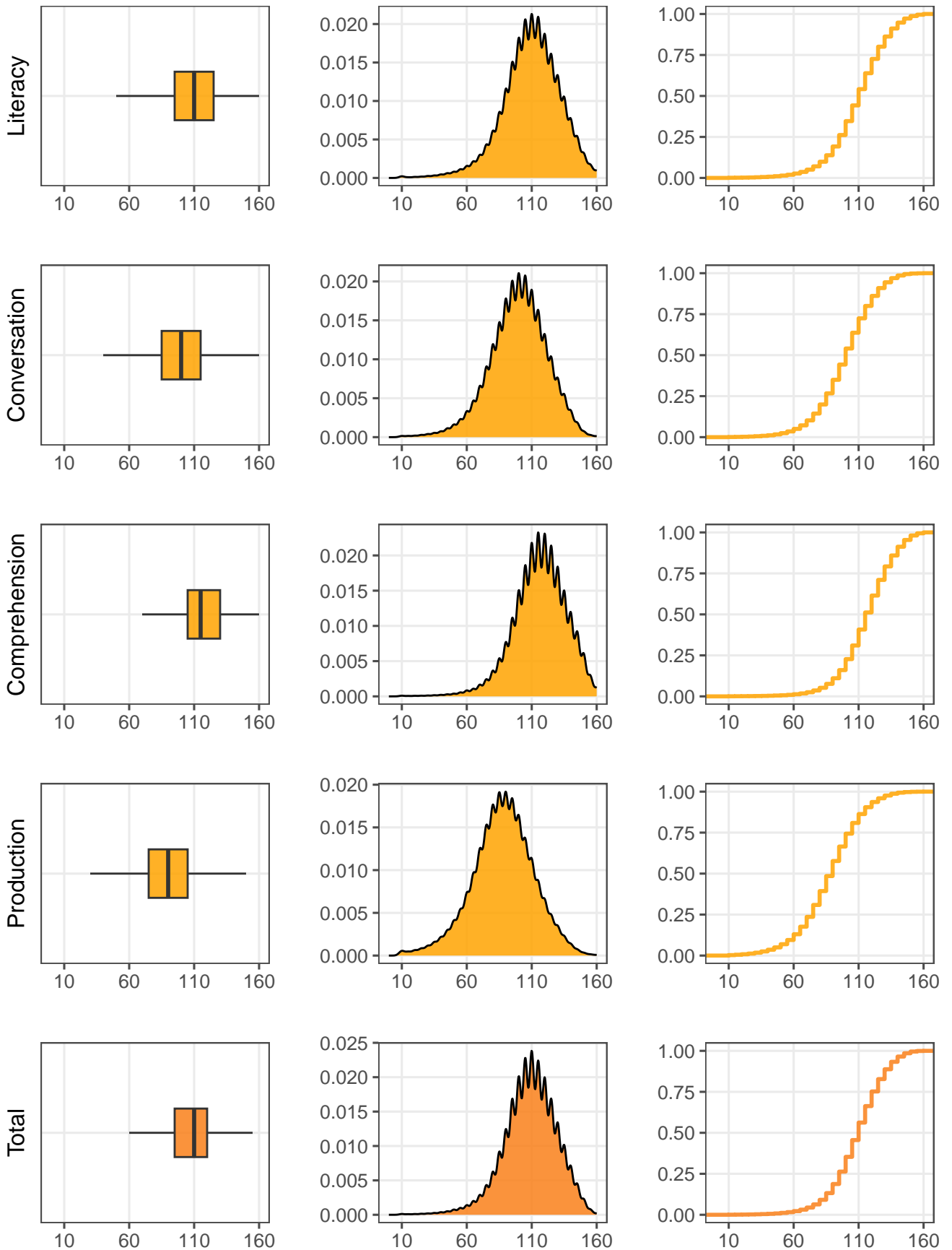


Figure 22. Boxplots (left), Density Plots (middle), and Empirical Cumulative Distribution Plots (right) of the Overall Score and Subscores.

The point estimates and confidence intervals of the reliabilities for the DET overall score and subscores are shown in Table 12. The subscore reliabilities are slightly lower than the overall score reliability. This finding is expected because subscores are calculated from a smaller number of items. The SEM is estimated based on the standard deviation of the overall score or subscore and the corresponding test–retest reliability estimate. The SEM is a statistic that reflects the accuracy or precision of a score, indicating how much a person’s score might vary if the test were taken multiple times. A smaller SEM indicates higher test reliability.

Table 12. Test-Retest Reliability and SEM Estimates (March 29, 2022 — August 05, 2022)

Score	Test–Retest	Lower CI	Upper CI	SEM
Literacy	0.90	0.88	0.92	6.64
Conversation	0.90	0.88	0.92	6.57
Comprehension	0.92	0.90	0.94	5.71
Production	0.90	0.88	0.91	7.14
Overall	0.93	0.91	0.95	4.95

8.3 Relationship with Other Tests

In 2022, correlational and concordance studies were conducted (Cardwell et al., 2023) to examine the relationship between DET scores and scores from TOEFL iBT and IELTS Academic—tests designed to measure similar constructs of English language proficiency and used for the same purpose of postsecondary admissions. The data for these studies are the results of certified DET sessions since the launch of the Integrated Reading task type on March 29, 2022, as well as associated TOEFL or IELTS scores for a subset of test takers.

