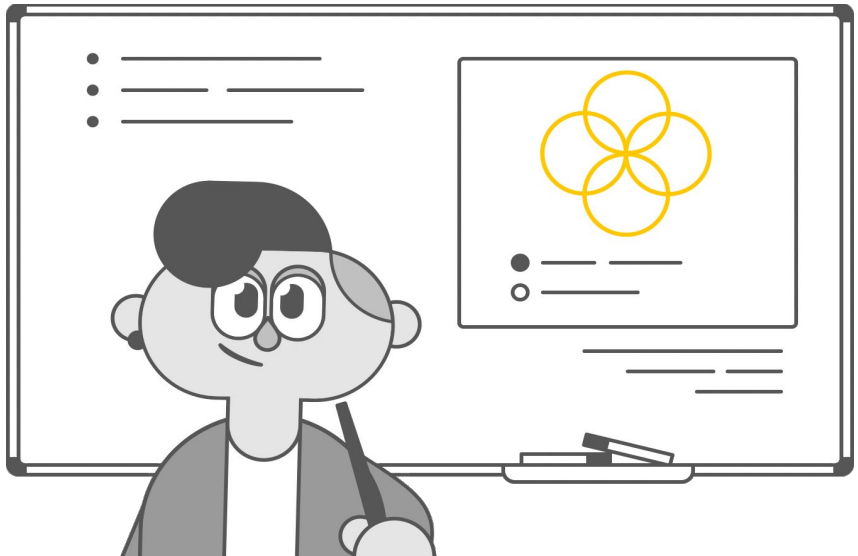
Duolingo English Test

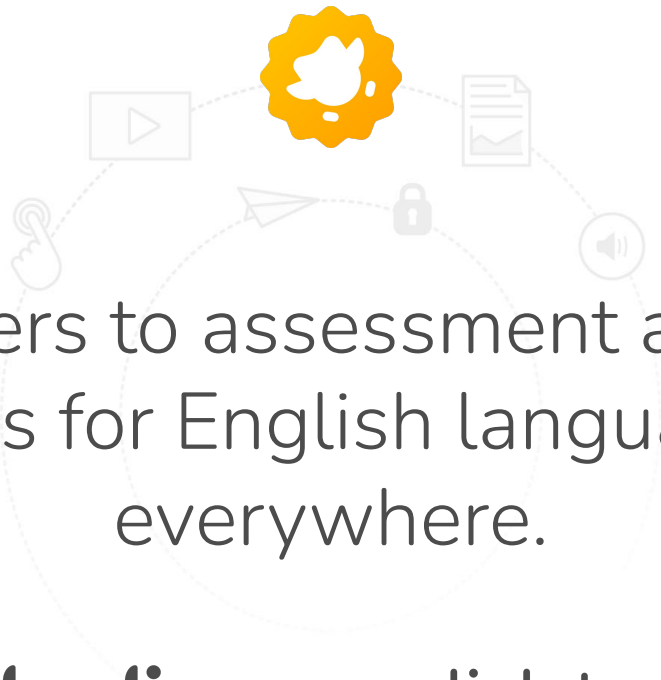
Dr Geoffrey T. LaFlair

November 18, 2022

Agenda

- Mission
- Overview of DET & Purpose
- Computer adaptive tests (CATs)
- Reliability & relationship with other tests
- Security & Quality assurance



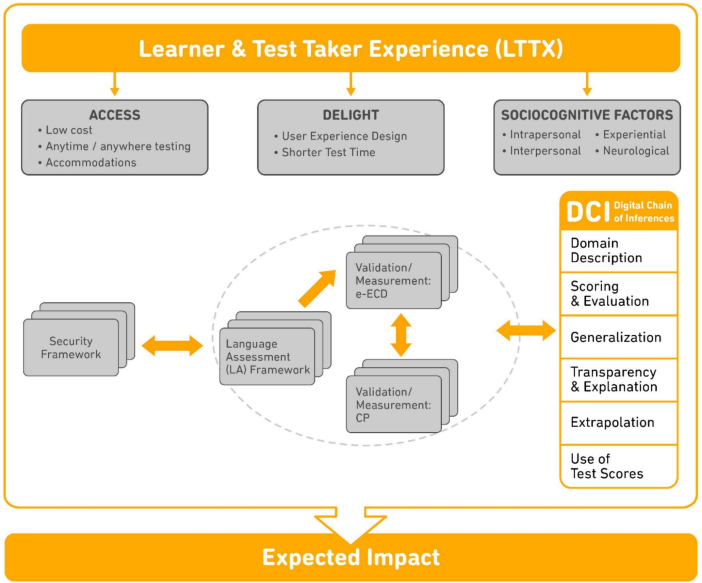


Lower barriers to assessment and increase opportunities for English language learners everywhere.

duolingo english test

Digital-first design

How can we use technology to enhance how we measure language ability?



