Duolingo English Test: Technical Manual

duolingo english test

Duolingo Research Report
October 22, 2020 (33 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: October 22, 2020).

Contents

[*] Duolingo, Inc.

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

1 Duolingo English Test

The Duolingo English Test is a measure of English language proficiency for communication in English-medium settings. It assesses test taker ability to use language skills that are required for literacy, conversation, comprehension, and production. 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 about 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 of 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 are delivered and scored;
- and the statistical characteristics of the test.

The test scores are intended to be interpreted as reflecting test taker 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 development of 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 AuthaGraph[*] (Rudis & Kunimune, 2020) maps in Figure 1 shows 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 a stark difference in how much more easily an internet-based test can be accessed than a test center[†]. 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

[*] https://en.wikipedia.org/wiki/AuthaGraph_projection

[†] Central Africa is underserved by both models.

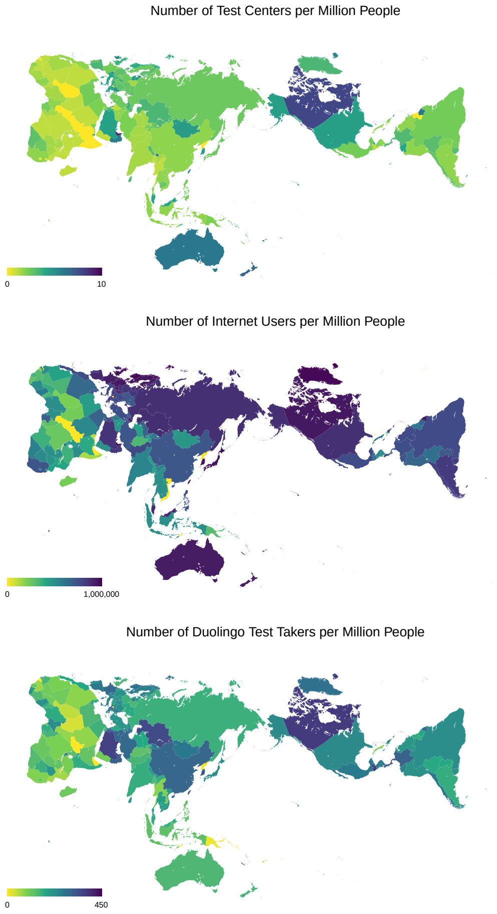

Number of Test Centers per Million People
0
10
Number of Internet Users per Million People
0
1,000,000
Number of Duolingo Test Takers per Million People
0
450

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

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 who could be served currently by more traditional test centers. The map in the bottom panel shows that the Duolingo English Test is beginning to realize this potential with for people taking the Duolingo English test from places with relatively low concentrations of test centers, such as countries in South and Central America (Colombia, 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, 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.

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 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 test taker identification information is collected, that test takers are informed of the test's administration rules, and that test takers agree to follow the rules.

3.3 Administration Rules

The 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 test taker 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. Test takers are also asked to be visible at all times to their cameras and to keep their camera and microphone enabled through the test administration.

- Leaving the camera preview

[‡] More information can be found in the Security, Proctoring, and Accommodations whitepaper.

- 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

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 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 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 99,451 people who took certified Duolingo English Tests between July 1, 2019 and June 30, 2020.

Test takers are asked to report their L1s during the onboarding process. The most frequent first languages of Duolingo English Test test takers include Mandarin, Spanish, Arabic, English[§], French, and Portuguese (see Table 1). There are 132 unique L1s represented by test takers of the Duolingo English Test and the test has been administered to test takers from 186 countries. The full tables of all test taker L1s and countries of origin can be found in the Appendix (Section 9).

Reporting gender identity during the onboarding process is optional, but reporting birth date is required. Table 2 shows that 50.57% of Duolingo English Test test takers identified

[§] 50% of these test takers come from India and Canada

Table 1. Most Frequent Test Taker L1s

First Language
Chinese - Mandarin
Spanish
Arabic
English
French
Farsi
Portuguese
Hindi
Korean
Russian

as female, 48.68% of test takers identified as male, and 0.75% chose not to report. Table 3 shows that 82% of the Duolingo English Test test takers are between 16 and 30 years of age.