DET assessment scientists designed a study to collect official TOEFL and IELTS score reports from DET test takers. Test takers could submit official score reports in exchange for payment or a credit to take the DET again (referred to subsequently as the “official score data”). Prior to any analysis, official score data were assembled, checked, and cleaned by Duolingo assessment scientists and a research assistant. In order to achieve recommended minimum sample sizes of 1,500* (Kolen & Brennan, 2004, p. 304) for both TOEFL and IELTS data, as well as to represent a greater range of test-taker ability, the official score data were supplemented with self-report data. DET test takers have the opportunity to voluntarily report TOEFL or IELTS results at the end of each test session. Table 13 reports the sizes of the final analytic samples after data cleaning (e.g., removing out-of-range scores and records with invalid subscore–overall score relationships) and restricting the data to DET–TOEFL and DET–IELTS score pairs from test dates less than four months apart.

Table 13. Sample Sizes for Correlation and Concordance Analyses (March 29, 2022 — August 05, 2022)

	TOEFL	IELTS
Official	328	1,643
Self-report	1,095	4,420

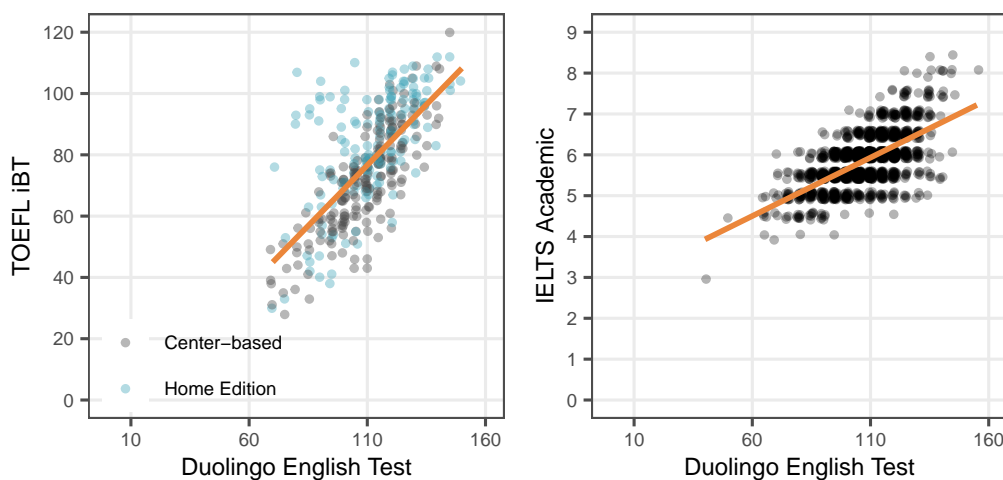
Correlation

Pearson’s correlation coefficients were estimated from official score data to evaluate the relationship between the DET and the TOEFL iBT and IELTS Academic (Table 14). The correlation coefficients show strong, positive relationships of DET scores with TOEFL iBT scores and with IELTS scores. These relationships are visualized in Figure 23. The left panel shows the relationship between the DET and TOEFL iBT, and the right panel shows the relationship between the DET and IELTS. Values in parentheses are the sample sizes corresponding to each condition.

*This recommended minimum is for the equivalent-groups design. The necessary minimum sample size for a single-group design is theoretically smaller, but a specific number is not given, and so we take 1,500 as the acceptable minimum.

Table 14. Correlations Between DET Scores and TOEFL / IELTS Scores (March 29, 2022 — August 05, 2022)

	TOEFL	IELTS
All candidates	.71 (328)	.65 (1,643)
Center-based	.82 (183)	—
Home Edition	.61 (145)	—

**Figure 23.** Relationship Between Test Scores

Concordance

Given that a sample size of 1,500 is the recommended minimum for building a concordance table using standard equipercentile equating methods (Kolen & Brennan, 2004, p. 304), self-report and official data were both included in the concordance study. Assessment scientists first used data of individuals who both self-reported an external score and submitted an official score report to estimate potential reporting bias in self-report data. MDIA was used to correct for this reporting bias. Follow-up analyses demonstrated that the resulting, adjusted scores had approximately the same properties as the official scores. The DET–IELTS concordance results computed on the official data and on the combined data were compared to confirm that the combined data set is unbiased. The sample of those who took both the DET and IELTS was sufficiently large to allow for this comparison. After correcting for reporting bias, the self-report and official data were then combined prior to performing final equating. For individuals with external scores in both the self-report and official score data, only the official score records were retained in the combined data.