Integrate digital affordances:

- Task design
- Test delivery
- Task interaction
- Task scoring
- Security
- Psychometrics

Overview

Duolingo English Test

Overview

- Test scores are used to make university admissions decisions
- Computer adaptive test (CAT) of English language proficiency
- Delivered online, remotely at any time of day
- Proctored after test completion (2-3 people)
- Score turn-around < 48 hours
- ML/NLP tools for item generation and difficulty estimation

DET Item types

Table 1. Summary of Item Formats on the Duolingo English Test

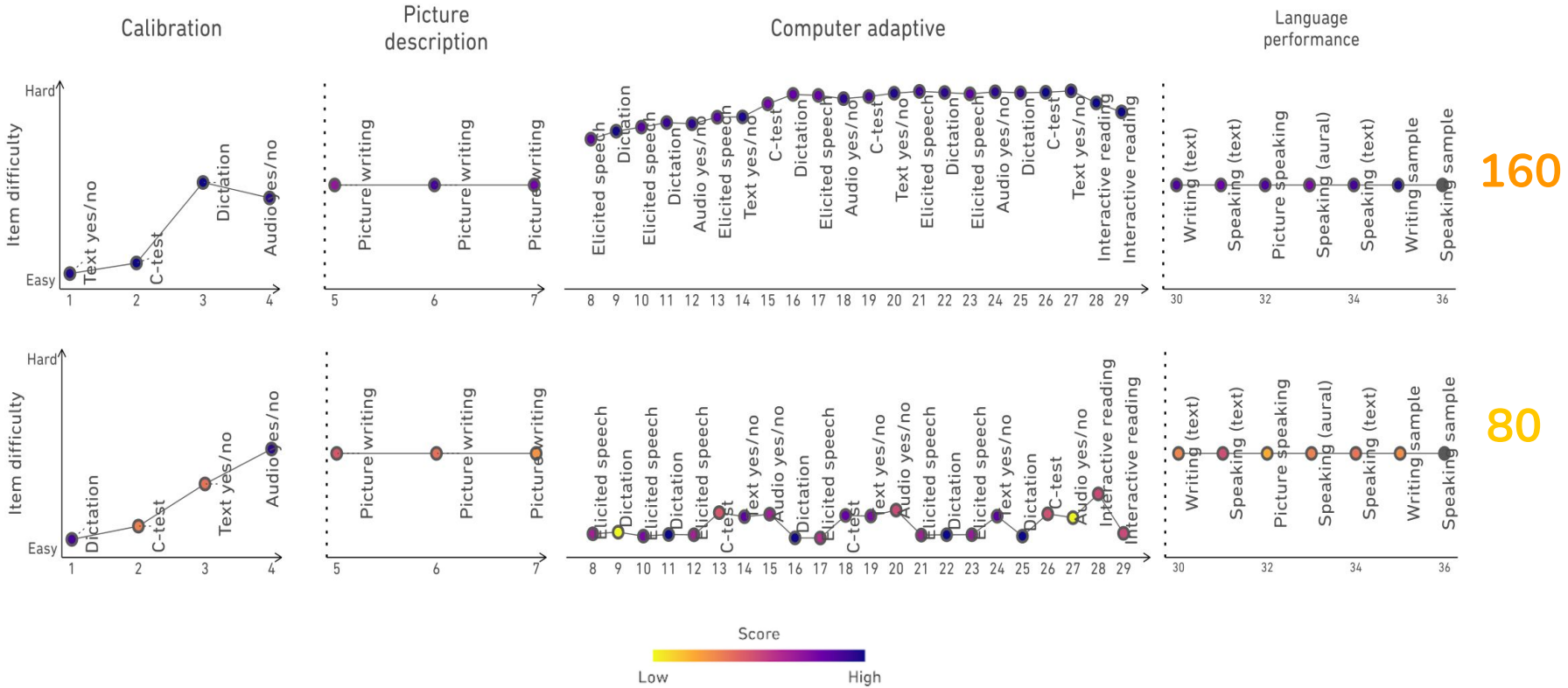
Item Type	Name for Test Takers	Type	Freq./Test
1. C-test	Read and Complete	CAT	4–6
2. Yes/no (text)	Read and Select	CAT	4–6
3. Yes/no (audio)	Listen and Select	CAT	4–6
4. Dictation	Listen and Type	CAT	4–6
5. Elicited imitation	Read Aloud	CAT	4–6
6. Interactive reading	Interactive Reading	non–CAT	2
7.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

Administration

- **CAT:** adaptively selected & feed into adaptive algorithm
- **IR (non-CAT):** Adaptively selected (do not feed into algorithm)
- **Perform/ungraded:** Randomly assigned

Computer adaptive testing

DET Administration: Four stages



Why CATs?

