

---

# Duolingo English Test: Technical Manual



Duolingo Research Report  
July 15, 2019 (25 pages)  
<https://englishtest.duolingo.com/research>

**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 15, 2019).

**Contents**

<b>1</b>	<b>Duolingo English Test</b>	<b>3</b>
<b>2</b>	<b>Accessibility</b>	<b>3</b>
<b>3</b>	<b>Test Administration and Security</b>	<b>4</b>
3.1	Test Administration . . . . .	4
3.2	Onboarding . . . . .	4
3.3	Administration Rules . . . . .	4
3.4	Proctoring & Reporting . . . . .	5
<b>4</b>	<b>Test Taker Demographics</b>	<b>5</b>
<b>5</b>	<b>Item Description</b>	<b>6</b>
5.1	C-test . . . . .	6
5.2	Yes/No Vocabulary . . . . .	7
5.3	Dictation . . . . .	7
5.4	Elicited Imitation (Read-aloud) . . . . .	8

---

\*Duolingo, Inc.

**Corresponding author:**  
Geoffrey T. LaFlair, PhD  
Duolingo, Inc. 5900 Penn Ave, Pittsburgh, PA 15206, USA  
Email: [englishtest-research@duolingo.com](mailto:englishtest-research@duolingo.com)

5.5	Extended Speaking . . . . .	10
5.6	Extended Writing . . . . .	10
<b>6</b>	<b>Development, Delivery, &amp; Scoring</b>	<b>10</b>
6.1	Item Development . . . . .	10
6.2	CAT Delivery . . . . .	13
6.3	Item Scoring . . . . .	14
6.4	Extended Speaking and Writing Tasks . . . . .	14
<b>7</b>	<b>Statistical Characteristics</b>	<b>15</b>
7.1	Score Distribution . . . . .	15
7.2	Reliability . . . . .	17
7.3	Relationship with Other Tests . . . . .	17
<b>8</b>	<b>Conclusion</b>	<b>19</b>
	<b>References</b>	<b>20</b>
	<b>Appendix</b>	<b>23</b>

## 1 Duolingo English Test

The Duolingo English Test is a measure of English language proficiency. The test has been 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. It has been designed to be efficient. It takes less than one hour to complete the entire process of taking the test (i.e., onboarding, test administration, uploading). It is a computer-adaptive test (CAT), and it uses item types that provide maximal information about English language proficiency. It is designed to be user-friendly; the onboarding, user interface, and item formats are easy to interact with.

This document provides an overview of the design the Duolingo English Test. It contains a discussion of:

- the test's accessibility, delivery, proctoring and security processes;
- the demographic information of the test taking population;
- the test's items, how they were created, and how they are delivered and scored;
- and the statistical characteristics of the Duolingo English Test.

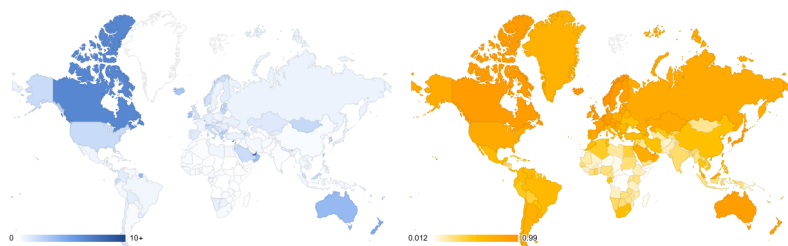
The test scores are intended to be interpreted as reflecting test takers' English language ability and used in a variety of settings, including for university admissions decisions.

## 2 Accessibility

Broad accessibility is one of the central motivations for the Duolingo English Test. Tests administered at test centers consume resources which limit accessibility: they require appointments at a physical testing center within certain hours on specific dates (and travel to the test center), and carry considerable registration fees. The maps in Figure 1 show the concentration of test centers in the world (left panel) compared to internet penetration in the world (right panel). The figure shows a stark difference in how much more easily an internet-based test can be accessed than a test center\*. While the ratio of internet access 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, and physical test centers can usually handle dozens of test-takers at once — it is still clear that the potential audience for the Duolingo English Test is orders of magnitude larger than those who could be served currently by more traditional test centers. By delivering assessments on-demand, 24 hours a day, to an estimated 2 billion internet-connected devices anywhere in the world for US\$49, we argue that the Duolingo English Test holds the potential to be the most accessible valid and secure language assessment platform in the world.

---

\*Central Africa is underserved by both models.



**Figure 1.** Global heatmaps comparing the number of physical test centers (left) per 1 million inhabitants (TOEFL®, IELTS®, and PTE® combined) vs. the percentage of internet users (right).

### 3 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 of these procedures are evaluated after the test has been administered and prior to score reporting.

#### 3.1 Test Administration

Test takers are required to take the test alone in a quiet environment. The Duolingo English Test can be taken in the Chrome and Opera browsers worldwide. In China, the test can be taken on the 360 and QQ browsers as well. An internet connection with at least 2 Mbps download speed and 1 Mbps upload speed is recommended for test sessions.

#### 3.2 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 the test takers' identification information is collected and that test takers are informed of the test's administration rules.