Two types of equating were compared in a single-group equating design: equipercentile (Kolen & Brennan, 2004) and kernel equating (von Davier et al., 2004). The equating study was conducted using the *equate* (Albano, 2016) and *kequate* (Andersson et al., 2013) packages in R (R Core Team, 2022). Additionally, the data were presmoothed using log-linear models (von Davier et al., 2004) prior to applying the equating methods. The equating methods were evaluated by looking at the final concordance as well as the standard error of equating (SEE), which were estimated via bootstrapping. The final concordance was very similar when comparing equipercentile and kernel equating methods. The standard errors were also very similar across equating methods, although kernel equating had slightly lower and more stable standard errors than equipercentile equating, especially for IELTS given the shorter scale. For these reasons, kernel equating was chosen as the final equating method.

The concordance with IELTS exhibits less error overall because the IELTS score scale contains fewer distinct score points (19 possible band scores between 1 and 9) than the DET (31 possible score values), meaning test takers with the same DET score are very likely to have the same IELTS score. Conversely, the TOEFL scale contains a greater number of distinct score points (121 unique score values), leading to relatively more cases where a particular DET score can correspond to multiple TOEFL scores, which inflates the SEE. The concordance tables can be found on the DET scores page (<https://englishtest.duolingo.com/scores>).

9 Quality Control

The unprecedented flexibility, complexity, and high-stakes nature of the Duolingo English Test pose quality assurance challenges. In order to ensure the test is of high quality at all times, it is necessary to continuously monitor processes associated with the DET ecosystem frameworks and key summary statistics of the test. Doing so allows for the prompt identification and correction of any anomalies.

9.1 Analytics for Quality Assurance in Assessment

The DET utilizes a custom-built psychometric quality assurance system, Analytics for Quality Assurance in Assessment (AQuAA; Liao et al., 2022), to continuously monitor test metrics and trends in the test data. AQuAA is an interactive dashboard that blends educational data mining techniques and psychometric theory, allowing the DET’s psychometricians and assessment scientists to continuously monitor and evaluate the interaction between the test items, the test administration and scoring algorithms, and the samples of test takers, ensuring scores are consistent over many test administrations. As depicted in Figure 24, test data such as test-taker demographics, item response durations, and item scores are automatically imported into AQuAA from DET databases. These data are then used to calculate various statistics, producing intermediate data files and data visualizations, which are regularly reviewed by a team of psychometricians in order to promptly detect and respond to any anomalous events.

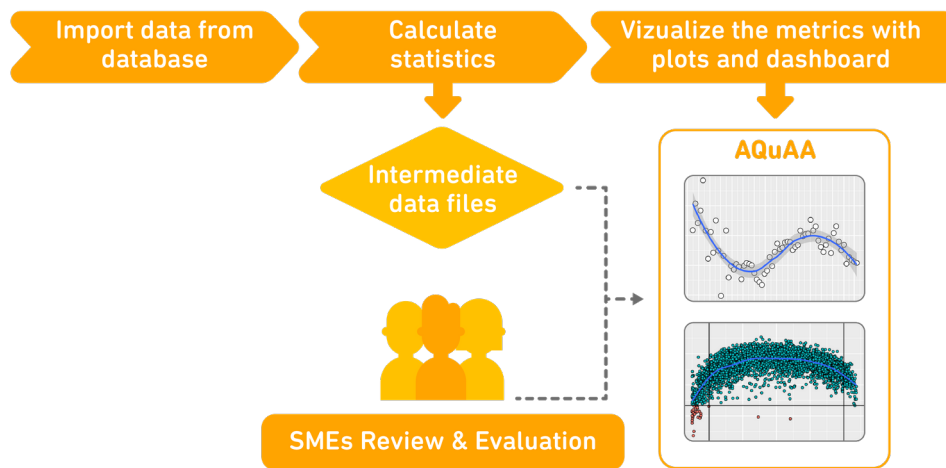


Figure 24. DET Quality Control Procedures

AQuAA monitors metrics over time in the following five categories, adjusting for seasonality effects.

1. **Scores:** Overall scores, subscores, and task type scores are tracked. Score-related statistics include the location and spread of scores, inter-correlations between scores, internal consistency reliability measures and SEM, and correlation with self-reported external measures.
2. **Test-taker profile:** The composition of the test-taker population is tracked over time, as demographic trends partially explain seasonal variability in test scores. Specifically tracked are the percentages of test takers by country, first language (L1), gender, age, intent in taking the test, and other background variables. In addition, many of the score statistics are tracked across major test-taker groups.
3. **Repeaters:** Repeaters are defined as those who take the test more than once within a 30-day window. The prevalence, demographic composition, and test performance of the repeater population are tracked. The performance of the repeater population is tracked with many of the same test score statistics identified above, with additional statistics that are specific to repeaters: testing location and distribution of scores from both the first and second test attempt, as well as their score change, and test–retest reliability (and SEM).
4. **Item analysis:** Item quality is quantified with four categories of item performance statistics—item difficulty, item discrimination, and item slowness (response time). Tracking these statistics allows for setting expectations about the item bank with respect to item performance, flagging items with extreme and/or inadequate performance, and detecting drift in measures of performance across time.