- **The algorithm**
  - Creates estimates of ability based on performance
  - Selects the next time to match that estimate of ability
- **Security**
  - Large enough so that unique tests are sufficiently different
  - Large enough so that harvesting cannot result in pre-knowledge
- **Test taker experience**
  - Cover the difficulty/ability continuum (tailored to proficiency)
  - Represent the construct at different levels of ability

Efficient and precise measure of English language proficiency

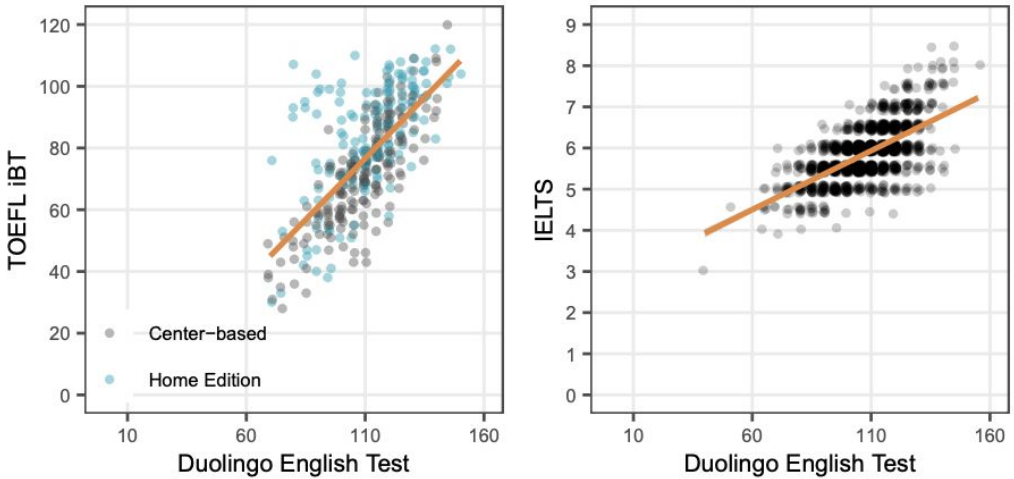
Reliability & Relationships

Reliability

- Weighted test-retest (2 days) correlation
- Country, L1, age, gender, OS, TOEFL scores, IELTS scores, Time 1 DET scores (means / variances)

Score	Test-Retest R
Overall	0.93
Literacy	0.90
Conversation	0.90
Comprehension	0.92
Production	0.90

Relationships



Official & Self-reported scores: March-August, 2022

Type	$TOEFL_n$	$IELTS_n$
Official	328	1,642
Self-rep	1,095	4,420

	$TOEFL_r$	$IELTS_r$
All	.71 (328)	.65 (1,643)
Center	.82 (183)	—
At home	.61 (145)	—

Concordance tables

Test concordance: Mapping between scales of tests of comparable constructs

Requires consideration of:

- Construct coverage of the tests
- Target population
- Sample size and representativeness
- Equating methods

The three tests have the same purpose—university admission—and cover a similar construct, English language skills for academic success and their scores can be compared.

DET	IELTS Academic
160	8.5–9
150–155	8
140–145	7.5
130–135	7
120–125	6.5
105–115	6
95–100	5.5
80–90	5
65–75	4.5
10–60	0–4

More information on concordance can be found at englishtest.duolingo.com/scores

Security
&
Quality assurance

The asynchronous advantage

Live proctoring relies on a chain of trust consisting of dozens of humans.

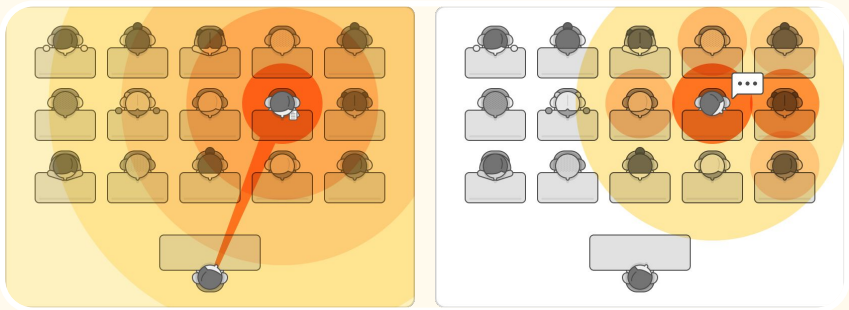
- A single point of failure can disrupt the entire chain.

