

Duolingo English Test: Technical Manual



Duolingo Research Report
July 14, 2021 (37 pages)
<https://englishtest.duolingo.com/research>

Ramsey Cardwell*, Geoffrey T. LaFlair*, and Burr Settles*

Abstract

The Duolingo English Test Technical Manual provides an overview of the design, development, administration, and scoring of the Duolingo English Test. Furthermore, it reports on test-taker demographics and the statistical characteristics of the test. This is a living document and will be updated regularly (last update: July 14, 2021).

Contents

1	Introduction	3
2	Purpose	3
3	Accessibility	3
4	Test Administration and Security	5
4.1	Test Administration	5
4.2	Onboarding	6
4.3	Administration Rules	6
4.4	Proctoring and Reporting	7
5	Test-Taker Demographics	7
6	Item Type Descriptions	11

*Duolingo, Inc.

Corresponding author:

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

6.1	C-test	11
6.2	Yes/No Vocabulary (Text)	12
6.3	Yes/No Vocabulary (Audio)	13
6.4	Dictation	13
6.5	Elicited Imitation (Read-aloud)	13
6.6	Extended Writing	14
6.7	Extended Speaking	16
7	Development, Delivery, and Scoring	17
7.1	Item Development	17
7.2	CAT Delivery	19
7.3	CAT Item Scoring	20
7.4	Extended Speaking and Writing Tasks	20
8	Test Performance Statistics	21
8.1	Score Distributions	21
8.2	Reliability	22
8.3	Relationship with Other Tests	24
9	Quality Control	26
10	Conclusion	29
11	Appendix	30
	References	35

1 Introduction

The Duolingo English Test 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. It takes about one hour to complete the entire process of taking the test (i.e., onboarding, responding to tasks, and uploading responses). The test uses item types that provide maximal information about English language proficiency while being feasible to develop, administer, and score at scale. It is designed to be user-friendly in terms of onboarding, user interface, and item formats.

This technical manual provides an overview of the design of the Duolingo English Test. It contains a presentation of:

- the test’s accessibility, delivery, proctoring and security processes;
- the demographic information of the test-taker population;
- the test’s items, how they were created, and how they are delivered and scored;
- and the statistical characteristics of the test.

2 Purpose

Duolingo English Test scores are intended to be interpreted as reflecting test-taker English language ability and to be used in a variety of settings, including for post-secondary admissions decisions.

3 Accessibility

Broad accessibility is one of the central motivations for the development of the Duolingo English Test. While tests administered at test centers require resources which limit accessibility—the time to be at a physical testing center within certain hours on specific dates, travel to the test center, and considerable registration fees—the Duolingo English Test can be taken online, on demand, 24 hours a day and 365 days a year.

The AuthaGraph (Rudis & Kunimune, 2020) maps in Figure 1 show the concentration of test centers in the world (top panel) compared to internet penetration in the world (middle panel), and the concentration of Duolingo English Test test takers (bottom panel; for all tests administered since August 1, 2017). The top two panels of Figure 1 show how much more easily an internet-based test can be accessed than a test center (although

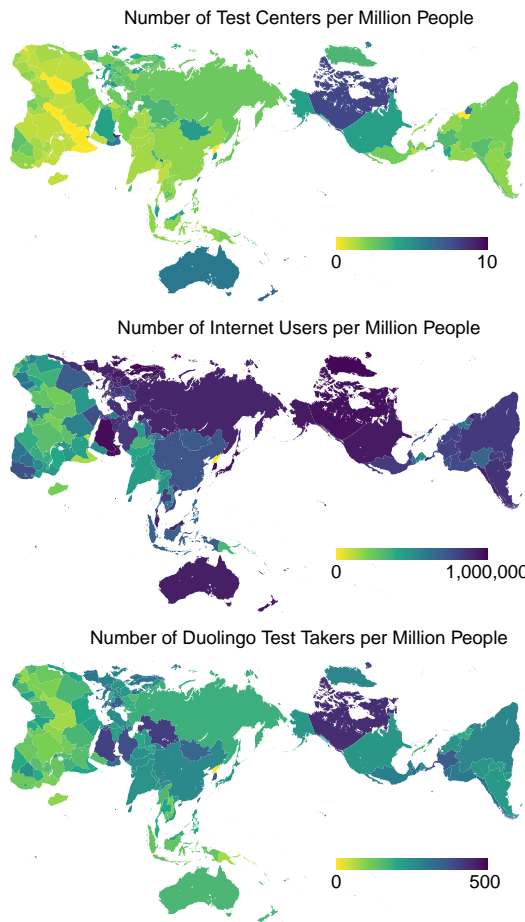


Figure 1. Heatmaps of test center accessibility as of 2018 (top), internet accessibility (middle), and concentration of Duolingo test takers (bottom)

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—not every internet user has access to a device that can run the Duolingo English Test, physical test centers can usually handle dozens of test-takers at once, and not all people need to take an English language proficiency assessment—it is still clear that the potential audience for the Duolingo English Test is orders of magnitude larger than those with convenient access to traditional test centers.

The map in the bottom panel shows that the Duolingo English Test is beginning to realize this potential, with people taking the Duolingo English Test from places with relatively

low concentrations of test centers, such as countries in South and Central America (Columbia, French Guiana, and Guatemala); in Central and East Asia (Kazakhstan and China); and Central and East Africa (Central African Republic and Zimbabwe). By delivering assessments on-demand, 24 hours a day, for US\$49, on any of the world's estimated 2 billion internet-connected computers, we argue that the Duolingo English Test holds the potential to be the most accessible, valid, and secure language assessment platform in the world.

4 Test Administration and Security

The Duolingo English Test is administered online, via the internet to test takers. The security of Duolingo English Test scores is ensured through a robust and secure onboarding process, 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 Duolingo English Test'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). The remainder of this section presents a summary of the information found in the [Security, Proctoring, and Accommodations](#) whitepaper.

4.1 Test Administration

Test takers are required to take the test alone in a quiet environment on a laptop or desktop computer 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. From the fall of 2020, test takers using Windows and macOS devices are required to take the test through the Duolingo English Test 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 Duolingo English Test website. For test takers using a Linux operating system, the Duolingo English Test can still be taken in the Chrome and Opera browsers worldwide or in the 360 and QQ browsers in China. The desktop app automatically prevents navigation away from the test and blocks tools such as spelling and grammar checkers. 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.

4.2 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.