5. **Item exposure:** An important statistic in this category is the item exposure rate, which is calculated as the number of test administrations containing a certain item divided by the total number of test administrations. Tracking item exposure rates can help flag under- or over-exposure of items. Values of item exposure statistics result from the interaction of various factors, including the size of the item bank and the item selection algorithm.

The quality assurance of the DET is a combination of automatic processes and human review processes. The AQuAA system is used as the starting point for the human review process, and the human review process, in turn, helps AQuAA to evolve into a more powerful tool to detect assessment validity issues. Figure 25 depicts the human review process following every week's update of AQuAA; assessment experts meet to review all metrics for any potential anomalies. Automatic flags have also been implemented to indicate results that warrant closer attention. The assessment experts review any flags individually to determine whether it is a false alarm or further action is required. If the alarm is believed to be caused by a validity issue, follow-up actions are taken to determine the severity and urgency of the issue, fix the issue, and document the issue. Improvements are regularly made to the automatic flagging mechanisms to minimize false positives and false negatives, thereby improving AQuAA's functionality.

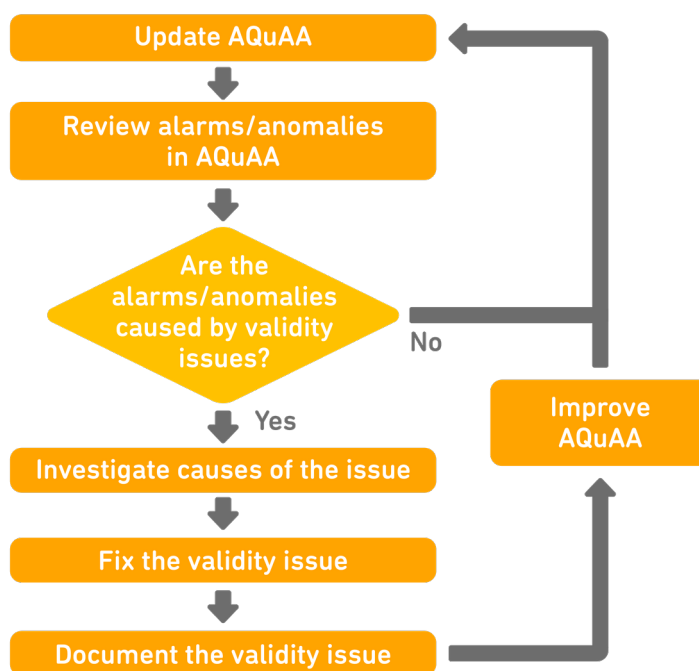


Figure 25. AQuAA Expert Review Process

While the primary purpose of AQuAA is to facilitate quality control, it also helps DET developers continually improve the exam. Insights drawn from AQuAA are used to direct the maintenance and improvement of other aspects of the assessment, such as item development. Additionally, the AQuAA system itself is designed to be flexible, with the possibility to modify and add metrics in order to adapt as the DET continues to evolve.

9.2 Proctoring Quality Assurance

In addition to psychometric quality assurance, DET proctoring quality is monitored regularly by assessment scientists and subject matter experts. A variety of tools and metrics are used to evaluate decision consistency among DET proctors and improve accuracy of decision-making in accordance with proctoring guidelines. These tools and metrics include:

Tools

- Monthly reports that track and evaluate proctors' decisions over the last 12 months

- Used to identify outlier proctors, who then undergo retraining with senior proctors
- Proctor calibration tool that evaluates proctors' decisions using the same test sessions automatically provides immediate feedback about the consensus answer (i.e., what the majority of proctors decide about a test session)
- Calibration meetings between senior and junior proctors, where senior proctors provide feedback on difficult proctoring sessions in a group setting
- Personal training sessions where more experienced proctors shadow less experienced proctors and provide feedback
- Weekly quizzes on proctoring process changes

Metrics

- Percentage of test sessions determined to have rule violations, cheating outcomes, identification issues, or technical errors across time
 - Changes in the test taker population (e.g., due to seasonal trends or market forces) can lead to differences in these trends
- Variability in proctors' decisions across all test sessions proctored, as well as on the same test sessions (e.g., see proctor calibration tool)
- Percentage of decisions overturned between proctors with more and less experience
- Outliers in the percentage of flagged test-taker behaviors, both in terms of under- and overuse (e.g., see monthly reports)
- Average number of minutes taken to proctor a test, controlling for decision type (i.e., rule violation, cheating, etc.) and accuracy of decision
- Test-taker score differences as a function of the type of test-taker behavior that is flagged

The tools and metrics used to monitor proctoring decisions help maintain high-quality, consistent proctoring by continually providing formative feedback to proctors and identifying proctors in need of additional training or re-calibration. Additionally, insights from proctoring quality assurance processes can lead to improvements in test administration and security. For instance, we can identify how and where test takers most often violate rules unintentionally and then modify instructions to minimize rule violation. Maintaining a high degree of consistency across proctors reinforces the security of the DET and ensures that test-taker sessions are reviewed equitably.