#### 3.3 Administration Rules

The list behaviors that are prohibited during an administration of the Duolingo English Test are listed below. In addition to these behavioral rules, there are rules for the test takers' internet browsers. The browsers are locked down after onboarding, which means that any navigation away from the browser invalidates the test session. Additionally, all browser plugins must be disabled.

- Interacting with anyone
- Allowing other people in the room
- Using headphones or earbuds
- Disabling the microphone or camera

- Looking off screen
- Moving out of frame of the camera
- Accessing any outside material or devices
- Leaving the web browser

3.4 Proctoring & 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. This process take 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 they elect to share their scores with. Test takers can share their scores with an unlimited number of institutions.

4 Test Taker Demographics

In this section, test taker demographics are reported. During the onboarding process of each test administration, test takers are asked to report their first language (L1), date of birth, and their gender identity. Their country of residence is logged when they show their proof of identification during the onboarding process. There were 23,460 people who took certified Duolingo English Tests between August 1, 2017 and June 30, 2019.

The most frequent L1s of Duolingo English Test test takers include Mandarin, Spanish, Arabic, and Portuguese (see Table 1). Duolingo English Test test takers represent 115 unique L1s and 168 unique countries. The full tables of all test taker L1s and countries of origin can be found in the Appendix.

Table 1. Most Frequent Test Taker L1s

First Language
Chinese - Mandarin
Spanish
Arabic
Portuguese
English
Korean
Japanese
French
Russian
Kurdish

Reporting gender identity during the onboarding process is optional, but reporting birth date is required. Table 2 shows that 34.60% of Duolingo English Test test takers identified as female, 41.29% of test takers identified as male, and 24.08% chose not to report. Table 3 shows that 78% of Duolingo English Test test takers are between 16 and 30 years of age.

**Table 2.** Counts and Percentages of Test Taker Gender

Gender	n	Percentage
Female	8,117	34.60%
Male	9,687	41.29%
Other	6	0.03%
Not reported	5,650	24.08%
Total	23,460	100.00%

**Table 3.** Counts and Percentages of Test Taker Age

Age	n	Percentage
< 16	1,338	5.70%
16 - 20	10,240	43.65%
21 - 25	4,999	21.31%
26 - 30	3,022	12.88%
31 - 40	2,731	11.64%
> 40	1,130	4.82%
Total	23,460	100.00%

## 5 Item Description

The test has seven different item types, which collectively tap into test takers' reading, writing, listening, and speaking abilities. Because the Duolingo English Test is a CAT, it will adjust in difficulty as the computer updates its real-time estimate of test takers' language proficiency as they progress through the test. There are five item types 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 of the Duolingo English Test, a test taker will see at minimum three of each CAT item type and at maximum of seven of each CAT item type. The median rate of occurrence of the CAT item types across all administrations is six times per test administration. Additionally, test takers respond to four writing prompts and three speaking prompts. They are not a 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 takers' English language proficiency in reading, writing, listening, and speaking.

### 5.1 C-test

The c-tests provide a measure of the test takers' reading ability (Klein-Braley 1997; Khodadady 2014). In this task, the first and last sentences are fully intact, while 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 2). The test taker needs to rely on context and discourse information to reconstruct the damaged words (which span multiple vocabulary and morpho-syntactic categories). It has been shown that c-tests are

1:32

Type the missing letters to complete the text below

Sam bought a new coat yesterday. He i s often c o l d, a n d he c h o s e a c o [ ] that i [ ] warmer t h [ ] his o [ ] jacket. His c o [ ] is g r [ ] and r [ ]. It l o [ ] good w i [ ] his r [ ] hat. He is going to wear it every day this winter.

NEXT

**Figure 2.** Example C-test Item

significantly correlated with many other major language proficiency tests, and additionally are related to spelling skills (Khodadady 2014).

## 5.2 Yes/No Vocabulary

This is a variant of the “yes/no” vocabulary test (Beeckmans et al. 2001). The test taker is presented with a set of English words mixed with pseudo-words that are designed to appear English-like, and must discriminate between them.<sup>†</sup> Such tests have been used to assess vocabulary knowledge at various CEFR levels (Milton 2010), and have been shown to predict language proficiency skills—the text version (see top panel in Figure 3) predicts listening, reading, and writing abilities; while the audio version (see bottom panel in Figure 3) predicts listening and speaking abilities in particular (Milton, Wade, and Hopkins 2010; Staehr 2008). These tests typically show a large set of stimuli (e.g., 60 words and 40 pseudo-words) of mixed difficulty at once. The format is made computer-adaptive by successively presenting multiple 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 6.1).

## 5.3 Dictation

In this exercise, the test taker listens to a spoken sentence or short passage and then transcribes it using the computer keyboard<sup>‡</sup> (see Figure 4). The test takers have one minute in total to listen to and transcribe what they heard. They can play the passage up to three times. This assesses

<sup>†</sup> 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.

<sup>‡</sup> Autocomplete, spell-checking, and other assistive device features or plugins are detected and disabled.