Table 2. Counts and Percentage of Test Taker Gender

Gender	n	Percentage
Female	50,292	50.57%
Male	48,409	48.68%
Other	750	0.75%
Total	99,451	100.00%

Table 3. Counts and Percentage of Test Taker Age

Age	n	Percentage
< 16	5,131	5.16%
16 - 20	33,932	34.12%
21 - 25	33,992	34.18%
26 - 30	13,416	13.49%
31 - 40	10,368	10.43%
> 40	2,612	2.63%
Total	99,451	100.00%

5 Item Descriptions

The test has seven different item types, which collectively measure test taker ability to use language skills that are required for literacy, conversation, comprehension, and production. Because the Duolingo English Test is a CAT, it will adjust in difficulty as the computer updates its real-time estimate of test taker 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 four 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 taker English language proficiency in reading, writing, listening, and speaking.

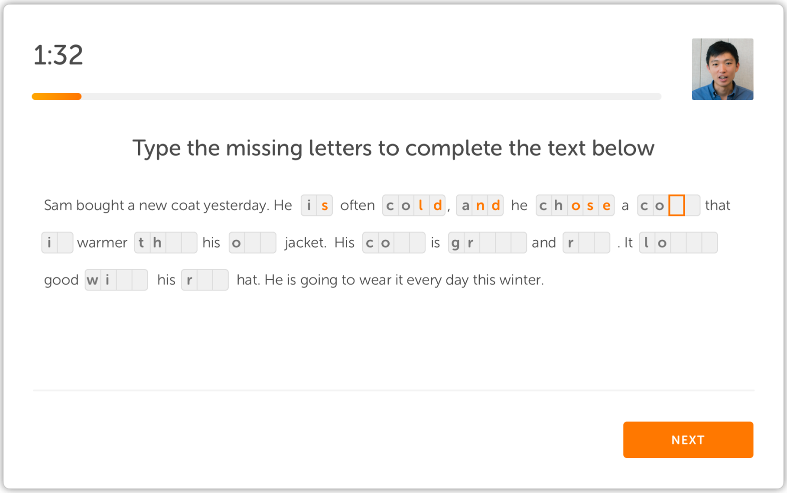
5.1 C-test

The c-tests provide a measure of test taker reading ability (Khodadady, 2014; Klein-Braley, 1997). 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). Test takers need 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 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, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001). Test takers are presented with a set of English words mixed with pseudo-words that are designed to appear English-like, and must discriminate between them[¶]. 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

[¶] 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 pseudowords that orthographically or phonetically resemble real English words too closely.



1:32

Type the missing letters to complete the text below

Sam bought a new coat yesterday. He is often cold, and he chose a co__ that i_ warmer th__ his o__ jacket. His co__ is gr___ and r__ . It lo___ good wi__ his r__ hat. He is going to wear it every day this winter.