4.3 Administration Rules

The behaviors that are prohibited during an administration of the Duolingo English Test are listed below. These rules require test takers to remain visible at all times to their cameras 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 or other non-malicious reasons, such as:

- 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 unless instructed
- 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

^{*}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

4.4 Proctoring and Reporting

After the test has been completed and uploaded, it undergoes a thorough proctoring review using human proctors with TESOL/applied linguistics expertise, which is supplemented by artificial intelligence to call proctors' attention to suspicious behavior. Each test session is reviewed in full by at least two independent proctors, with a third proctor brought in in the event of a disagreement between the first two proctors. 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.

5 Test-Taker Demographics

This section summarizes test-taker demographics based on all certified Duolingo English Test sessions between July 31, 2020 and July 13, 2021. 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 1 shows that 49.39% of Duolingo English Test test takers identified as female, 50.45% of test takers identified as male, and 0.16% selected "Other."

Table 1. Percentages of Test Taker Gender

Gender	Percentage
Female	49.39%
Male	50.45%
Other	0.16%
Total	100.00%

The gender distribution of test takers varies considerably across countries. Figure 2 depicts the proportion of reported gender identities for all countries with more than 100 test takers, ranging from 77% male to 67% female.

The median test-taker age is 22. Table 2 shows that 79% of Duolingo English Test test takers are between 16 and 30 years of age at the time of test administration.

Test takers are asked to report their L1s during the onboarding process. The most common first languages of Duolingo English Test test takers include Mandarin, Spanish, Arabic,

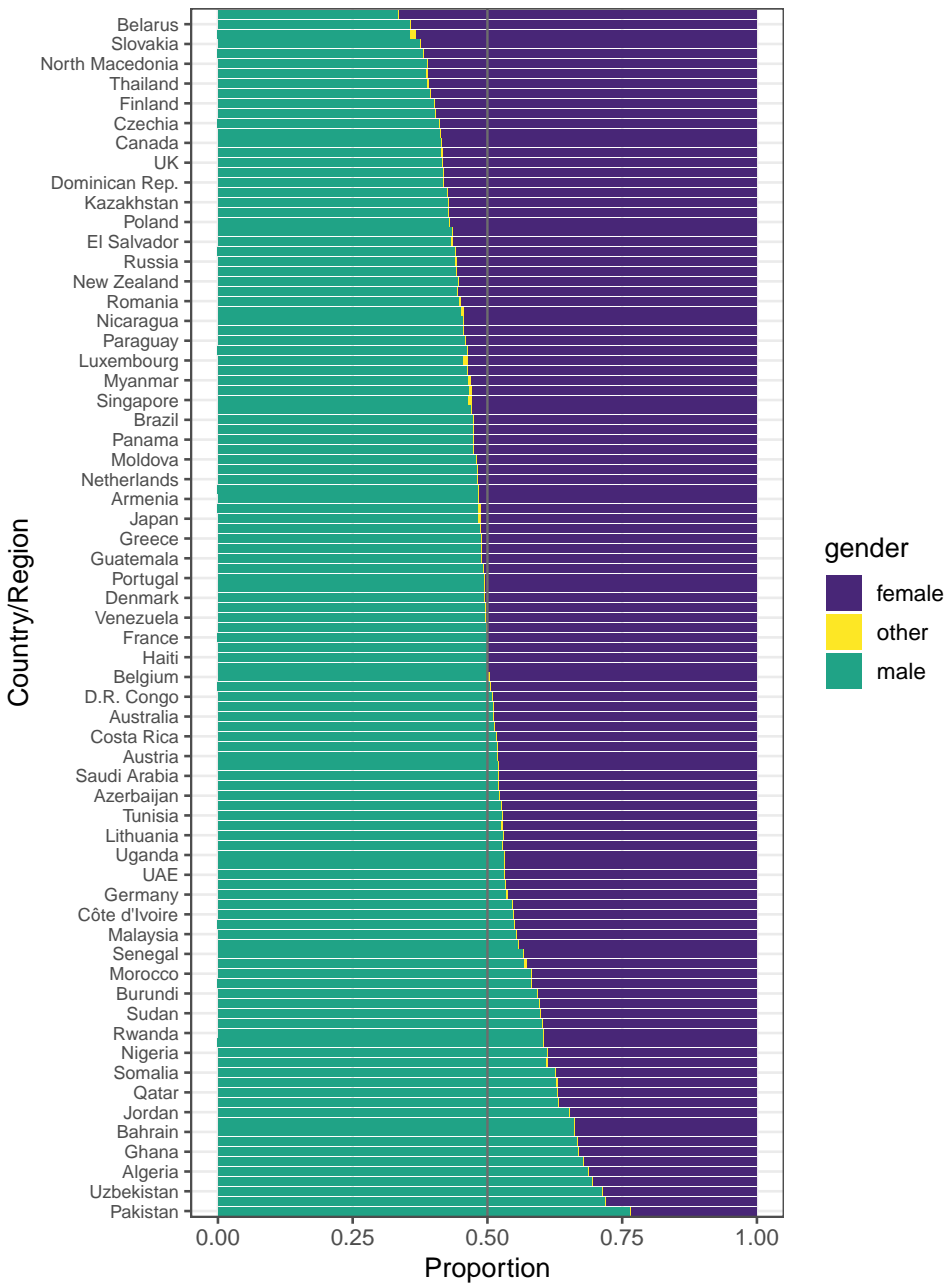


Figure 2. Proportion of reported gender identities for all countries and territories with >100 test takers (only every other country labeled)

Table 2. Percentages of Test Taker Age

Age	Percentage
< 16	3.10%
16 - 20	34.36%
21 - 25	29.76%
26 - 30	15.36%
31 - 40	13.33%
> 40	4.10%
Total	100.00%

English[‡], French, and Portuguese (see Table 3). There are 146 unique L1s represented by test takers of the Duolingo English Test, and the test has been administered to test takers from 210 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 3. Most Frequent Test-Taker L1s

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

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 82% 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 Duolingo English Test. Tables 4 and 5 display, for such test takers, the top 10 testing locations and the top 10 ID issuing countries, respectively.

Test takers are also asked to optionally indicate their intention for taking the Duolingo English Test, with the choice of applying to a school (secondary, undergraduate, or

[‡]55% of English-L1 test takers come from India and Canada

Table 4. Most Frequent Testing Locations for Test Takers Residing Outside Their Country of Origin

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

Table 5. Most Frequent ID Issuing Countries for Test Takers Residing Outside Their Country of Origin

Top ID origins
China
India
Romania
South Korea
Brazil
Italy
USA
Saudi Arabia
Colombia
Viet Nam

graduate) and job-related purposes. Table 6 presents the distribution of test-taker intentions.