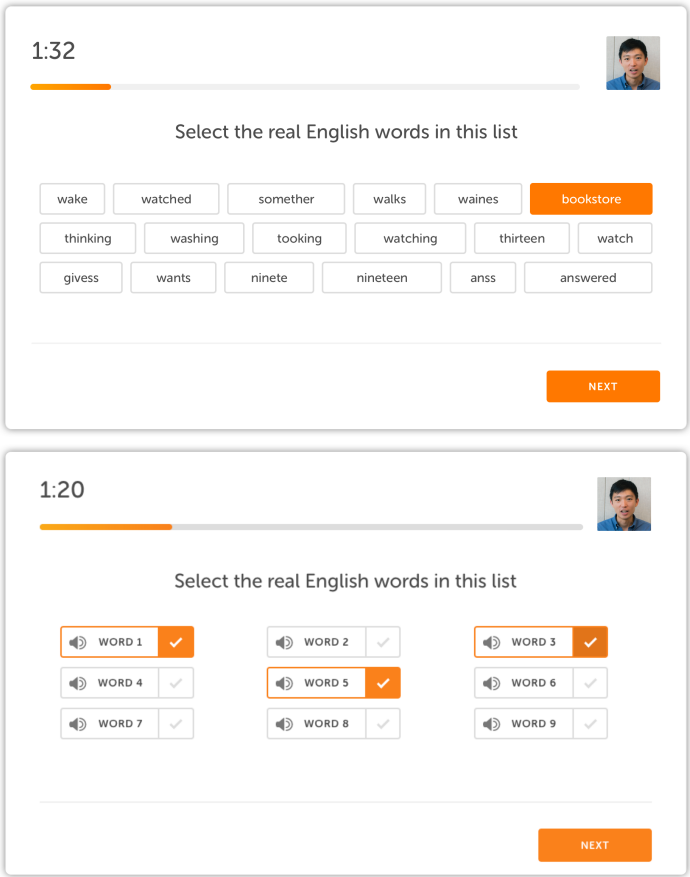


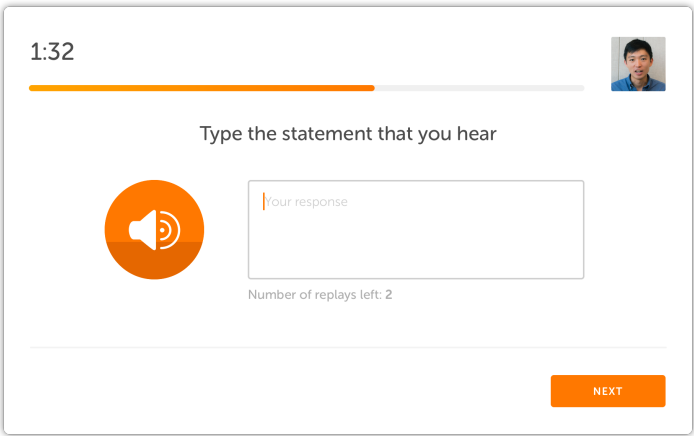
Figure 3. Example Yes/No Vocabulary Items

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 and Bent 2002; Buck 2001; Smith and Kosslyn 2007). Dictation tasks have also been found to be associated with language learner intelligibility in speech production (Bradlow and Bent 2008).

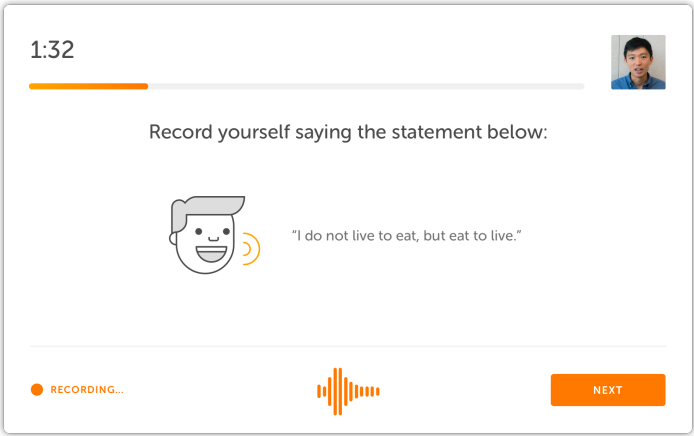
5.4 Elicited Imitation (Read-aloud)

The read-aloud variation of the elicited imitation task—example in Figure 5—is a measure of test taker reading and speaking abilities (Vinther 2002; Jessop, Suzuki, and Tomita 2007; Litman, Strik, and Lim 2018). It requires the 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





**Figure 4.** Example Dictation Item



**Figure 5.** Example Elicited Imitation Item

like intonation, rhythm, and stress (Anderson-Hsieh, Johnson, and Koehler 1992; Derwing, Munro, and Wiebe 1998; Field 2005; Hahn 2004). Furthermore, intelligibility is correlated with overall spoken comprehensibility (Munro and Derwing 1995; Derwing and Munro 1997; Derwing, Munro, and Wiebe 1998), meaning that this item format can capture aspects of speaking proficiency. We use state-of-the-art 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.