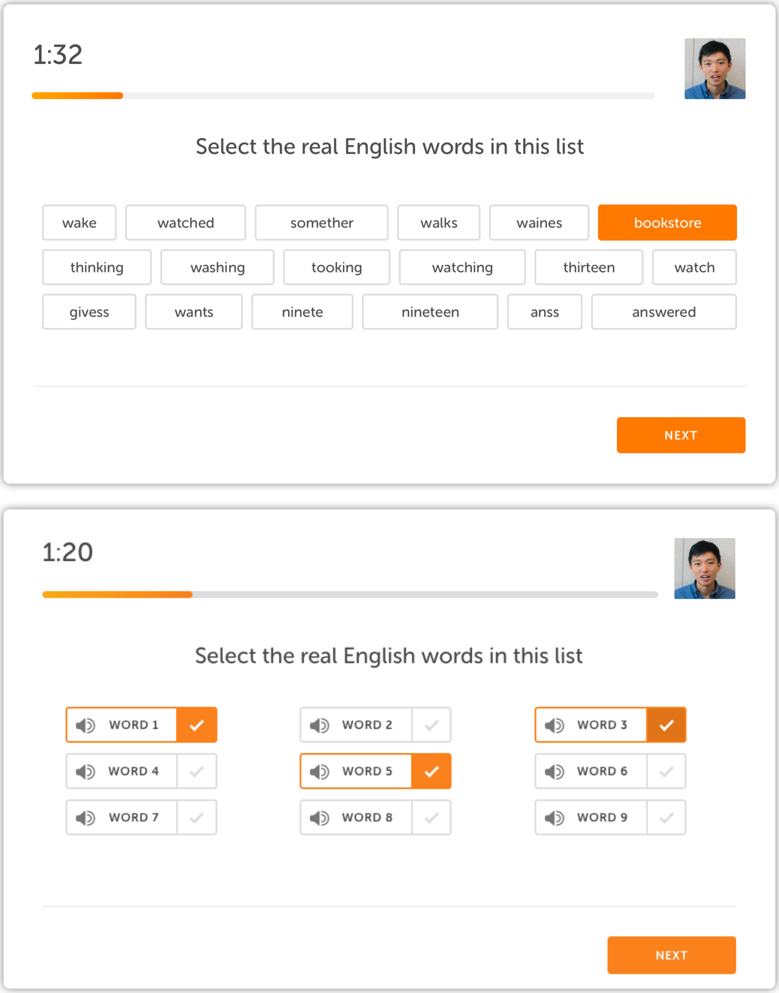
Figure 2. Example C-test Item

abilities; while the audio version (see bottom panel in Figure 3) predicts listening and speaking abilities in particular (McLean, Stewart, & Batty, 2020; Milton et al., 2010; Staehr, 2008). These tests typically show a large set of stimuli (say, 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, test takers listen to a spoken sentence or short passage and then transcribe it using the computer keyboard[‖] (see Figure 4). Test takers have one minute in total to listen to 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).

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



1:32

Select the real English words in this list

wake | watched | something | walks | waines | ☑ bookstore
thinking | washing | tooking | watching | thirteen | watch
givess | wants | ninete | nineteen | anss | answered

1:20

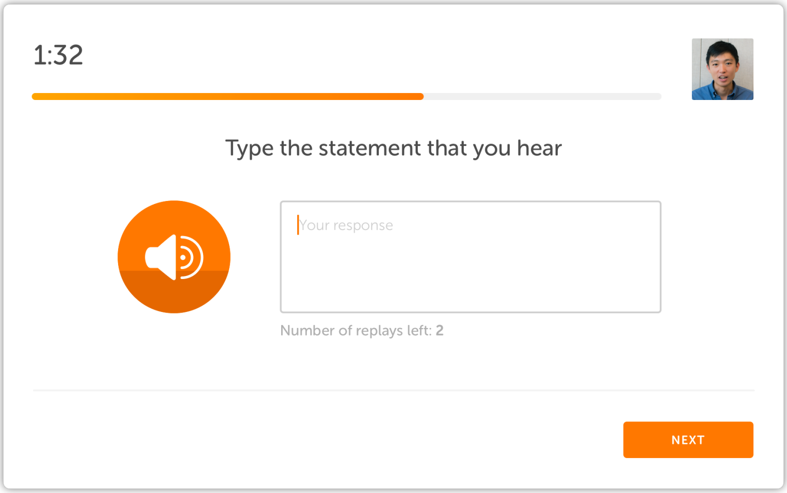
Select the real English words in this list

WORD 1 ☑ | WORD 2 ✓ | WORD 3 ☑
WORD 4 ✓ | WORD 5 ☑ | WORD 6 ✓
WORD 7 ✓ | WORD 8 ✓ | WORD 9 ✓

Figure 3. Example Yes/No Vocabulary Items

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 (Jessop, Suzuki, & Tomita, 2007; Litman, Strik, & Lim, 2018; Vinther, 2002). 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 like intonation, rhythm, and stress (Anderson-Hsieh,



1:32

Type the statement that you hear

Your response

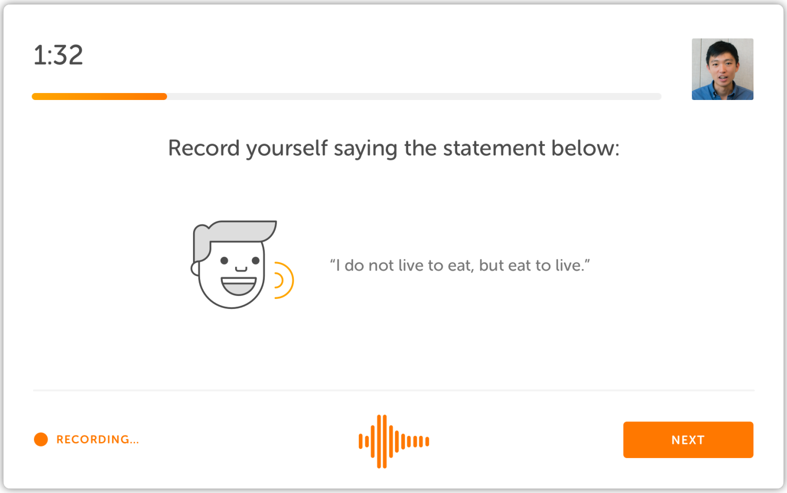
Number of replays left: 2

Figure 4. Example Dictation Item

Johnson, & Koehler, 1992; Derwing, Munro, & Wiebe, 1998; Field, 2005; Hahn, 2004). Furthermore, intelligibility is correlated with overall spoken comprehensibility (Derwing & Munro, 1997; Derwing et al., 1998; Munro & Derwing, 1995), 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 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)



Record yourself saying the statement below:

“I do not live to eat, but eat to live.”