Synchronous proctoring doesn't allow for multiple rounds of review

- No evidence of test taker behavior other than the proctor's narrative, based on memory.

Since proctors and test takers are in the same room, they cannot be anonymous.

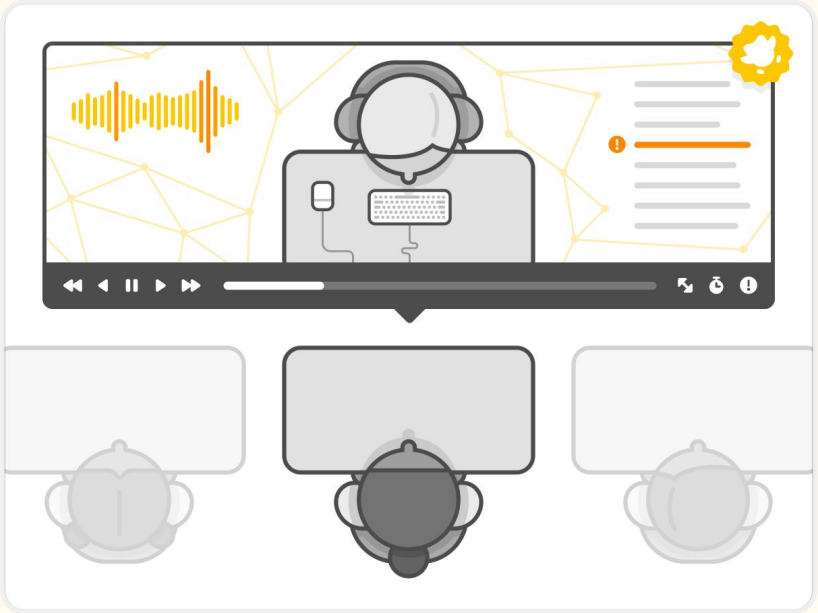
A single instance of cheating can mean that an entire batch of test takers may be unable to get a valid score.



englishtest.duolingo.com/security

How Duolingo keeps digital tests secure

- Each test session is adaptive and recorded via the computer, microphone, keyboard and mouse, and then reviewed in multiple rounds of proctoring.
- Trained ESL proctors review asynchronously and remotely, and are anonymous to test takers and to one another.
- Each test session is discrete and unique, so one person's attempt at cheating can't invalidate another test taker's score.
- Because each session is recorded, there's more evidence available if cheating occurred.



englishtest.duolingo.com/security

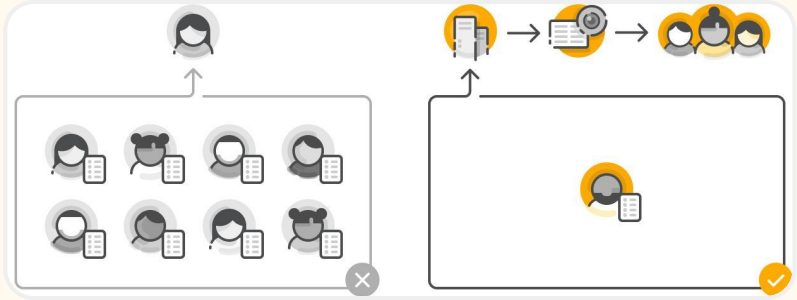
Advanced proctoring process

Expert human proctors, with the help of AI, examine each test session for over **150 different behaviors** over multiple, independent rounds of review.

Using the test video, audio, screen recording, keystrokes, mouse movement, and other recorded variables, proctors examine:

- The test taker's environment
- Eye movement
- Background noise
- Irregular behavior
- Other suspicious activities

Completed within 48 hours after test submission.