## 5.5 Extended Speaking

The extended speaking tasks are measures of test taker English speaking abilities. At the end of the CAT portion of the test, the test takers respond to four speaking prompts: one picture description task and three independent speaking tasks, two with a written prompt and one with an aural prompt (see Figure 6). 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. All of these task types require the test taker 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 and Palmer 1996).

## 5.6 Extended Writing

The extended writing tasks are measures of the test takers writing English abilities. Test takers respond to four writing prompts that require extended responses: three picture description tasks and one independent task with a written prompt (see Figure 7). Similar to the speaking tasks, these are drawn from different levels of difficulty conditional on the estimated ability level of the test taker. 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 (Cushing-Weigle 2002). The independent tasks require test takers to describe, recount, or make an argument; these require the test takers to demonstrate more discursive knowledge of writing in addition to language knowledge (Cushing-Weigle 2002).


## 6 Development, Delivery, & Scoring

This section explains how the computer-adaptive items in the test were developed, how the computer-adaptive test 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.


### 6.1 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 (and very large). As a result, it is not feasible to estimate  $b_i$  (item difficulty) statistically from actual administrations due to data sparsity, and 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


1:32



Describe aloud the image below




RECORDING...



NEXT

1:32




Speak your answer to the question below

Talk about a hobby or activity that you enjoy.


- What is it?
- How long have you been doing it?
- Who do you do it with?
- Why is it important to you?

RECORDING...




START

1:32




Speak the answer to the question you hear



Number of replays left: 2


RECORDING...




NEXT

Figure 6. Example Speaking Items

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

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 which are described in greater detail in (Settles, Hagirawa, and LaFlair, under revision).

## 6.2 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 (Wainer 2000; Segall 2005). Because computer-adaptive administration gives items to test takers conditional on their estimated ability, CATs have been shown to be shorter (Thissen and Mislevy 2000) and provide uniformly precise scores for most test takers when compared to fixed-form tests (Weiss and Kingsbury 1984).

To do this, we employ a generalization of item response theory (IRT). The conditional probability of an observed item score sequence  $\mathbf{g} = \langle g_1, g_2, \dots, g_t \rangle$  given  $\theta$  is the product of all the item-specific item response function (IRF) probabilities (assuming local item independence):

$$p(\mathbf{g}|\theta) = \prod_{i=1}^t p_i(\theta)^{g_i} (1 - p_i(\theta))^{1-g_i}, \quad (1)$$

where  $g_i$  denotes the scored response to item  $i$  (typically  $g_i = 1$  if correct,  $g_i = 0$  if incorrect), and  $1 - p_i(\theta)$  is the probability of an incorrect response under the IRF model. An implication of local independence is that the probability of responses for two separate test items  $i$  and  $j$  are independent of each other, controlling for the effect of  $\theta$ .

The purpose of a CAT is to estimate the ability ( $\theta$ ) of test takers as precisely as possible with as few test items as possible. The precision of our  $\theta$  estimate depends on the item sequence  $\mathbf{g}$ : test takers of higher ability  $\theta$  are best assessed by items with higher difficulty  $b_i$  (and likewise for lower values of  $\theta$  and  $b_i$ ). The true value of a test taker's ability ( $\theta$ ) 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 and added to  $\mathbf{g}$ , the process repeats until a stopping criterion is satisfied.

The maximum-likelihood estimation (MLE) approach to finding  $\hat{\theta}_t$  and selecting the next item is based on the log-likelihood function:

$$\begin{aligned} LL(\hat{\theta}_t) &= \log \prod_{i=1}^t p_i(\hat{\theta}_t)^{g_i} (1 - p_i(\hat{\theta}_t))^{1-g_i} \\ &= \sum_{i=1}^t g_i \log p_i(\hat{\theta}_t) + (1 - g_i) \log(1 - p_i(\hat{\theta}_t)). \end{aligned} \quad (2)$$

The first line directly follows from Equation (1), and is a typical formulation in the IRT literature. The rearrangement on the second line more explicitly relates the objective to minimizing *cross-entropy* (de Boer et al. 2005), a measure of disagreement between two probability distributions. This is because our test items are scored probabilistically (see Section 6.3). As a result,  $g_i$  is a probabilistic response ( $0 \leq g_i \leq 1$ ) rather than a binary response ( $g_i \in \{0, 1\}$ ). The MLE optimization in Equation (2) seeks to find the  $\hat{\theta}_t$  that yields an IRF prediction  $p_i(\hat{\theta}_t)$  that is most similar to each scored response  $g_i \in \mathbf{g}$ . This generalization,

combined with concise and predictive item formats, helps to minimize test administration time significantly.

Duolingo English Tests are variable-length, meaning that exam time and number of items can vary with each administration. 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 30-45 minutes long (including speaking and writing; excluding onboarding and uploading), and the median test consists of about 27 computer-adaptive (and eight extended response items) items with over 200 measurements<sup>§</sup>

Once the algorithm converges, the final reported score is not the provisional MLE point-estimate given by Equation (2) used during CAT administration. Rather,  $p(\theta|\mathbf{g})$  is computed for the CAT items for each possible  $\theta \in [0, 10]$  and normalized into a posterior distribution in order to create a weighted average score for each item type. These weighted average scores of each CAT item type are then used to create a composite score with the scores of the speaking and writing tasks.