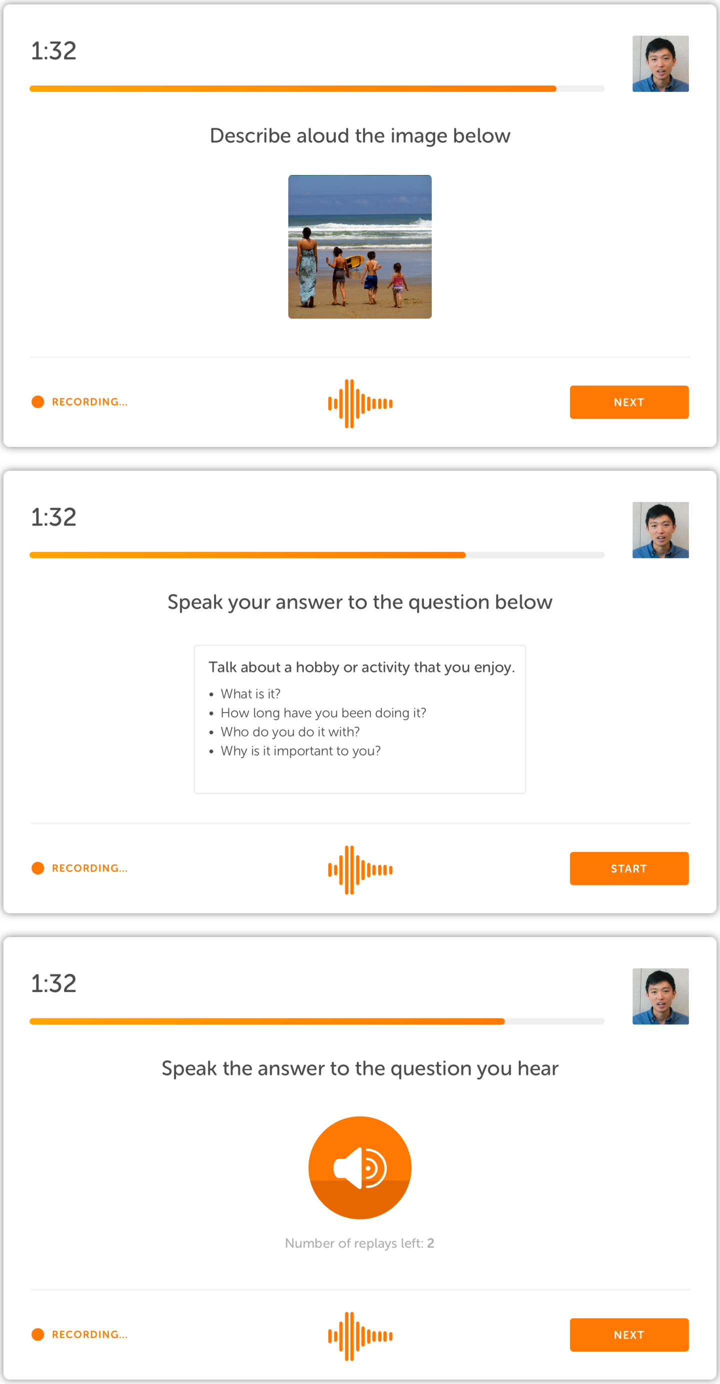
Figure 5. Example Elicited Imitation Item

5.6 Extended Writing

The extended writing tasks are measures of test taker English writing 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. 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.