10 Conclusion

This version of the Technical Manual was produced on December 15, 2023. It provides a detailed overview of all facets of the Duolingo English Test and reports evidence for the DET's validity, reliability, and fairness as outlined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Updated versions of this document will be released to reflect changes to the test and new research findings.

11 Appendix

Table 15. Test-Taker L1s in Alphabetical Order (December 15, 2022 — December 14, 2023)

Afrikaans	Efik	Kannada	Marshallese	Spanish
Akan	English	Kanuri	Mende	Sundanese
Albanian	Estonian	Kashmiri	Minangkabau	Swahili
Amharic	Ewe	Kazakh	Mongolian	Swedish
Arabic	Farsi	Khmer	Mossi	Tagalog
Armenian	Fijian	Kikuyu	Nauru	Tajik
Assamese	Finnish	Kinyarwanda	Nepali	Tamil
Aymara	French	Kirundi	Northern Sotho	Tatar
Azerbaijani	Fulah	Kongo	Norwegian	Telugu
Bambara	Ga	Konkani	Oriya	Thai
Bashkir	Galician	Korean	Oromo	Tibetan
Basque	Ganda	Kosraean	Palauan	Tigrinya
Belarusian	Georgian	Kurdish	Pohnpeian	Tonga
Bemba	German	Kurukh	Polish	Tswana
Bengali	Greek	Lao	Portuguese	Turkish
Bikol	Guarani	Latvian	Punjabi	Turkmen
Bosnian	Gujarati	Lingala	Pushto	Twi
Bulgarian	Hausa	Lithuanian	Romanian	Uighur
Burmese	Hebrew	Luba-Lulua	Russian	Ukrainian
Catalan	Hiligaynon	Luo	Samoan	Umbundu
Cebuano	Hindi	Luxembourgish	Santali	Urdu
Chichewa (Nyanja)	Hungarian	Macedonian	Serbian	Uzbek
Chinese - Cantonese	Icelandic	Madurese	Sesotho	Vietnamese
Chinese - Mandarin	Igbo	Malagasy	Shona	Wolof
Chuvash	Iloko	Malay	Sindhi	Xhosa
Croatian	Indonesian	Malayalam	Sinhalese	Yapese
Czech	Italian	Maltese	Slovak	Yoruba
Danish	Japanese	Mandingo	Slovenian	Zhuang
Dutch	Javanese	Marathi	Somali	Zulu

Table 16. Test-Taker Country Origins in Alphabetical Order (December 15, 2022 — December 14, 2023)

Afghanistan	Djibouti	Lesotho	Saint Lucia
Albania	Dominica	Liberia	Saint Vincent and the Grenadines
Algeria	Dominican Republic	Libya	Samoa
American Samoa	Ecuador	Liechtenstein	San Marino
Andorra	Egypt	Lithuania	Sao Tome and Principe
Angola	El Salvador	Luxembourg	Saudi Arabia
Anguilla	Equatorial Guinea	Macao	Senegal
Antigua and Barbuda	Eritrea	Madagascar	Serbia
Argentina	Estonia	Malawi	Seychelles
Armenia	Eswatini	Malaysia	Sierra Leone
Aruba	Ethiopia	Maldives	Singapore
Australia	Faroe Islands	Mali	Sint Maarten (Dutch)
Austria	Fiji	Malta	Slovakia
Azerbaijan	Finland	Marshall Islands	Slovenia
Bahamas	France	Mauritania	Solomon Islands
Bahrain	Gabon	Mauritius	Somalia
Bangladesh	Gambia	Mexico	South Africa
Barbados	Georgia	Micronesia (Federated States)	South Sudan
Belarus	Germany	Monaco	Spain
Belgium	Ghana	Mongolia	Sri Lanka
Belize	Gibraltar	Montenegro	State of Palestine
Benin	Greece	Montserrat	Sudan
Bermuda	Greenland	Morocco	Suriname
Bhutan	Grenada	Mozambique	Sweden
Bolivarian Republic of Venezuela	Guatemala	Myanmar	Switzerland
Bolivia	Guinea	Namibia	Taiwan
Bosnia and Herzegovina	Guinea-Bissau	Nepal	Tajikistan
Botswana	Guyana	Netherlands	Thailand
Brazil	Haiti	New Zealand	Timor-Leste
Brunei Darussalam	Holy See	Nicaragua	Togo
Bulgaria	Honduras	Niger	Tonga
Burkina Faso	Hong Kong	Nigeria	Trinidad and Tobago
Burundi	Hungary	North Macedonia	Tunisia
Cabo Verde	Iceland	Norway	Turkey
Cambodia	India	Oman	Turkmenistan
Cameroon	Indonesia	Pakistan	Turks and Caicos Islands
Canada	Iraq	Palau	Uganda
Cayman Islands	Ireland	Panama	Ukraine
Central African Republic	Isle of Man	Papua New Guinea	United Arab Emirates
Chad	Israel	Paraguay	United Kingdom of Great Britain and Northern Ireland
Chile	Italy	Peru	United Republic of Tanzania
China	Jamaica	Philippines	United States of America
Colombia	Japan	Poland	Uruguay
Comoros	Jersey	Portugal	Uzbekistan
Congo	Jordan	Puerto Rico	Vanuatu
Congo (Democratic Republic)	Kazakhstan	Qatar	Viet Nam
Costa Rica	Kenya	Republic of Korea	Virgin Islands (British)
Croatia	Kiribati	Republic of Moldova	Virgin Islands (U.S.)
Cuba	Kuwait	Romania	Yemen
Cyprus	Kyrgyzstan	Russian Federation	Zambia
Czechia	Lao People's Democratic Republic	Rwanda	Zimbabwe
Côte d'Ivoire	Latvia	Saint Helena, Ascension and Tristan da Cunha	Åland Islands
Denmark	Lebanon	Saint Kitts and Nevis	