### 6.3 Item Scoring

All test items are scored automatically via statistical procedures developed specifically for each format. For example, the yes/no vocabulary (see Figure 3) format 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 et al. 2001, Zimmerman et al. 1977). However, traditional yes/no tests assume that all stimuli are given at once, which is not the case in 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 test taker by item response and converted it to a score  $g_i$ , which can be interpreted as “the test taker can accurately discriminate between English words and pseudo-words at this score/difficulty level with probability  $g_i$ ,” where  $g_i \in [0, 1]$ .

Similarly, 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<sup>¶</sup> to provide its probabilistic score  $g_i$ .

### 6.4 Extended Speaking and Writing Tasks

The writing and speaking tasks are scored by automated scoring algorithms developed by ML and NLP experts at Duolingo. There are two separate algorithms: one for the speaking tasks and one for the writing tasks. Currently, the scores for the tasks are estimated at the portfolio level—meaning that the speaking score that is included in the composite score represents the test takers' performance on the four speaking tasks and the writing score represents the test takers' performance on the four writing tasks.

<sup>§</sup>For 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).

<sup>¶</sup>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  $r = -0.75$  ( $p < 0.001$ ).

The speaking and writing scoring systems evaluate each task based on the features listed below.

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

The writing scoring algorithm was trained on 3,626 writing performances, and the speaking scoring algorithm was trained on 3,966 performances. Both sets of performances were scored by human raters with TESOL/applied linguistics training. The algorithms were then evaluated through a process known as cross-validation. In this process, they are trained on a portion of the data (90%; the training set) and then evaluated on the remaining portion (10%; the test set). This design is called 10-fold cross-validation because the analysis is repeated 10 times on different configurations of 90/10 training/test sets.

This analysis used Cohen’s  $\kappa$  as the index of agreement (results in Table 4). It is a measure of probability of agreement with chance agreement factored out. The first row shows the rate of human-human agreement. The last two rows show rates of human-machine agreement. The  $\kappa$  index reflects agreement when the training set is used as the test set; this is expected to be higher than the cross-validated analysis. The  $\kappa_{xv}$  index shows the rates of agreement when using cross-validated analysis. All human-machine relationships show high rates of agreement ( $\kappa > 0.70$ ) between the algorithms’ scores and human rater scores.

**Table 4.** Machine-Human Agreement

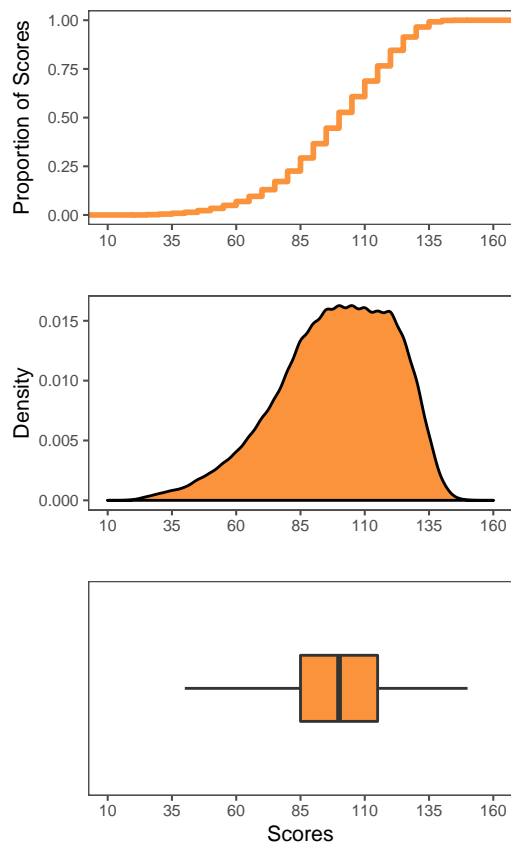
Scorers	Index	Writing	Speaking
Human:Human	$\kappa$	0.68	0.77
Human:Machine	$\kappa$	0.82	0.79
Human:Machine	$\kappa_{xv}$	0.73	0.77

## 7 Statistical Characteristics

This section provides an overview of the statistical characteristics of the Duolingo English Test. It includes information about the total score distribution and three reliability measures: test-retest, internal consistency, and standard error of measure.

### 7.1 Score Distribution

The following reports on an analysis of the composite scores for test administered between August 1, 2017 to June 30, 2019 that have been rescored to reflect the current operational scale: 10-160 in five point increments. This is the operational scale for all tests administered on or after July 15, 2019.



**Figure 8.** Distributions of Test Taker Scores

Figure 8 shows the distribution of test scores on the 10-160 point scale (on the x-axis of each plot). The top panel shows the empirical cumulative density function (ECDF) of the test scores. Where a test score meets the line in the ECDF, it shows the proportion of scores at or below that point. The middle panel shows the density function of the test scores, and the bottom panel shows a box plot of the total test scores.

The plots in Figure 8 show some negative skew, which is reflected in the descriptive statistics in Table 5. The mean and the median test score are 98.87 and 100 respectively, and the interquartile range is 30. Table 9 in the Appendix shows the percentage and cumulative percentage of test takers at each score point.