Describe aloud the image below

Speak your answer to the question below

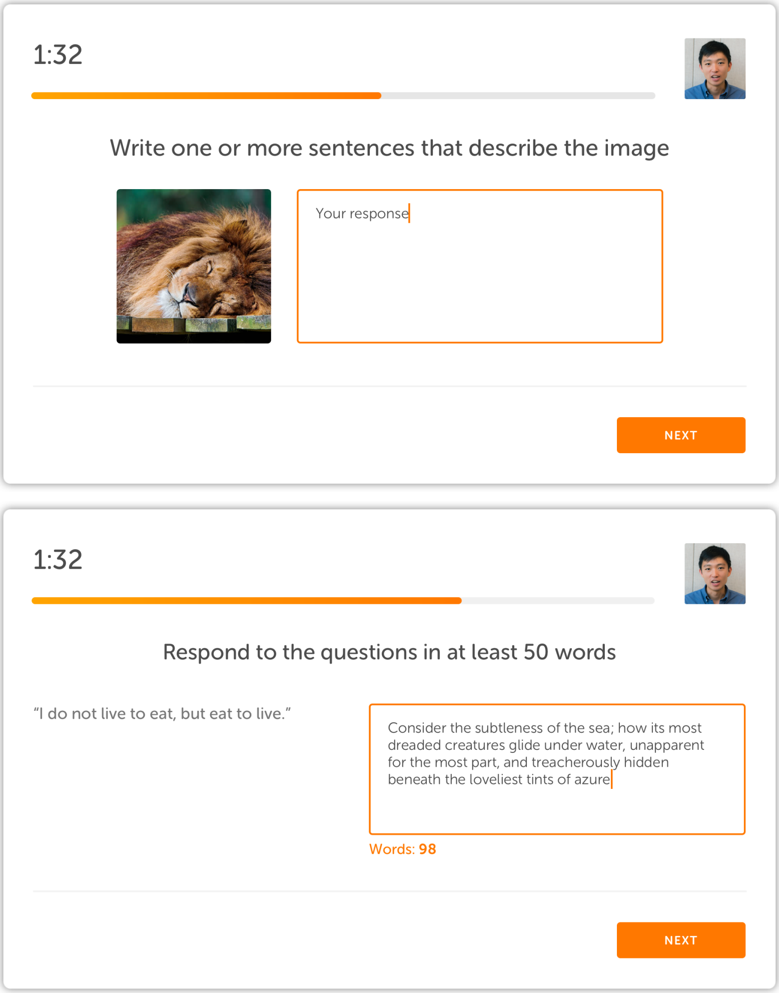
Talk about a hobby or activity that you enjoy.

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

Speak the answer to the question you hear

Number of replays left: 2

Figure 6. Example Speaking Items



Write one or more sentences that describe the image

Your response

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**

Figure 7. Example Writing Items

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 $\hat{b}_i$ (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 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, LaFlair, & Hagiwara (2020).

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 computer-adaptive testing (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).

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 θ is the product of all the item-specific item response function (IRF) probabilities (assuming local item independence):

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

where g_i denotes the graded 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 θ.

The purpose of a CAT is to estimate the ability (θ) of test takers as precisely as possible with as few test items as possible. The precision of our θ estimate depends on the item sequence $\mathbf{g}$: 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 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:

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

The first line directly follows from (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, Kroese, Mannor, & Rubinstien, 2005), a measure of disagreement between two probability distributions. This is because our test items are graded probabilistically (see Section 6.3. As a result, g_i as 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 graded 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**

Once the algorithm converges, the final reported score is not the provisional MLE point-estimate given by (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 total score, and the subscores, with the scores of the speaking and writing tasks.

**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).

6.3 CAT Item Scoring

All test items are graded 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, Broder, Shaughnessy, & Underwood, 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[††] to provide its probabilistic grade 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 total score represents test taker performance on the four speaking tasks and the writing score represents test taker performance on the four writing tasks.

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 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 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 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 κ 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 κ 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 κ_{xv} index shows the rates of agreement when using cross-validated analysis (i.e., there are separate training and test sets). All human-machine relationships show high rates of agreement ($\kappa > 0.70$) between the algorithm's scores and the human rater scores.

Table 4. Machine-Human Agreement

Scorers	Index	Writing	Speaking
Human:Human	κ	0.68	0.77
Human:Machine	κ	0.82	0.79
Human:Machine	κ_{xv}	0.73	0.77

7 Statistical Characteristics

This section provides an overview of the statistical characteristics of the Duolingo English Test. This section includes information about the score distribution and test score reliability for the total score and subscores. The analyses of the subscores were conducted on data from tests that were administered between July 1, 2019 and June 18, 2020.

7.1 Score Distributions

Figure 8 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. Where a test score meets the line in the ECDF, it shows the proportion of scores at or below that point.