Table 6. Test-Taker Intention

Intention	Percentage
Undergrad	40.38%
Grad	37.59%
Secondary School	3.85%
Work	1.38%
None of the Above	4.92%
(No Response)	11.89%

6 Item Type Descriptions

The Duolingo English Test has ten different graded item types, which collectively measure test-taker ability to use language skills required for literacy, conversation, comprehension, and production. Because the Duolingo English Test is a CAT, the difficulty of items adjusts as the computer updates its real-time estimate of test-taker language proficiency over the course of a test administration. Of the ten graded item types, five are in the computer-adaptive portion of the test. The CAT item types include c-test, audio yes/no vocabulary, visual yes/no vocabulary, dictation, and elicited imitation. During each administration, a test taker will see at minimum three of each CAT item type and at maximum seven of each CAT item type. The median rate of occurrence of each of the CAT item types across all administrations is six times per test administration. In addition to the five CAT item types, test takers respond to four writing prompts and four speaking prompts, which are not part of the computer-adaptive portion of the test. However, the writing and speaking prompts also vary in difficulty, and their selection is based on the CAT's estimate of test-taker ability. These items work together to measure test-taker English language proficiency in reading, writing, listening, and speaking. All Duolingo English Test item types are summarized in Table 7 below and described one by one in the subsequent sections.

Table 7. Summary of item formats on the Duolingo English Test

Item Type	Name for Test Takers	Type	Freq./Test	Skill(s)
1. C-test	Read and Complete	CAT	5–7	R,W
2. Yes/no (text)	Read and Select	CAT	5–7	L,R,W
3. Yes/no (audio)	Listen and Select	CAT	5–7	L,S
4. Dictation	Listen and Type	CAT	5–7	L,W
5. Elicited imitation	Read Aloud	CAT	5–7	R,S
6. Picture description	Write About the Photo	Perform	3	W
7. Text-independent	Read, Then Write	Perform	1	W
8. Picture description	Speak About the Photo	Perform	1	S
9. Text-independent	Read, Then Speak	Perform	1	S
10. Audio-independent	Listen, Then Speak	Perform	2	S
11. Speaking sample	Speaking Sample	Ungraded	1	S
12. Writing sample	Writing Sample	Ungraded	1	W

6.1 C-test

The c-test item type provides a measure of test-taker reading ability (Khodadady, 2014; Klein-Braley, 1997). In this task, the first and last sentences 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 (see Figure 3). Test takers need to rely on context and discourse information to reconstruct the damaged words (which span multiple lexical and morpho-syntactic categories). It has been shown that c-tests are significantly correlated with many other major language proficiency tests, and additionally are related to spelling skills (Khodadady, 2014).

Figure 3. Example C-test Item

6.2 Yes/No Vocabulary (Text)

This item type is a variant of the “yes/no” vocabulary test (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001). Such tests have been used to assess vocabulary knowledge at various CEFR levels (Milton, 2010). In the text variant of this item type (top panel of Figure 4), test takers are presented with a set of written English words mixed with pseudo-words that are designed to appear English-like, and must discriminate between them[§]. The text yes/no vocabulary item type has been shown to predict listening, reading, and writing abilities (Milton, Wade, & Hopkins, 2010; Staehr, 2008).

Traditional yes/no vocabulary tests simultaneously present a large set of mixed-difficulty stimuli (e.g., 60 words and 40 pseudo-words). The format is made computer-adaptive

[§]We use an LSTM recurrent neural network trained on the English dictionary to create realistic pseudo-words, filtering out any real words, acceptable regional spellings, and pseudo-words that orthographically or phonetically resemble real English words too closely.

by presenting multiple, smaller sets (items/testlets), each containing a few stimuli of the same difficulty (e.g., B1-level words with pseudo-words that should be B1-level if they existed; more on how this is done in Section 7.1).

6.3 Yes/No Vocabulary (Audio)

The audio variant of the yes/no vocabulary item type is conceptually equivalent to the text variant, except that the words and pseudo-words are presented auditorily. Test takers see an arrangement of speaker symbols labeled “word 1,” “word 2,” etc. (bottom panel of Figure 4) and must click on each symbol to hear an audio recording of the word. Test takers can replay the recordings as many times as desired. The audio yes/no vocabulary item type has been shown to predict listening and speaking abilities in particular (McLean, Stewart, & Batty, 2020; Milton, Wade, & Hopkins, 2010).

6.4 Dictation

In this exercise, test takers listen to a spoken sentence or short passage and then transcribe it using the computer keyboard[¶] (see Figure 5). Test takers have one minute to listen to the stimulus and transcribe what they heard. They can play the passage up to three times. This assesses test-taker ability to recognize individual words and to hold them in memory long enough to accurately reproduce them; both 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).

6.5 Elicited Imitation (Read-aloud)

The read-aloud variation of the elicited imitation task—example in Figure 6—is a measure of test-taker reading and speaking abilities (Jessop, Suzuki, & Tomita, 2007; Litman, Strik, & Lim, 2018; Vinther, 2002). It requires test takers to read, understand, and speak a sentence. Test takers respond to this task by using the computer’s microphone to record themselves speaking a written sentence. The goal of this task is to evaluate intelligible speech production, which is affected by segmental/phonemic and suprasegmental properties like intonation, rhythm, and stress (Anderson-Hsieh, Johnson, & Koehler, 1992; Derwing, Munro, & Wiebe, 1998; Field, 2005; Hahn, 2004). Furthermore, intelligibility is correlated with overall spoken comprehensibility (Derwing & Munro, 1997; Derwing, Munro, & Wiebe, 1998; Munro & Derwing, 1995), meaning that this item format can capture aspects of speaking proficiency. We use state-of-the-art

[¶]Autocomplete, spell-checking, and other assistive device features or plugins are detected and disabled.

The figure consists of two screenshots of a Duolingo vocabulary exercise. Each screenshot shows a timer at the top left, a progress bar, and a user profile picture at the top right. The instruction is "Select the real English words in this list".

Top Screenshot (1:32): The instruction is "Select the real English words in this list". The list of words is: wake, watched, somether, walks, waines, bookstore, thinking, washing, tooking, watching, thirteen, watch, givess, wants, ninete, nineteen, anss, answered. The word "bookstore" is highlighted in orange.

Bottom Screenshot (1:20): The instruction is "Select the real English words in this list". The list of words is: WORD 1, WORD 2, WORD 3, WORD 4, WORD 5, WORD 6, WORD 7, WORD 8, WORD 9. Each word has a speaker icon and a checkmark icon. The word "WORD 5" is highlighted in orange.

Figure 4. Example Yes/No Vocabulary Items

speech technologies to extract features of spoken language, such as acoustic and fluency features that predict these properties (in addition to basic automatic speech recognition), thus evaluating the general clarity of speech.