**Table 5.** Descriptive Statistics: Scale 10 - 160

n	Mean	SD	Median	IQR
23,460	98.83	22.49	100	30



## 7.2 Reliability

The reliability of the Duolingo English Test is evaluated by examining the relationship between scores from repeated test sessions (test-retest reliability), the relationship among the different halves of each test (split-half reliability; a measure of internal consistency), and the standard error of measure (SEM). The SEM is the range of test scores in which the test takers true score exists. It is a function of the test's reliability and standard deviation as shown in Equation (3).

$$SEM = SD * \sqrt{1 - reliability} \quad (3)$$

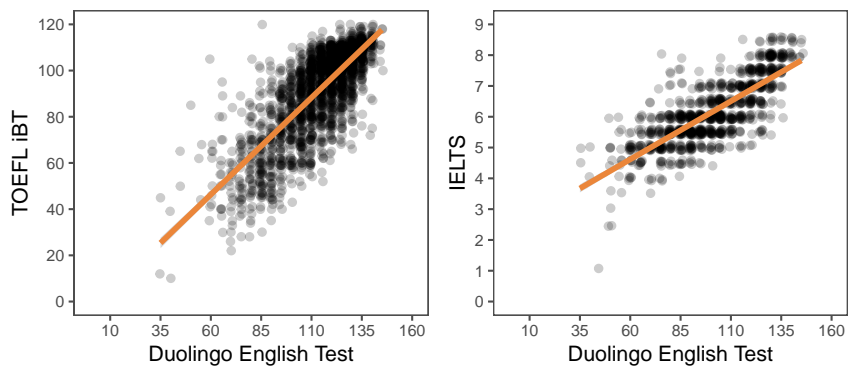
There have been 776 test takers who have taken the test twice within 30 days. The correlation between test scores at time one and test scores at time two is strong, positive, and significant ( $r = 0.85, p < 0.001$ ). The internal consistency of the Duolingo English Test is evaluated using split-half methods on the computer-adaptive portion only. A representative balance item types on each half is controlled for. The split-half reliability of the Duolingo English Test is also strong, positive, and significant ( $n = 23,460, r = 0.91, p < 0.001$ ). Using the test-retest reliability coefficient results in a conservative estimate of the SEM ( $\pm 8.71$  score points), which means that a test taker's true score falls within a range of 8.71 score points above or below their estimated ability.

## 7.3 Relationship with Other Tests

While the Duolingo English Test is being uploaded after the test administration, we ask test takers to share their recent TOEFL iBT and IELTS test scores (data collection started in August 2018). This data was used to evaluate the relationship between these two tests and the Duolingo English Test and to create concordance tables for the total test scores.

**Correlation** Pearson's correlations coefficients were estimated to evaluate the relationship between the Duolingo English Test and the TOEFL iBT and IELTS. The correlation coefficients for both revealed strong, positive, and significant relationships between the Duolingo English Test scores and the TOEFL iBT scores ( $n = 2,319, r = 0.77, p < 0.001$ ) and IELTS scores ( $n = 991, r = 0.78, p < 0.001$ ). These relationships are visualized in Figure 9. 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.

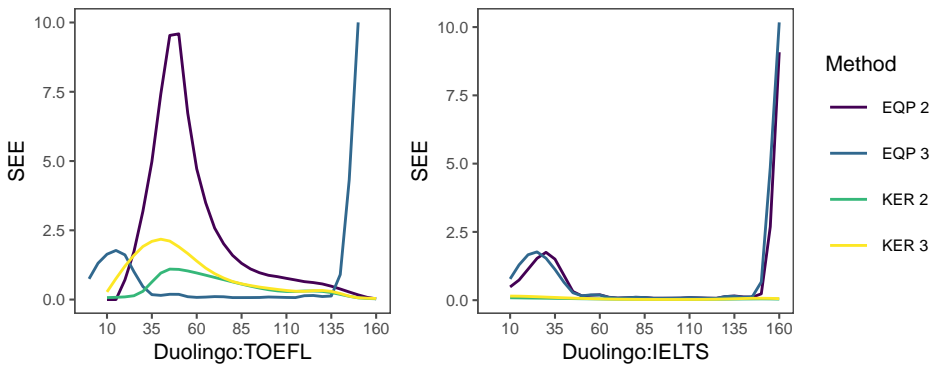
**Concordance** The same data from the correlation study was used to create concordance tables for Duolingo English Test users. Two types of equating were compared: equipercentile (Kolen and Brennan 2014) and kernel equating (Davies, Holland, and Thayer 2003). Within each equating type two methods were 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 *equate* (Albano 2016) and *kequate* (Andersson, Bränberg, and Wiberg 2013) packages in R (R Core Team 2018) were used to conduct the equating study.