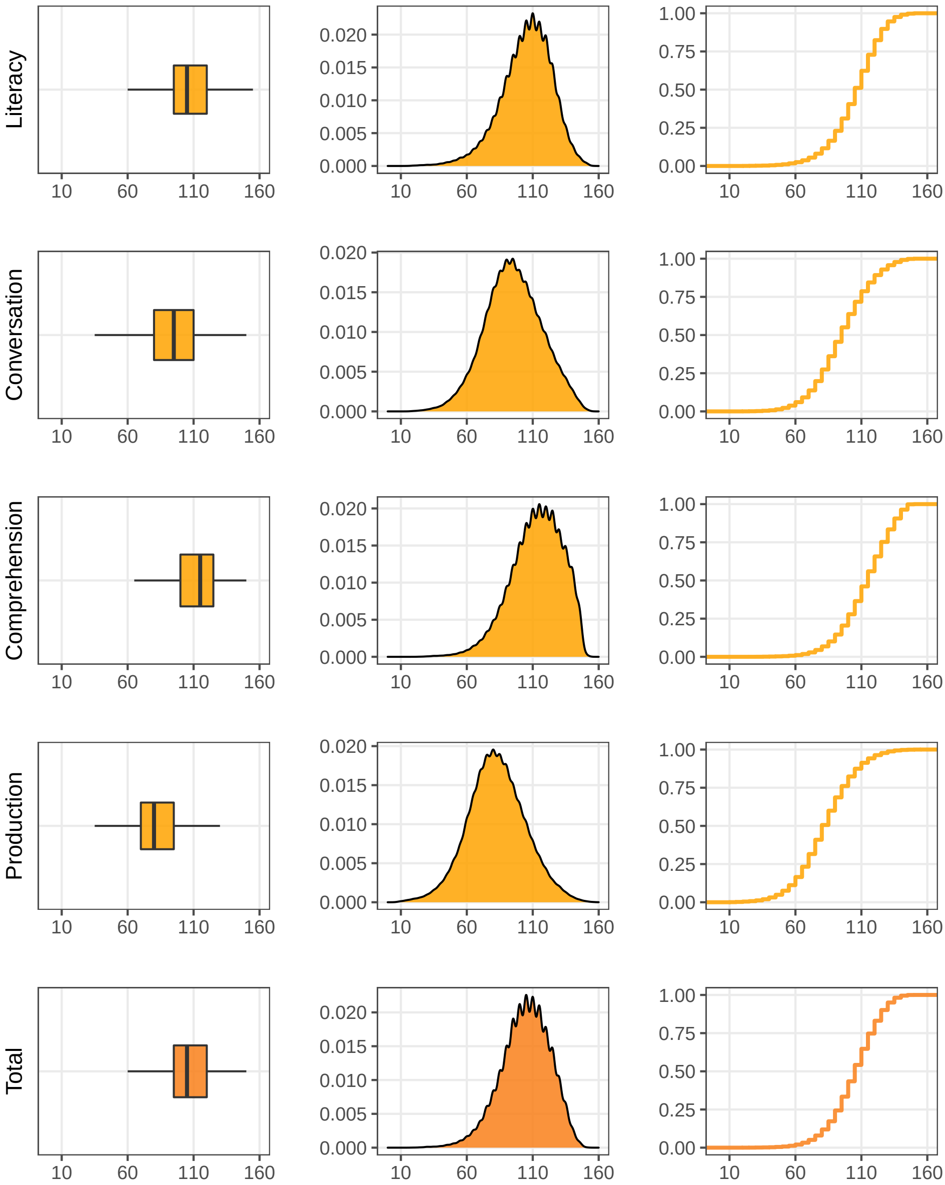


Figure 8. Boxplots (left), Density Plots (middle), and Empirical Cumulative Distribution Plots (right) of the Total Score and Subscores.

The plots in Figure 8 show some negative skew, which is reflected in the descriptive statistics in Table 5. The total score mean and the median test score are 104.42 and 105 respectively, and the interquartile range is 25. Tables 10-14 in the Appendix show the percentage and cumulative percentage of the total test scores and subscores. These are a numerical, tabled representations of the plots in Figure 8.

Table 5. Descritive Statistics for Total and Subscores (n = 99,415)

Score	Mean	SD	25th Percentile	Median	75th Percentile
Comprehension	112.91	19.27	100	115	125
Conversation	95.23	21.22	80	95	110
Literacy	105.18	19.03	95	105	120
Production	82.69	21.60	70	80	95
Total	104.42	18.48	95	105	120

7.2 Reliability

The reliability of the Duolingo English Test is evaluated by examining the relationship between repeated 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 data for each of these measures come from a subset of the 99,451 certified tests administered between July 1, 2019 and June 30, 2020. There were 10,187 people who have taken the test twice within 30 days. The data for the internal consistency index come from 8,041 test session in which the speaking and writing portfolios have been separated into their four prompts which were rescored using a retrained version of the automated scoring systems. This process allowed for the random assignment of half of each of the speaking and writing prompts to either side of the split-half test. The coefficients for the subscores and the total score in Table 6 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 (3), where x is a total score or subscore, SD is the standard deviation of the total score or subscore, and α is split-half 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 all subscores except Production.