6.6 Extended Writing

The five extended writing tasks are measures of test-taker English writing abilities and include three written picture description tasks, one independent writing task based on a

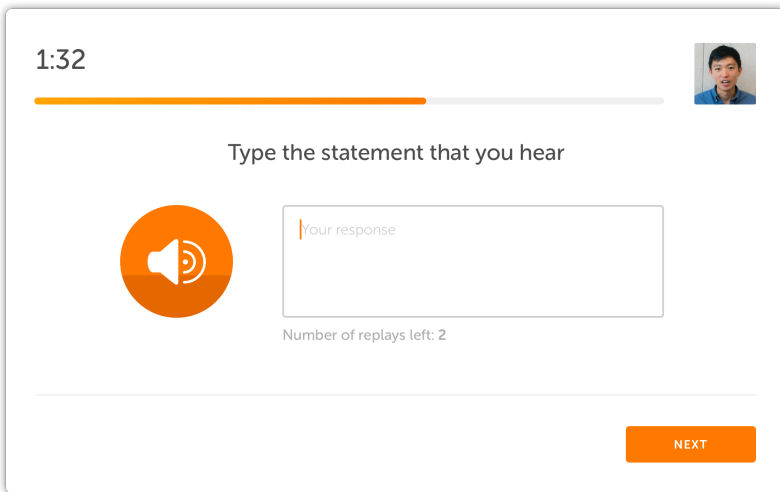


Figure 5. Example Dictation Item

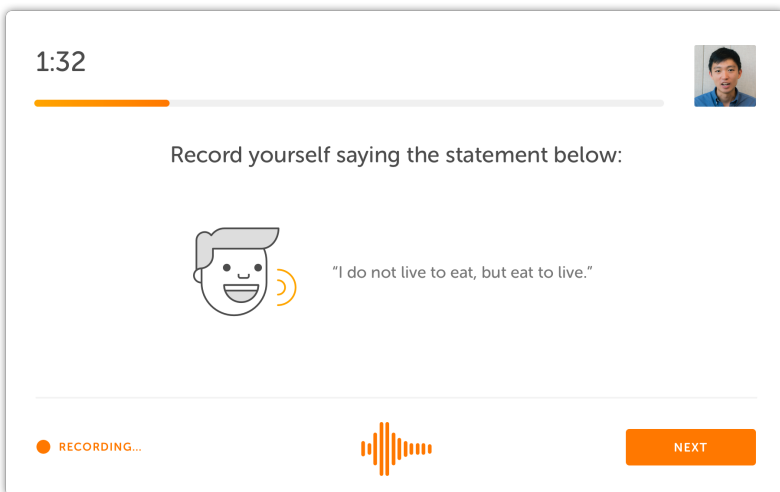



Figure 6. Example Elicited Imitation Item

written prompt, and one ungraded writing sample (see Figure 7). Each of the task types have items that are calibrated for high, intermediate, and low proficiency levels. The difficulty level of the tasks that test takers receive is conditional on their estimated ability in the CAT portion of the test. The stimuli in the picture description tasks were selected by people with graduate-level degrees in applied linguistics. They are designed to give test takers the opportunity to display their full range of written language abilities. The


independent tasks ask test takers to describe something, recount an experience, or argue a point of view, which require test takers to demonstrate more discursive knowledge of writing in addition to language knowledge (Cushing-Weigle, 2002).

1:32 

Write one or more sentences that describe the image

Your response

NEXT

1:32 

Respond to the questions in at least 50 words

"I do not live to eat, but eat to live."

Consider the subtleness of the sea; how its most dreaded creatures glide under water, unapparent for the most part, and treacherously hidden beneath the loveliest tints of azure

Words: 98

NEXT

Figure 7. Example Writing Items

6.7 Extended Speaking

The extended speaking tasks are measures of test-taker English speaking abilities. After the CAT portion of the test, test takers respond to five speaking prompts: one picture

description task and three independent speaking tasks—two with a written prompt and one with an aural prompt (see Figure 8)—as well as an ungraded speaking sample. Similar to the writing tasks, these are drawn from different levels of difficulty conditional on the estimated ability level of the test taker at the end of the CAT portion. All of 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, text structure) and functional elements of their pragmatic language knowledge (e.g., ideational knowledge) (Bachman & Palmer, 1996).

7 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.

7.1 Item Development

In order to create enough items of each type at varying levels of difficulty, the Duolingo English Test item pool is automatically generated. 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. The resulting data matrix is therefore sparse, making it infeasible to estimate \hat{b}_i (item difficulty) empirically. Furthermore, it is not scalable to have each item manually reviewed by CEFR-trained experts. Instead, we employ statistical machine learning (ML) and natural language processing (NLP) to automatically project items onto the Duolingo English Test scale. Each of the items has an estimated level of difficulty on a continuous scale between zero and ten. These levels were assigned to the items based on one of two ML/NLP models—a vocabulary model and a passage model—that were trained as part of the test development process. The vocabulary model was used to estimate the item difficulty of the yes/no vocabulary tasks. The passage model was used to estimate the difficulty of the other item types. The two models are used to predict \hat{b}_i values for the different CAT item types as a function of various psycholinguistically-motivated predictor variables, including:

- syntactic variables (dependency parse tree depth, number and direction of dependencies, verb tenses, sentence length, etc.);
- morphological variables (character-level language model statistics, word length in characters and syllables, etc.);
- lexical variables (word-level language model statistics).

The variables were processed using various NLP pipelines described in greater detail in Settles, LaFlair, & Hagiwara (2020).

The figure displays three sequential screenshots of a Duolingo speaking practice interface. Each screenshot features a 1:32 timer at the top left and a user profile picture at the top right. A progress bar is located below the timer.

- Top Screenshot:** The instruction is "Describe aloud the image below". Below the text is a photograph of a family (a woman, a man, and two children) on a beach. At the bottom, there is a "RECORDING..." indicator, a waveform icon, and a "NEXT" button.
- Middle Screenshot:** The instruction is "Speak your answer to the question below". Below this is a text box containing the prompt "Talk about a hobby or activity that you enjoy." followed by a bulleted list of questions: "What is it?", "How long have you been doing it?", "Who do you do it with?", and "Why is it important to you?". At the bottom, there is a "RECORDING..." indicator, a waveform icon, and a "START" button.
- Bottom Screenshot:** The instruction is "Speak the answer to the question you hear". Below the text is a large orange speaker icon. Underneath the speaker icon, it says "Number of replays left: 2". At the bottom, there is a "RECORDING..." indicator, a waveform icon, and a "NEXT" button.