**Figure 9.** Relationship between Test Scores

**Table 6.** Standard Error of Equating Summary

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



**Figure 10.** Relationship between Test Scores

The equating procedure that was selected to create the concordance tables was the one that minimized the mean standard error of equating. Table 6 shows that the the kernel equating that preserved the first two moments and the bivariate score relationship introduces the least amount of error. Figure 10 shows that the conditional error across the Duolingo English Test score range is very small for kernel equating as well. As a result, we used the “KER 2” equating relationship

to create our concordance tables. The concordance tables can be found at the Duolingo English Test scores page (<https://englishtest.duolingo.com/scores>).

## 8 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.

## 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., R. Johnson, and K. Koehler. 1992. “The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentals, Prosody, and Syllable Structure.” *Language Learning* 42: 529–55.
- Andersson, Björn, Kenny Bränberg, and Marie Wiberg. 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/>.
- Bachman, L. F., and A. S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.
- Beeckmans, R., J. Eyckmans, V. Janssens, M. Dufranne, and H. Van de Velde. 2001. “Examining the Yes/No Vocabulary Test: Some Methodological Issues in Theory and Practice.” *Language Testing* 18 (3): 235–74.
- Bradlow, A. R., and T. Bent. 2002. “The Clear Speech Effect for Non-Native Listeners.” *Journal of the Acoustical Society of America* 112: 272–84.
- . 2008. “Perceptual Adaptation to Non-Native Speech.” *Cognition* 106: 707–29.
- Buck, G. 2001. *Assessing Listening*. Cambridge: Cambridge University Press.
- Cushing-Weigle, S. 2002. *Assessing Writing*. Cambridge: Cambridge University Press.
- Davies, Alina A von, Paul W Holland, and Dorothy T Thayer. 2003. *The Kernel Method of Test Equating*. NY: Springer Science & Business Media.
- de Boer, P. T., D. P. Kroese, S. Mannor, and R. Y. Rubinstien. 2005. “A Tutorial on the Cross-Entropy Method.” *Annals of Operations Research* 34: 19–67.
- Derwing, T. M., and M. J. Munro. 1997. “Accent, Intelligibility, and Comprehensibility: Evidence from Four L1s.” *Studies in Second Language Acquisition* 19 (1): 1–16.
- Derwing, T. M., M. J. Munro, and G. Wiebe. 1998. “Evidence in Favor of a Broad Framework for Pronunciation Instruction.” *Language Learning* 48: 393–410.
- Fawcett, T. 2006. “An Introduction to ROC Analysis.” *Pattern Recognition Letters* 27 (8): 861–74.
- Field, J. 2005. “Intelligibility and the Listener: The Role of Lexical Stress.” *TESOL Quarterly* 39: 399–423.
- Hahn, L. D. 2004. “Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals.” *TESOL Quarterly* 38: 201–23.
- Jessop, L., W. Suzuki, and Y. Tomita. 2007. “Elicited Imitation in Second Language Acquisition Research.” *Canadian Modern Language Review* 64 (1): 215–38.

- Khodadady, E. 2014. “Construct Validity of C-tests: A Factorial Approach.” *Journal of Language Teaching and Research* 5 (November).
- Klein-Braley, C. 1997. “C-Tests in the Context of Reduced Redundancy Testing: An Appraisal.” *Language Testing* 14 (1): 47–84.
- Kolen, Michael J, and Robert L Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices*. NY: Springer Science & Business Media.
- Litman, D., H. Strik, and G. S. Lim. 2018. “Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities.” *Language Assessment Quarterly*, 1–16.
- Milton, J. 2010. “The Development of Vocabulary Breadth Across the CEFR Levels.” In *Communicative Proficiency and Linguistic Development: Intersections Between SLA and Language Testing Research*, edited by I. Bartning, M. Martin, and I. Vedder, 211–32. Eurosla.
- Milton, J., J. Wade, and N. Hopkins. 2010. “Aural Word Recognition and Oral Competence in English as a Foreign Language.” In *Insights into Non-Native Vocabulary Teaching and Learning*, edited by R. Chacón-Beltrán, C. Abello-Contesse, and M. Torreblanca-López, 52:83–98. Bristol: Multilingual Matters.
- Munro, M. J., and T. M. Derwing. 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*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Segall, D. O. 2005. “Computerized Adaptive Testing.” In *Encyclopedia of Social Measurement*, edited by K. Kempf-Leonard. New York, NY: Elsevier.
- Settles, B., M. Hagirawa, and G. T. LaFlair. under revision. “Machine Learning Driven Language Assessment.” *Transactions in Applied Computational Linguistics*.
- Smith, E. E., and S. M. Kosslyn. 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: 139–52.
- Thissen, D., and R. J. Mislevy. 2000. “Testing Algorithms.” In *Computerized Adaptive Testing: A Primer*, edited by H. Wainer. 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., and G. G. Kingsbury. 1984. “Application of Computerized Adaptive Testing to Educational Problems.” *Journal of Educational Measurement* 21: 361–75.

Zimmerman, J., P. K. Broder, J. J. Shaughnessy, and B. J. Underwood. 1977. "A Recognition Test of Vocabulary Using Signal-Detection Measures, and Some Correlates of Word and Nonword Recognition." *Intelligence* 1 (1): 5–31.