Table 17. Percentage Distribution Overall Score (May 01, 2023 — December 14, 2023)

Total	Percentage	Cumulative percentage
160	0.08%	100.00%
155	0.36%	99.92%
150	1.03%	99.56%
145	2.03%	98.53%
140	3.20%	96.50%
135	4.52%	93.30%
130	6.01%	88.78%
125	7.55%	82.77%
120	8.93%	75.22%
115	10.05%	66.29%
110	10.67%	56.24%
105	10.24%	45.57%
100	9.01%	35.34%
95	7.56%	26.33%
90	5.55%	18.77%
85	4.10%	13.22%
80	2.79%	9.12%
75	1.92%	6.34%
70	1.29%	4.41%
65	0.89%	3.12%
60	0.62%	2.23%
55	0.45%	1.61%
50	0.34%	1.16%
45	0.23%	0.82%
40	0.17%	0.59%
35	0.12%	0.42%
30	0.09%	0.30%
25	0.07%	0.21%
20	0.05%	0.15%
15	0.04%	0.10%
10	0.05%	0.05%

Table 18. Subscore Percentage Distributions (May 01, 2023 — December 14, 2023)

	Conversation	Literacy	Comprehension	Production
160	0.06%	0.44%	0.54%	0.04%
155	0.13%	0.81%	1.34%	0.07%
150	0.35%	1.50%	2.72%	0.16%
145	0.87%	2.51%	4.13%	0.35%
140	1.60%	3.59%	5.35%	0.67%
135	2.55%	4.88%	6.67%	1.09%
130	3.52%	6.21%	8.23%	1.68%
125	4.77%	7.46%	9.54%	2.27%
120	6.13%	8.71%	10.31%	3.16%
115	7.55%	9.72%	10.41%	4.18%
110	8.78%	9.93%	9.73%	5.42%
105	9.62%	9.57%	8.32%	6.56%
100	9.79%	8.48%	6.71%	7.85%
95	9.30%	6.98%	4.97%	8.82%
90	8.26%	5.37%	3.42%	9.18%
85	6.80%	3.94%	2.40%	9.15%
80	5.52%	2.83%	1.62%	8.46%
75	4.17%	2.00%	1.08%	7.29%
70	3.04%	1.39%	0.74%	5.98%
65	2.15%	1.00%	0.48%	4.59%
60	1.54%	0.72%	0.38%	3.51%
55	1.06%	0.52%	0.25%	2.57%
50	0.75%	0.37%	0.17%	1.96%
45	0.52%	0.29%	0.13%	1.42%
40	0.35%	0.21%	0.10%	1.03%
35	0.25%	0.15%	0.07%	0.76%
30	0.19%	0.12%	0.06%	0.55%
25	0.13%	0.08%	0.04%	0.40%
20	0.09%	0.06%	0.04%	0.31%
15	0.07%	0.04%	0.03%	0.22%
10	0.07%	0.11%	0.04%	0.28%

Table 19. Subscore Cumulative Percentage Distributions (May 01, 2023 — December 14, 2023)

	Conversation	Literacy	Comprehension	Production
160	100.00%	100.00%	100.00%	100.00%
155	99.94%	99.56%	99.46%	99.96%
150	99.82%	98.75%	98.12%	99.89%
145	99.47%	97.24%	95.40%	99.73%
140	98.60%	94.74%	91.27%	99.38%
135	97.01%	91.15%	85.93%	98.71%
130	94.45%	86.27%	79.26%	97.62%
125	90.93%	80.06%	71.03%	95.94%
120	86.16%	72.60%	61.49%	93.67%
115	80.03%	63.89%	51.18%	90.51%
110	72.48%	54.16%	40.77%	86.33%
105	63.69%	44.23%	31.04%	80.91%
100	54.08%	34.66%	22.72%	74.35%
95	44.28%	26.18%	16.01%	66.50%
90	34.98%	19.20%	11.04%	57.67%
85	26.72%	13.83%	7.62%	48.50%
80	19.92%	9.89%	5.22%	39.34%
75	14.39%	7.06%	3.60%	30.89%
70	10.22%	5.06%	2.52%	23.60%
65	7.18%	3.67%	1.78%	17.61%
60	5.04%	2.67%	1.30%	13.02%
55	3.49%	1.95%	0.93%	9.51%
50	2.43%	1.43%	0.68%	6.94%
45	1.68%	1.06%	0.51%	4.98%
40	1.16%	0.77%	0.38%	3.56%
35	0.81%	0.56%	0.28%	2.53%
30	0.56%	0.41%	0.21%	1.77%
25	0.37%	0.29%	0.15%	1.21%
20	0.24%	0.21%	0.11%	0.81%
15	0.15%	0.15%	0.07%	0.50%
10	0.07%	0.11%	0.04%	0.28%