Figure 8. Example Speaking Items

7.2 CAT Delivery

Once items are generated, calibrated (\hat{b}_i estimates are made), and placed in the item pool, the Duolingo English Test 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 θ 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 (θ) is unknown before test administration. As a result, an iterative, adaptive algorithm is required. First, the algorithm makes a provisional estimate of $\hat{\theta}_t$ based on responses to a set of items at the beginning of the test — increasing in difficulty — to time point t . Then the difficulty of the next item is selected as a function of the current estimate: $b_{t+1} = f(\hat{\theta}_t)$. Once that item is scored, the process repeats until a stopping criterion is satisfied.

The Duolingo English Test uses maximum-likelihood estimation (MLE) to estimate $\hat{\theta}_t$ and select the next item. The MLE optimization seeks to find the $\hat{\theta}_t$ that is most probable given a test taker's item-level scores. This approach, combined with concise and predictive item formats, helps to minimize test administration time significantly.

Duolingo English Test sessions are variable-length, meaning that exam duration and number of items vary across administrations. The iterative, adaptive procedure continues until either the variance of the $\hat{\theta}_t$ estimate drops below a certain threshold, or the test exceeds a maximum length in terms of minutes or items. Most tests are less than 45 minutes long (including speaking and writing; excluding onboarding and uploading), and the median test consists of 30 computer-adaptive (and eight extended response) items with over 200 measurements¹.

Once the algorithm converges, the final reported score is not the provisional MLE point-estimate used during CAT administration. Rather, for each CAT item type, the probability is computed for each possible $\theta \in [0, 10]$ and normalized into a posterior distribution in order to create a weighted average score. These weighted average scores of each CAT item type are then used with the scores of the speaking and writing tasks to compute a total score and the four subscores.

¹For example, each word (or pseudo-word) in the vocabulary format, and each damaged word in the c-test passage format, is considered a separate “measurement” (or sub-item).

7.3 CAT Item Scoring

All test items are graded automatically via statistical procedures appropriate for the item type. For example, the yes/no vocabulary format (see Figure 4) is traditionally scored using the sensitivity index d' , a measure of separation between signal (word) and noise (pseudo-word) distributions from signal detection theory (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001; Zimmerman, Broder, Shaughnessy, & Underwood, 1977). However, traditional yes/no tests assume that all stimuli are given at once, which is not the case in the Duolingo English Test's adaptive variant. This index, d' , is easily computed for fewer stimuli, and it has a probabilistic interpretation under receiver-operator characteristics (ROC) analysis (Fawcett, 2006). That is, d' is calculated for each yes/no item response and converted into a score g_i , which can be interpreted as "the test taker can accurately discriminate between English words and pseudo-words at this difficulty level with probability g_i ," where $g_i \in [0, 1]$.

The responses to the dictation, elicited imitation, and c-test tasks are aligned against an expected reference text, and similarities and differences in the alignment are evaluated. The output of the comparison is used in a (binary) logistic regression model** to provide the probabilistic grade g_i .

7.4 Extended Speaking and Writing Tasks

The writing and speaking tasks are scored by automated scoring models developed by ML and NLP experts at Duolingo. There is a separate scoring model for each of the three speaking task types and two writing task types. The speaking and writing scoring systems evaluate each item response based on the following categories of features:

- Grammatical accuracy
- Grammatical complexity
- Lexical sophistication
- Lexical diversity
- Task relevance
- Length
- Fluency & acoustic features (speaking)

Numerical values on each feature are computed for each extended speaking and writing task response, and the task-level score is computed as a weighted sum of the features. Scores on the writing and speaking tasks then contribute to a test taker's final overall score and subscores; writing task scores contribute to the subscores Production and Literacy,

**the weights for this model were trained on aggregate human judgments of correctness and intelligibility on tens of thousands of test items. The correlation between model predictions and human judgments is $r = 0.75$ ($p < 0.001$).

while the speaking task scores contribute to Production and Conversation. One way to evaluate the validity of the automated scoring procedures is to examine the correlations of automated scores with independent measures of the same construct. Table 8 summarizes the correlations of automated writing scores with TOEFL and IELTS writing subscores, and automated speaking scores with TOEFL and IELTS speaking subscores. These correlations are based on Duolingo English Test takers' self-reported results from the TOEFL (n = 2,746) and IELTS (n = 12,797) and weighted averages of item-level scores on writing and speaking tasks. The *Pearson Cor.* column contains the raw Pearson correlation coefficients, while the *Corrected Cor.* column presents the correlations after correcting for restriction of range, given that higher-ability test takers are more likely to report TOEFL/IELTS results.

The moderate-to-strong correlations presented in Table 8 are comparable to those reported between TOEFL and IELTS subscores (Educational Testing Service, 2010) and suggest that the Duolingo English Test automated writing and speaking scores measure a construct similar to that of the TOEFL and IELTS writing and speaking subscores. It should be noted that the TOEFL and IELTS scores used in these correlations were from tests taken up to 90 days before the Duolingo English Test. This gap between test administrations, as well as the self-reported nature of the TOEFL and IELTS scores, introduces error into the data, making the resulting correlations lower than they likely would be if data were collected under controlled conditions.

Table 8. Correlations of Duolingo English Test automated speaking and writing grades with relevant subscores of other tests

Automated grade \diamond Criterion score	Pearson Cor.	Corrected Cor.
Writing & TOEFL writing	0.53	0.59
Writing & IELTS writing	0.42	0.47
Speaking & TOEFL speaking	0.60	0.64
Speaking & IELTS speaking	0.54	0.59

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 total score and subscores. The analyses of the subscores were conducted on data from tests that were administered between July 31, 2020 and July 13, 2021.

8.1 Score Distributions

Figure 9 shows the distribution of scores for the total 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 total score using three different visualization techniques. The left panels show a box plot 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 plots in Figure 9 show some negative skew, which is reflected in the descriptive statistics in Table 9. The total score mean and the median test score are 107.4 and 110 respectively, and the interquartile range is 25. Tables 14–16 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 9.

Table 9. Descriptive Statistics for Total and Subscores (n = 99,415)

Score	Mean	SD	25th Percentile	Median	75th Percentile
Comprehension	116.03	19.99	105	120	130
Conversation	98.54	22.01	85	100	115
Literacy	107.45	20.06	95	110	120
Production	85.18	22.61	70	85	100
Total	107.40	19.29	95	110	120

8.2 Reliability

The reliability of the Duolingo English Test 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 for each of these measures come from a subset of the 301,196 certified tests administered between July 31, 2020 and July 13, 2021. 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 homogenous, lower-ability subpopulation than the full testing population. Unless addressed, this reduction in 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 heterogeneity across test takers in true proficiency growth. Unless addressed, this excess heterogeneity also will tend to artificially reduce estimated reliabilities based on repeaters.