## Appendix

**Table 7.** Test Taker L1s in Alphabetical Order

Afrikaans	Dutch	Italian	Mandingo	Swahili
Akan	Efik	Japanese	Marathi	Swedish
Albanian	English	Kannada	Mongolian	Tagalog
Amharic	Ewe	Kashmiri	Nepali	Tajik
Arabic	Farsi	Kazakh	Northern Sotho	Tamil
Armenian	Finnish	Khmer	Norwegian	Tatar
Azerbaijani	French	Kikuyu	Oriya	Telugu
Bambara	Fulah	Kinyarwanda	Oromo	Thai
Belarusian	Ga	Kirundi	Polish	Tibetan
Bemba	Galician	Kongo	Portuguese	Tigrinya
Bengali	Ganda	Konkani	Punjabi	Tswana
Bikol	Georgian	Korean	Pushto	Turkish
Bosnian	German	Kurdish	Romanian	Turkmen
Bulgarian	Greek	Lao	Russian	Twi
Burmese	Guarani	Latvian	Serbian	Uighur
Catalan	Gujarati	Lingala	Sesotho	Ukrainian
Cebuano	Hausa	Lithuanian	Shona	Urdu
Chichewa (Nyanja)	Hebrew	Luo	Sinhalese	Uzbek
Chinese - Cantonese	Hindi	Luxembourgish	Slovak	Vietnamese
Chinese - Mandarin	Hungarian	Macedonian	Slovenian	Wolof
Croatian	Icelandic	Malagasy	Somali	Xhosa
Czech	Igbo	Malay	Spanish	Yoruba
Danish	Indonesian	Malayalam	Sundanese	Zulu

**Table 8.** Test Taker Country Origins in Alphabetical Order

Afghanistan	Cyprus	Lebanon	Russian Federation
Albania	Czechia	Lesotho	Rwanda
Algeria	Denmark	Liberia	Saudi Arabia
American Samoa	Dominica	Libya	Senegal
Angola	Dominican Republic	Lithuania	Serbia
Antigua and Barbuda	Ecuador	Luxembourg	Sierra Leone
Argentina	Egypt	Macao	Singapore
Armenia	El Salvador	Madagascar	Sint Maarten (Dutch)
Australia	Eritrea	Malawi	Slovakia
Austria	Ethiopia	Malaysia	Slovenia
Azerbaijan	Fiji	Mali	Somalia
Bahamas	Finland	Malta	South Africa
Bahrain	France	Mauritania	South Sudan
Bangladesh	Gabon	Mauritius	Spain
Barbados	Gambia	Mexico	Sri Lanka
Belarus	Georgia	Mongolia	State of Palestine
Belgium	Germany	Montenegro	Sudan
Belize	Ghana	Morocco	Suriname
Benin	Greece	Mozambique	Sweden
Bhutan	Guatemala	Myanmar	Switzerland
Bolivian Republic of Venezuela	Guinea	Nepal	Syrian Arab Republic
Bolivia	Haiti	Netherlands	Taiwan
Bosnia and Herzegovina	Honduras	New Zealand	Tajikistan
Botswana	Hong Kong	Nicaragua	Thailand
Brazil	Hungary	Niger	Togo
Bulgaria	Iceland	Nigeria	Trinidad and Tobago
Burkina Faso	India	North Macedonia	Tunisia
Burundi	Indonesia	Norway	Turkey
Cabo Verde	Iran (Islamic Republic)	Oman	Turkmenistan
Cambodia	Iraq	Pakistan	Uganda
Cameroon	Ireland	Panama	Ukraine
Canada	Israel	Papua New Guinea	United Arab Emirates
Central African Republic	Italy	Paraguay	United Kingdom of Great Britain and Northern Ireland
Chile	Jamaica	Peru	United Republic of Tanzania
China	Japan	Philippines	United States of America
Colombia	Jordan	Poland	Uruguay
Congo	Kazakhstan	Portugal	Uzbekistan
Congo (Democratic Republic)	Kenya	Puerto Rico	Viet Nam
Costa Rica	Kuwait	Qatar	Virgin Islands (British)
Côte d'Ivoire	Kyrgyzstan	Republic of Korea	Yemen
Croatia	Lao People's Democratic Republic	Republic of Moldova	Zambia
Cuba	Latvia	Romania	Zimbabwe



**Table 9.** Percentage Distribution: Scale 10 - 160

Score	Percentage	Cumulative percentage
150	0.01%	100.00%
145	0.09%	99.99%
140	0.72%	99.90%
135	2.66%	99.17%
130	5.14%	96.52%
125	6.83%	91.38%
120	7.99%	84.55%
115	7.80%	76.56%
110	8.00%	68.76%
105	8.09%	60.76%
100	8.09%	52.67%
95	7.98%	44.58%
90	7.33%	36.60%
85	6.73%	29.27%
80	5.37%	22.54%
75	4.19%	17.17%
70	3.42%	12.98%
65	2.63%	9.56%
60	1.99%	6.92%
55	1.55%	4.93%
50	1.13%	3.38%
45	0.87%	2.25%
40	0.52%	1.38%
35	0.41%	0.86%
30	0.28%	0.45%
25	0.15%	0.17%
20	0.02%	0.02%