$$SEM_x = SD_x \ * \sqrt{1 - \alpha_x} \tag{3}$$

[‡‡] Coefficients are adjusted to full-length estimates using the Spearman-Brown prophecy formula

Table 6. Test-retest (n = 10,187), Internal Consistency (n = 8,041), and SEM Estimates

Score	Test-retest	Internal Consistency	SEM	SEM (rounded)
Literacy	0.80	0.88	6.48	5
Conversation	0.78	0.93	5.67	5
Comprehension	0.76	0.95	4.12	5
Production	0.81	0.75	10.85	10
Total	0.82	0.95	3.92	5

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

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 a strong, positive relationship between the Duolingo English Test scores and the TOEFL iBT scores ($r = 0.77$; $n = 2{,}319$) and IELTS scores ($r = 0.78$; n = 991). 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.

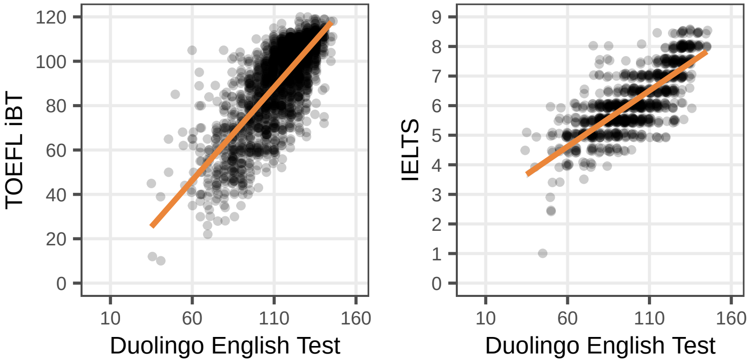


Figure 9. Relationship between Test Scores

Concordance

The same data from the correlation study was used to create concordance tables for Duolingo English Test score users. Two types of equating were compared: equipercentile (Kolen & Brennan, 2014) and kernel equating (Davier, 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. To conduct the equating study the *equate* (Albano, 2016) and *kequate* (Andersson, Bränberg, & Wiberg, 2013) packages in R (R Core Team, 2018) were used.

Table 7. Standard Error of Equating Summary

	TOEFL			IELTS	
Method	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

The equating procedure that was selected to create the concordance tables was the one that minimized the mean standard error of equating. Table 7 shows that this was the kernel equating that preserved the first two moments and the bivariate score relationship. Figure 10 shows that the conditional error across the Duolingo English Test score range is very small for kernel equating as well. The concordance tables can be found at the Duolingo English Test scores page(https://englishtest.duolingo.com/scores).

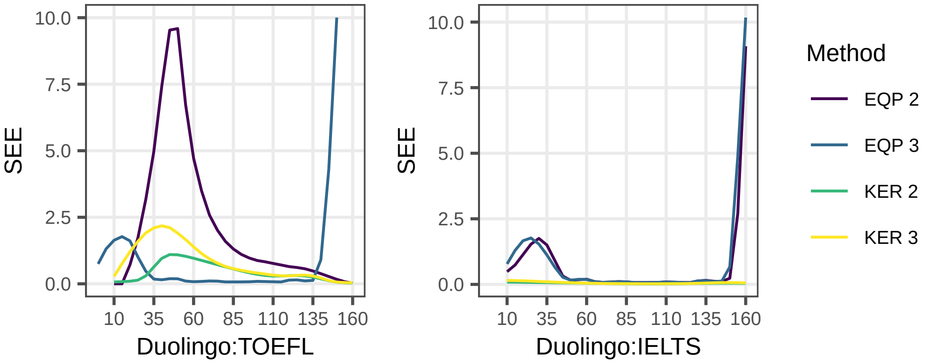


Figure 10. Conditional Standard Error of Equating

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.