In order to address the challenges inherent to test–retest reliability, the analysis was conducted on a sample of repeaters who took the Duolingo English Test twice within 15 days using a weighting and model-averaging procedure to mitigate the impacts of both

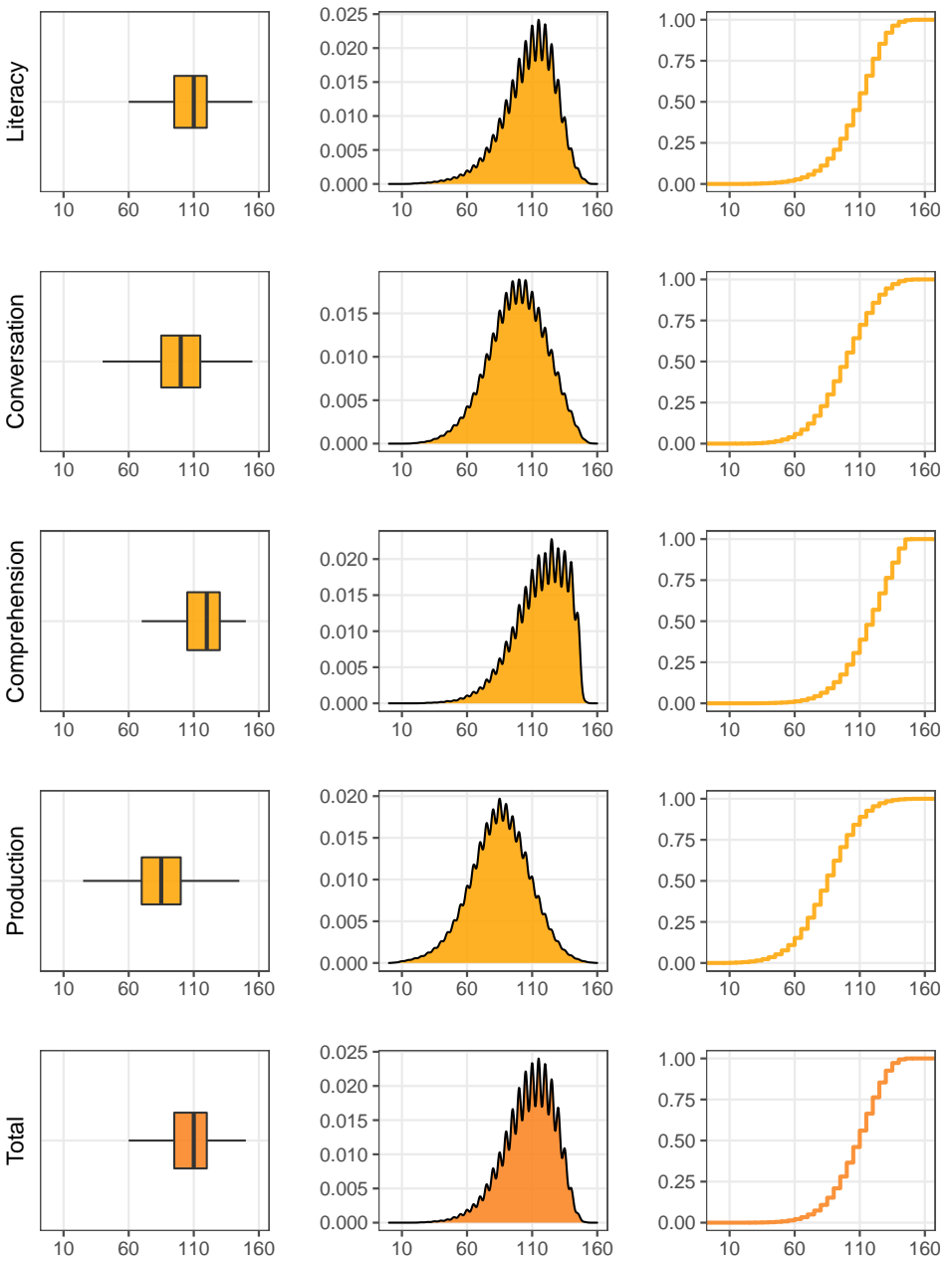


Figure 9. Boxplots (left), Density Plots (middle), and Empirical Cumulative Distribution Plots (right) of the Total Score and Subscores.

demographic non-representativeness and learning heterogeneity. The sample was first divided into 16 subsets corresponding to the number of days between the first and second test attempts. The Minimum Discriminant Information Adjustment (MDIA, [Haberman, 1984](#)) was used to weight each subset to match test-taker attributes of *all* first-time test takers. Specifically, MDIA finds weights for the subset so that the weighted subset matches all first-time test takers with respect to country, first language, age, gender, Windows vs MacOS, TOEFL overall scores, IELTS overall scores, and the means and variances of the Duolingo English Test scores on the first attempt. Weighting in this manner mitigates the potential biasing effects of repeater self-selection on estimated test–retest reliabilities ([Haberman & Yao, 2015](#)). A weighted test–retest correlation was calculated separately on each subset for the total score and all four subscores. Four polynomial regression models (linear, quadratic, cubic, and quartic) were then fit to each set of 16 correlations. The values in [Table 10](#) are the reliabilities for the Duolingo English Test overall score and subscores, estimated as the weighted (as a function of model BIC) average of the four model-based predictions of the test–retest correlation that would be observed if all test takers retook the test within 24 hours.

The coefficients for the subscores and the total score in [Table 10](#) show that the subscore reliability coefficients are slightly lower than the total score reliability. This is expected because they are calculated on a smaller number of items. The SEM is estimated using [Equation \(1\)](#), where x is a total score or subscore, SD is the standard deviation of the total score or subscore, and $\hat{\rho}_{XX'}$ is the test–retest reliability coefficient of the total score or subscore. When the results are rounded to the nearest 5-point increment—the Duolingo English Test score scale increases in 5-point increments—the range for the SEM is $+/- 5$, or one score unit, for the total score and the Literacy subscore, and $+/- 10$ for the remaining subscores.

$$SEM_x = SD_x * \sqrt{1 - \hat{\rho}_{XX'}} \quad (1)$$

Table 10. Test-Retest and SEM Estimates

Score	Test–Retest	SEM	SEM (rounded)
Literacy	0.88	6.95	5
Conversation	0.86	8.23	10
Comprehension	0.86	7.48	5
Production	0.86	8.46	10
Total	0.90	6.10	5

8.3 Relationship with Other Tests

In 2019, correlational and concordance studies were conducted to examine the relationship between Duolingo English Test scores and scores from TOEFL iBT and

IELTS. The data for these studies comprise self-reported TOEFL and IELTS test scores from Duolingo English Test test takers.