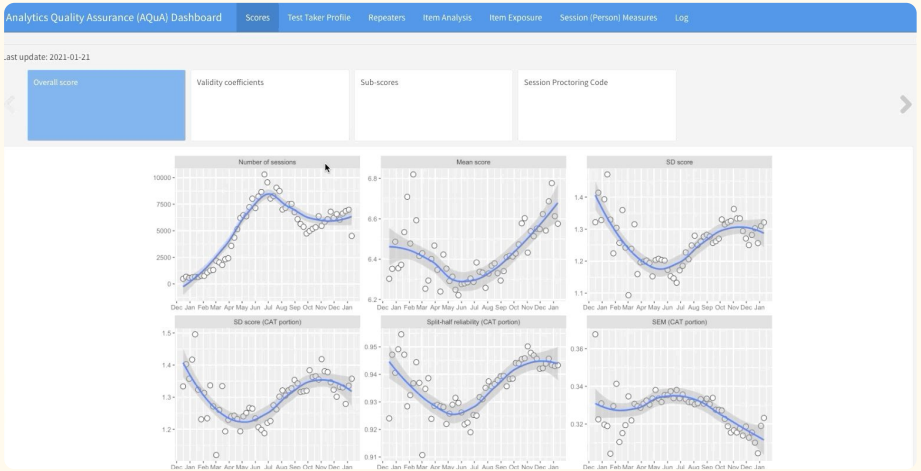
englishtest.duolingo.com/security

AQuAA

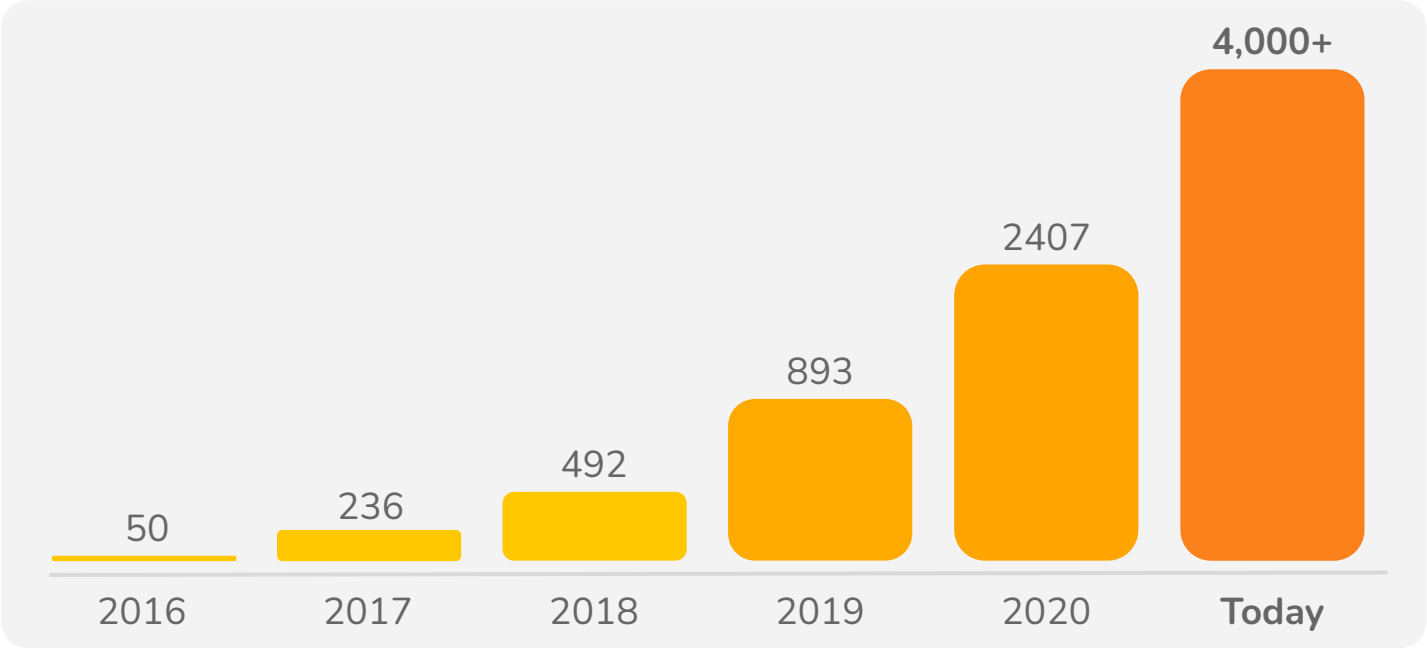
Analytics for **Qu**ality **A**ssurance in **A**ssessment

- An internally developed dashboard that tracks and reports on all validity-related metrics of our test daily across the globe

- Provides evidence for the validity of the interpretations and uses of our test scores (e.g. using the DET for higher ed admissions).



Thanks! Questions?



Number of Accepting Institutions
Source: englishtest.duolingo.com/institutions

Email Enquiry

Browse Research