References

- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. AERA.
- Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, 74(8), 1–36. <https://doi.org/10.18637/jss.v074.i08>
- Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42, 529–555. <https://doi.org/10.1111/j.1467-1770.1992.tb01043.x>
- Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, 55(6), 1–25. <http://www.jstatsoft.org/v55/i06/>
- Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology*, 10, 853. <https://doi.org/10.3389/fpsyg.2019.00853>
- Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5. <https://doi.org/10.3389/frai.2022.903077>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the yes/no vocabulary test: Some methodological issues in theory and practice. *Language Testing*, 18(3), 235–274. <https://doi.org/10.1177/026553220101800301>
- Belzak, W. C. (2023). The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educational Measurement: Issues and Practice*, 42(1), 24–33.
- Belzak, W. C., Naismith, B., & Burstein, J. (2023). Ensuring fairness of human- and AI-generated test items. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 701–707). Springer Nature Switzerland.
- Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/9781108686136>
- Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, 14(1), 14–31. <https://doi.org/10.1080/10904018.2000.10499033>
- Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, 112, 272–284. <https://doi.org/10.1121/1.1487837>
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106, 707–729. <https://doi.org/10.1016/j.cognition.2007.04.005>
- Buck, G. (2001). *Assessing listening*. Cambridge University Press.
- Burstein, J. (2023). *Responsible AI standards*. Duolingo. <https://go.duolingo.com/ResponsibleAI>
- Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test* (Duolingo Research Report DRR-22-01). Duolingo. <https://go.duolingo.com/ecosystem>
- Byram, M., & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: The globalisation of language education policy*. Multilingual Matters.
- Cardwell, R. L., Nydick, S. W., Lockwood, J., & von Davier, A. A. (2023). Practical considerations when building concordances between english tests. *Language Testing*.
- Carr, N. T. (2023). *Language background and its effect on performance on a digital age test*. <https://padlet.com/ncarr2/nathan-carr-s-conference-presentation-handouts-veqkqh7kklmk/wish/2569574650>
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. CRC press.
- Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge University Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – companion volume*. Council of Europe Publishing. <https://www.coe.int/lang-cefr>
- Daller, M., Müller, A., & Wang-Taylor, Y. (2021). The C-test as predictor of the academic success of international students. *International Journal of Bilingual Education and Bilingualism*, 24(10), 1502–1511. <https://doi.org/10.1080/13670050.2020.1747975>

- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16. <https://www.jstor.org/stable/44488664>
- Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48, 393–410. <https://doi.org/10.1111/0023-8333.00047>
- Duolingo English Test. (2021). *Duolingo English Test: Security, proctoring, and accommodations*. Duolingo. <https://duolingo-papers.s3.amazonaws.com/other/det-security-proctoring-whitepaper.pdf>
- Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23(3), 290–325. <https://doi.org/10.1191/0265532206lt330oa>
- Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39, 399–423. <https://doi.org/10.2307/3588487>
- Goodwin, S., Attali, Y., LaFlair, G. T., Runge, A., Park, Y., von Davier, A. A., & Yancey, K. P. (2023). *Duolingo English Test: Writing construct* (Duolingo Research Report DRR-22-03). Duolingo. <https://go.duolingo.com/scored-writing>
- Goodwin, S., & Naismith, B. (2023). *Assessing listening on the Duolingo English Test* (Duolingo Research Report DRR-23-02). Duolingo. <http://duolingo-testcenter.s3.amazonaws.com/media/resources/listening-whitepaper.pdf>
- Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, 12, 971–988. <https://doi.org/10.1214/aos/1176346715>
- Haberman, S. J., & Yao, L. (2015). Repeater analysis for combining information from different assessments. *Journal of Educational Measurement*, 52, 223–251. <https://doi.org/10.1111/jedm.12075>
- Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38, 201–223. <https://doi.org/10.2307/3588378>
- Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, 64(1), 215–238. <https://doi.org/10.3138/cmlr.64.1.215>
- Karimi, N. (2011). C-test and vocabulary knowledge. *Language Testing in Asia*, 1(4), 7. <https://doi.org/10.1186/2229-0443-1-4-7>
- Khodadady, E. (2014). Construct validity of C-tests: A factorial approach. *Journal of Language Teaching and Research*, 5. <https://doi.org/10.4304/jltr.5.6.1353-1362>
- Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, 14(1), 47–84. <https://doi.org/10.1177/026553229701400104>
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating methods and practices*. Springer-Verlag.
- Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12–24. <https://doi.org/10.1016/j.jslw.2016.10.003>
- LaFlair, G. T. (2020). *Duolingo English Test: Subscores* (Duolingo Research Report DRR-20-03). Duolingo. <https://duolingo-papers.s3.amazonaws.com/reports/subscore-whitepaper.pdf>
- LaFlair, G. T., Langenfeld, T., Baig, B., Horie, A. K., Attali, Y., & von Davier, A. A. (2022). Digital-first assessments: A security framework. *Journal of Computer Assisted Learning*. <https://doi.org/10.1111/jcal.12665>
- LaFlair, G. T., Runge, A., Attali, Y., Park, Y., Church, J., & Goodwin, S. (2023). *Interactive listening—The Duolingo English Test* (Duolingo Research Report DRR-23-01). Duolingo.
- Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability? *Journal of Research in Reading*, 15(2), 95–103. <https://doi.org/10.1111/j.1467-9817.1992.tb00025.x>
- Liao, M., Attali, Y., Lockwood, J. R., & Davier, A. A. von. (2022). Maintaining and monitoring quality of a continuously administered digital assessment. *Frontiers in Education*, 7. <https://doi.org/10.3389/educ.2022.857496>
- Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3), 294–309. <https://doi.org/10.1080/15434303.2018.1472265>
- McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021). Jump-starting item parameters for adaptive language tests. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 883–899. <https://doi.org/10.18653/v1/2021.emnlp-main.67>
- McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, 37(3), 389–411. <https://doi.org/10.1177/0265532219898380>
- Messick, S. (1989). Validity. In *Educational measurement, 3rd ed* (pp. 13–103). American Council on Education.
- Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 0–0). US Department of Education, Office of Educational Research; Improvement.
- Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (Vol. 1, pp. 211–232). EuroSLA.
- Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *EUROSLA monographs series 2* (pp. 57–78). European Second Language Association.

- Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (Vol. 52, pp. 83–98). Multilingual Matters.
- Molenaar, D., Cúri, M., & Bazán, J. L. (2022). Zero and one inflated item response theory models for bounded continuous data. *Journal of Educational and Behavioral Statistics*, 47(6), 693–735. <https://doi.org/10.3102/10769986221108455>
- Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45, 73–97. <https://doi.org/10.1111/j.1467-1770.1995.tb00963.x>
- Norris, J. (2018). *Developing C-tests for estimating proficiency in foreign language research*. Peter Lang. <https://doi.org/10.3726/b13235>
- Park, Y., Cardwell, R., Goodwin, S., Naismith, B., LaFlair, G. T., Loh, Kai-Ling, & Yancey, K. P. (2023). *Assessing speaking on the Duolingo English Test* (Duolingo Research Report DRR-23-03). <https://go.duolingo.com/listening>
- Park, Y., LaFlair, G. T., Attali, Y., Runge, A., & Goodwin, S. (2022). *Duolingo English Test: Interactive reading* (Duolingo Research Report DRR-22-02). Duolingo. <https://duolingo-papers.s3.amazonaws.com/other/mpr-whitepaper.pdf>
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Roche, T., & Harrington, M. (2014). Vocabulary knowledge and its relationship with EAP proficiency and academic achievement in an english-medium university in oman. In R. Al-Mahrooqi & A. Roscoe (Eds.), *Focusing on EFL reading: Theory and practice* (pp. 27–41). Cambridge Scholars Publishing.
- Rudis, B., & Kunimune, J. (2020). *Imago: Hacky world map GeoJSON based on the imago projection*. <https://git.rud.is/hrbrmstr/imago>
- Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, 45(1), 63–80. <http://www.jstor.org/stable/41307616>
- Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). Elsevier. <https://doi.org/10.1016/B0-12-369398-5/00444-8>
- Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8, 247–263. <https://doi.org/10.1162/tacl/a/00310>
- Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey* (No. 08-18). ETS. <https://www.ets.org/Media/Research/pdf/RM-08-18.pdf>
- Smith, E. E., & Kosslyn, S. M. (2007). *Cognitive psychology: Mind and brain*. Pearson/Prentice Hall.
- Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36, 139–152. <https://doi.org/10.1080/09571730802389975>
- Stowell, J. R., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research*, 42(2), 161–171. <https://doi.org/10.2190/EC.42.2.b>
- Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer (2nd edition)* (pp. 103–135). Routledge.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments. Synthesis report*.
- Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, 12(1), 54–73. <https://doi.org/10.1111/1473-4192.00024>
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer Science & Business Media.
- Wainer, H. (2000). *Computerized adaptive testing: A primer (2nd edition)*. Routledge.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361–375. <https://doi.org/j.1745-3984.1984.tb01040.x>
- Young, R. (2011). Interactional competence in language learning, teaching, and testing. In *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426–443).