Correlation

Pearson's correlation coefficients were estimated to evaluate the relationship between the Duolingo English Test and the TOEFL iBT and IELTS. Both correlation coefficients revealed strong, positive relationships of Duolingo English Test scores with TOEFL iBT scores ($r = 0.77$; $n = 2,319$) and with IELTS scores ($r = 0.78$; $n = 991$). These relationships are visualized in Figure 10. The left panel shows the relationship between the Duolingo English Test and TOEFL iBT, and the right panel shows the relationship between the Duolingo English Test and IELTS.

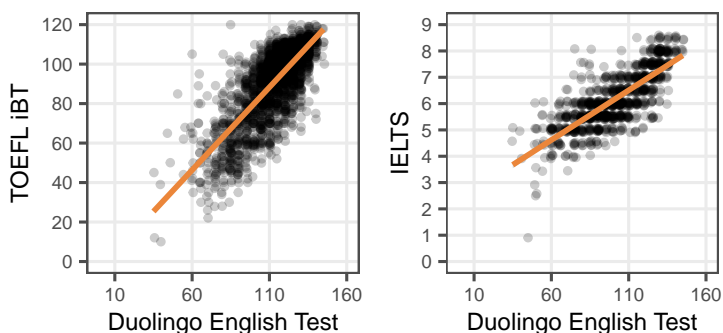


Figure 10. Relationship between Test Scores

Concordance

The same data from the correlation study were used to create concordance tables for Duolingo English Test score users. Two types of equating were compared: equipercentile (Kolen & Brennan, 2014) and kernel equating (Davies, Holland, & Thayer, 2003). Within each equating type two methods were evaluated: 1) loglinear pre-smoothing that preserved the first and second moments as well as the bivariate relationship between the test scores and 2) loglinear pre-smoothing that preserved the first, second, and third moments as well as the bivariate relationship between the test scores. The equating study was conducted using the *equate* (Albano, 2016) and *kequate* (Andersson, Bränberg, & Wiberg, 2013) packages in R (R Core Team, 2018).

The equating procedure that was selected to create the concordance tables was the one that minimized the mean standard error of equating. Table 11 shows that this was the kernel equating that preserved the first two moments and the bivariate score relationship. The conditional error across the Duolingo English Test score range is very small for kernel equating as well. As can be seen in Figure 10, data on TOEFL and IELTS scores are

Table 11. Standard Error of Equating (SEE) Summary

Method	TOEFL		IELTS		
	Mean	SD	Method	Mean	SD
EQP 2	2.20	2.76	EQP 2	0.73	1.68
EQP 3	0.84	1.91	EQP 3	0.87	1.97
KER 2	0.45	0.34	KER 2	0.05	0.02
KER 3	0.81	0.70	KER 3	0.06	0.04

extremely sparse for test takers with Duolingo English Test scores below 65, and equating error is thus larger in the 10–60 range of the Duolingo English Test score scale.

For score points between 65 and 160, the conditional standard error of equating (SEE) of the KER 2 method is between 0.03 and 0.88 for TOEFL, and between 0.02 and 0.05 for IELTS. 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 Duolingo English Test (31 possible score values), meaning test takers with the same Duolingo English Test 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 Duolingo English Test score can correspond to multiple TOEFL scores, which inflates the SEE. The concordance tables can be found on the Duolingo English Test scores page (<https://englishtest.duolingo.com/scores>).

9 Quality Control

The unprecedented flexibility, complexity, and high-stakes nature of the Duolingo English Test poses quality assurance challenges. In order to ensure the test is of high quality at all times, it is necessary to continuously monitor the key summary statistics of the test and be able to react promptly when needed. The Duolingo English Test therefore utilizes a custom-built quality assurance system, Analytics for Quality Assurance in Assessment (AQuAA), 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 Duolingo English Test’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 11, test data such as test-taker demographics, item response durations, and item scores are automatically imported into AQuAA from Duolingo English Test 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.

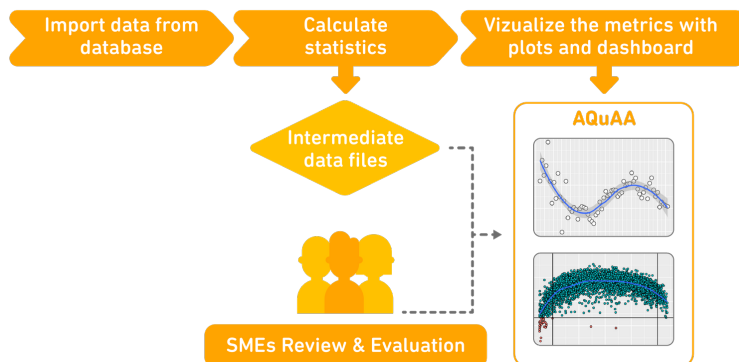


Figure 11. Duolingo English Test Quality Control Procedures

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

1. **Scores:** Overall scores, sub-scores, and item type scores are tracked. Score-related statistics include the location and spread of scores, inter-correlations between scores, internal consistency reliability measures and standard error of measurement (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, 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 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 Duolingo English Test 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 12 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.

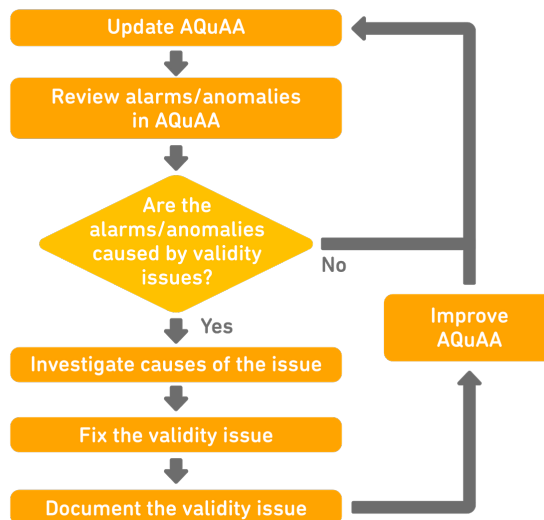


Figure 12. AQuAA Expert Review Process

While the primary purpose of AQuAA is to facilitate quality control, it also helps Duolingo English Test 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 Duolingo English Test continues to evolve.