9 Appendix

Table 8. Test Taker L1s in Alphabetical Order

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

Table 9. Test Taker Country Origins in Alphabetical Order

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

Table 10. Percentage Distribution Total Score

Total	Percentage	Cumulative percentage
150	0.02%	100.00%
145	0.42%	99.98%
140	1.37%	99.55%
135	3.13%	98.19%
130	4.93%	95.06%
125	7.01%	90.13%
120	8.39%	83.12%
115	10.00%	74.73%
110	10.55%	64.73%
105	10.73%	54.18%
100	10.05%	43.45%
95	9.04%	33.41%
90	7.08%	24.36%
85	5.35%	17.28%
80	3.96%	11.93%
75	2.88%	7.96%
70	1.84%	5.09%
65	1.22%	3.25%
60	0.79%	2.04%
55	0.49%	1.25%
50	0.29%	0.76%
45	0.20%	0.47%
40	0.11%	0.27%
35	0.07%	0.16%
30	0.06%	0.09%
25	0.02%	0.03%
20	0.01%	0.01%

Table 11. Percentage Distribution Literacy

Literacy	Percentage	Cumulative percentage
155	0.00%	100.00%
150	0.25%	100.00%
145	0.70%	99.75%
140	1.51%	99.05%
135	2.87%	97.54%
130	4.92%	94.68%
125	7.38%	89.76%
120	9.55%	82.38%
115	10.51%	72.83%
110	11.16%	62.31%
105	10.63%	51.15%
100	9.44%	40.53%
95	8.08%	31.09%
90	6.50%	23.01%
85	4.91%	16.51%
80	3.54%	11.59%
75	2.56%	8.05%
70	1.77%	5.49%
65	1.22%	3.72%
60	0.80%	2.50%
55	0.61%	1.71%
50	0.38%	1.10%
45	0.27%	0.72%
40	0.17%	0.45%
35	0.10%	0.28%
30	0.09%	0.18%
25	0.06%	0.09%
20	0.03%	0.04%
15	0.00%	0.00%

Table 12. Percentage Distribution Conversation

Conversation	Percentage	Cumulative percentage
150	0.15%	100.00%
145	0.67%	99.85%
140	1.36%	99.18%
135	2.09%	97.82%
130	2.81%	95.73%
125	3.67%	92.93%
120	4.82%	89.26%
115	5.75%	84.45%
110	6.92%	78.69%
105	8.01%	71.77%
100	8.70%	63.76%
95	9.51%	55.06%
90	9.42%	45.54%
85	8.69%	36.13%
80	7.61%	27.44%
75	6.09%	19.83%
70	4.55%	13.74%
65	3.14%	9.19%
60	2.26%	6.06%
55	1.50%	3.80%
50	0.96%	2.30%
45	0.60%	1.33%
40	0.31%	0.73%
35	0.21%	0.42%
30	0.12%	0.21%
25	0.06%	0.09%
20	0.03%	0.03%
15	0.00%	0.00%

Table 13. Percentage Distribution Comprehension

Comprehension	Percentage	Cumulative percentage
150	0.14%	100.00%
145	3.48%	99.86%
140	5.79%	96.38%
135	7.15%	90.59%
130	8.21%	83.44%
125	9.49%	75.23%
120	9.74%	65.74%
115	9.89%	56.00%
110	9.61%	46.11%
105	8.66%	36.51%
100	7.37%	27.85%
95	5.90%	20.48%
90	4.51%	14.58%
85	3.25%	10.07%
80	2.34%	6.82%
75	1.60%	4.49%
70	1.04%	2.88%
65	0.70%	1.85%
60	0.42%	1.14%
55	0.28%	0.72%
50	0.16%	0.45%
45	0.11%	0.29%
40	0.07%	0.17%
35	0.06%	0.10%
30	0.03%	0.04%
25	0.01%	0.01%
20	0.00%	0.00%

Table 14. Percentage Distribution Production

Production	Percentage	Cumulative percentage
155	0.02%	100.00%
150	0.07%	99.98%
145	0.17%	99.91%
140	0.35%	99.74%
135	0.63%	99.39%
130	1.03%	98.76%
125	1.39%	97.73%
120	2.12%	96.34%
115	2.91%	94.22%
110	3.87%	91.31%
105	5.08%	87.45%
100	6.25%	82.37%
95	7.45%	76.12%
90	8.71%	68.67%
85	9.36%	59.96%
80	9.69%	50.61%
75	9.34%	40.91%
70	8.31%	31.58%
65	6.77%	23.27%
60	5.32%	16.50%
55	3.63%	11.19%
50	2.67%	7.56%
45	1.75%	4.89%
40	1.16%	3.14%
35	0.77%	1.98%
30	0.48%	1.21%
25	0.30%	0.73%
20	0.22%	0.43%
15	0.14%	0.20%
10	0.07%	0.07%

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*. Campbridge: Cambridge University Press.

Cushing-Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Davier, A. A. von, Holland, P. W., & Thayer, D. T. (2003). *The kernel method of test equating*. NY: Springer Science & Business Media.

de Boer, P. T., Kroese, D. P., Mannor, S., & Rubinstien, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research*, *34*, 19–67.

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.

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.

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.