10 Conclusion

The research reported here illustrates evidence for the validity of the interpretations and uses of the Duolingo English Test. Updated versions of this document will be released as we continue our research.

11 Appendix

Table 12. Test-Taker L1s in Alphabetical Order

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

Table 13. Test-Taker Country Origins in Alphabetical Order

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

Table 14. Percentage Distribution Total Score

Total	Percentage	Cumulative percentage
150	0.04%	100.00%
145	0.57%	99.96%
140	2.16%	99.38%
135	4.65%	97.22%
130	7.21%	92.57%
125	9.03%	85.36%
120	9.95%	76.33%
115	10.29%	66.38%
110	10.03%	56.08%
105	9.50%	46.05%
100	8.47%	36.54%
95	7.14%	28.07%
90	5.72%	20.94%
85	4.41%	15.21%
80	3.41%	10.81%
75	2.40%	7.40%
70	1.65%	4.99%
65	1.16%	3.34%
60	0.81%	2.19%
55	0.54%	1.38%
50	0.35%	0.84%
45	0.21%	0.49%
40	0.13%	0.29%
35	0.09%	0.16%
30	0.04%	0.07%
25	0.02%	0.03%
20	0.00%	0.00%

Table 15. Subscore Percentage Distributions

	Conversation	Literacy	Comprehension	Production
160	0.00%	0.00%	0.00%	0.00%
155	0.00%	0.00%	0.00%	0.03%
150	0.12%	0.30%	0.12%	0.09%
145	5.55%	1.01%	5.55%	0.22%
140	8.63%	2.27%	8.63%	0.42%
135	9.30%	4.31%	9.30%	0.72%
130	9.43%	6.76%	9.43%	1.20%
125	10.01%	9.06%	10.01%	1.85%
120	9.16%	10.39%	9.16%	2.74%
115	9.03%	10.65%	9.03%	3.75%
110	8.13%	10.27%	8.13%	4.83%
105	7.11%	9.29%	7.11%	6.22%
100	5.99%	8.11%	5.99%	7.37%
95	4.63%	6.73%	4.63%	8.24%
90	3.76%	5.43%	3.76%	8.97%
85	2.72%	4.23%	2.72%	9.26%
80	2.05%	3.17%	2.05%	8.65%
75	1.45%	2.34%	1.45%	7.86%
70	0.97%	1.67%	0.97%	6.80%
65	0.70%	1.20%	0.70%	5.51%
60	0.47%	0.87%	0.47%	4.33%
55	0.30%	0.61%	0.30%	3.27%
50	0.18%	0.43%	0.18%	2.38%
45	0.14%	0.31%	0.14%	1.72%
40	0.08%	0.22%	0.08%	1.21%
35	0.05%	0.15%	0.05%	0.83%
30	0.03%	0.09%	0.03%	0.58%
25	0.01%	0.06%	0.01%	0.38%
20	0.00%	0.04%	0.00%	0.27%
15	0.00%	0.00%	0.00%	0.17%
10	0.00%	0.00%	0.00%	0.10%
5	0.00%	0.00%	0.00%	0.02%

Table 16. Subscore Cumulative Percentage Distributions

	Conversation	Literacy	Comprehension	Production
160	100.00%	100.00%	100.00%	100.00%
155	100.00%	100.00%	100.00%	100.00%
150	100.00%	100.00%	100.00%	99.97%
145	99.88%	99.69%	99.88%	99.88%
140	94.33%	98.68%	94.33%	99.66%
135	85.70%	96.41%	85.70%	99.23%
130	76.40%	92.10%	76.40%	98.51%
125	66.97%	85.34%	66.97%	97.31%
120	56.96%	76.28%	56.96%	95.46%
115	47.80%	65.90%	47.80%	92.72%
110	38.77%	55.25%	38.77%	88.98%
105	30.64%	44.98%	30.64%	84.14%
100	23.52%	35.68%	23.52%	77.93%
95	17.54%	27.57%	17.54%	70.55%
90	12.91%	20.84%	12.91%	62.31%
85	9.15%	15.40%	9.15%	53.34%
80	6.42%	11.17%	6.42%	44.08%
75	4.37%	8.00%	4.37%	35.43%
70	2.92%	5.65%	2.92%	27.57%
65	1.96%	3.99%	1.96%	20.77%
60	1.26%	2.78%	1.26%	15.26%
55	0.79%	1.92%	0.79%	10.93%
50	0.48%	1.31%	0.48%	7.67%
45	0.31%	0.88%	0.31%	5.28%
40	0.17%	0.56%	0.17%	3.57%
35	0.09%	0.34%	0.09%	2.36%
30	0.04%	0.19%	0.04%	1.52%
25	0.01%	0.10%	0.01%	0.94%
20	0.00%	0.04%	0.00%	0.56%
15	0.00%	0.00%	0.00%	0.30%
10	0.00%	0.00%	0.00%	0.13%
5	0.00%	0.00%	0.00%	0.02%

References

- 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.
- 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. Retrieved from <http://www.jstatsoft.org/v55/i06/>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: 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.
- Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, 112, 272–284.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106, 707–729.
- Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
- Cushing-Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Davies, A. A. von, Holland, P. W., & Thayer, D. T. (2003). *The kernel method of test equating*. NY: Springer Science & Business Media.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16.
- Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48, 393–410.
- Educational Testing Service. (2010). *Linking TOEFL iBT scores to IELTS scores—A research report*. Educational Testing Service Princeton, NJ.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39, 399–423.
- 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.
- Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, 64(1), 215–238.
- Khodadady, E. (2014). Construct validity of C-tests: A factorial approach. *Journal of Language Teaching and Research*, 5.
- Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, 14(1), 47–84.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. NY: Springer Science & Business Media.
- Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 1–16.
- McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing, OnlineFirst*. <https://doi.org/10.1177/0265532219898380>
- 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* (pp. 211–232). Eurosla.
- 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). Bristol: Multilingual Matters.
- Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45, 73–97.
- R Core Team. (2018). *R: A language and environment for statistical computing*. Retrieved from <https://www.R-project.org/>
- Rudis, B., & Kunimune, J. (2020). *Imago: Hacky world map GeoJSON based on the imago projection*. Retrieved from <https://git.rud.is/hrbrmstr/imago>
- Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. New York, NY: Elsevier.

- 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
- Smith, E. E., & Kosslyn, S. M. (2007). *Cognitive psychology: Mind and brain*. Upper Saddle River, NJ: Pearson/Prentice Hall.
- Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36, 139–152.
- Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer*. Routledge.
- Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, 12(1), 54–73.
- Wainer, H. (2000). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Routledge.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361–375.
- Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, 1(1), 5–31.