



**English Language Assessment
for Prospective International Graduate Students:
Promoting Greater Access While Ensuring Valid and Secure Results**


March 25, 2021

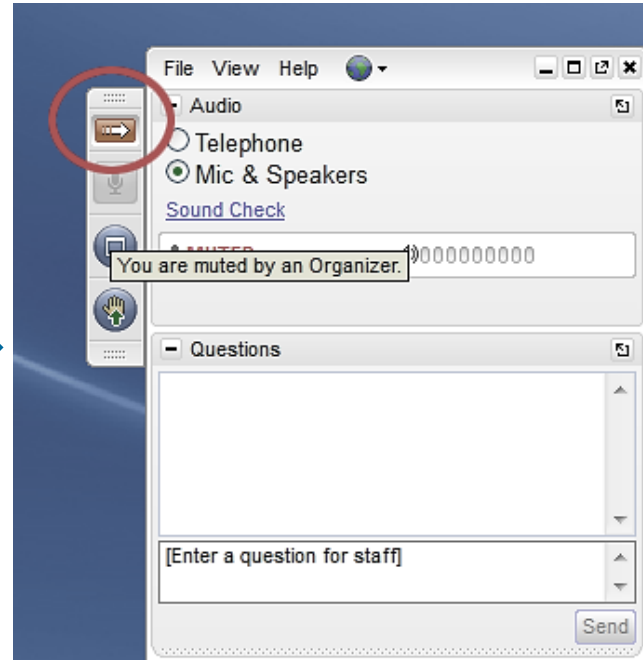
Webinar sponsored by



duolingo english test

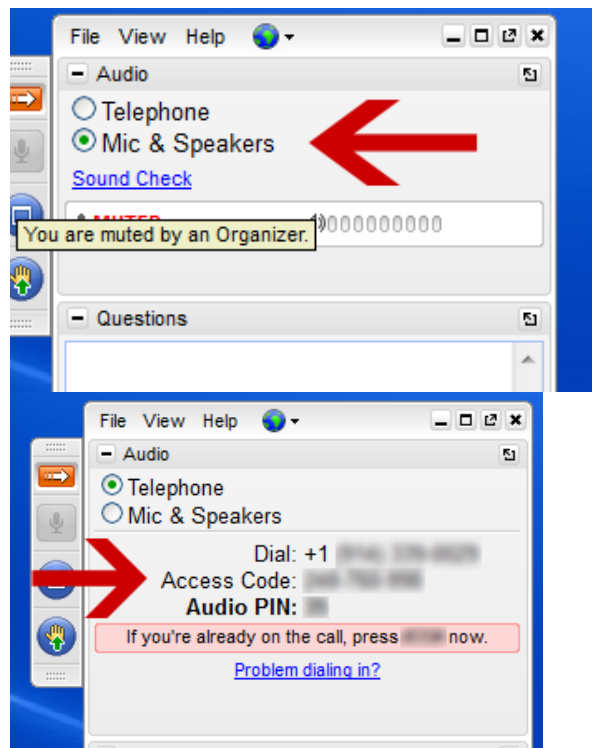
Technical Support

- Webinar **recording and slides** will be **emailed** to participants and **posted** on the CGS website
- Please **submit questions** through the GoToWebinar **control panel** 
- **Technical issues?** Troubleshooting help available:
http://support.citrixonline.com/en_US/webinar



Audio Troubleshooting

- Having trouble hearing us? Try switching to a different audio connection. You can change from Telephone to Mic & Speakers or vice versa without leaving the session.
- If you experience trouble with a telephone connection, click “Problem dialing in?” for an alternate phone number to dial.





Julia Kent
Vice President, Best Practices and Strategic Initiatives
CGS

Hello! 🖐️



Ryan Munson
Graduate Engagement Manager



Jennifer Spirer
Senior Associate Director of
Admissions and Recruitment
Carnegie Mellon University
Information Networking Institute

Agenda

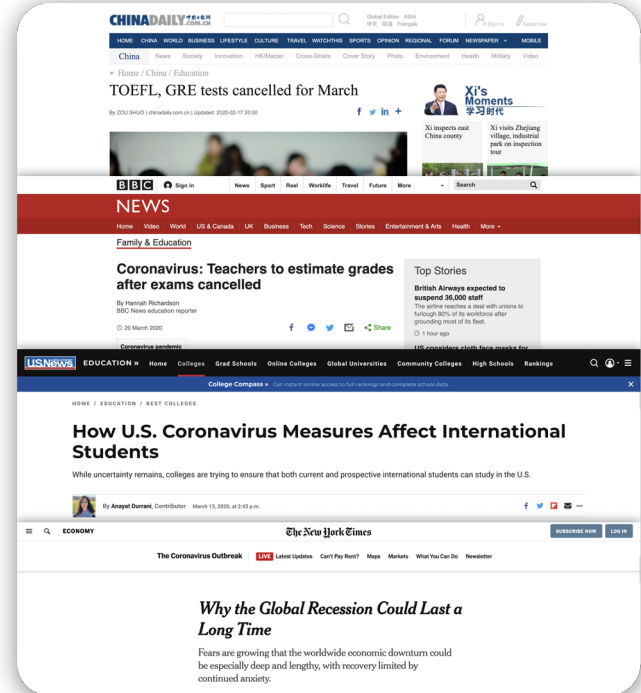
- International Education in 2021
- Duolingo English Test background and overview
- Research and Security
- Use in practice
- Q&A

International Education in 2021



What we are hearing

- App completion
- Secondary school exit requirements
- Visas
- Travel restrictions
- Economic distress



What we're seeing

+1,500

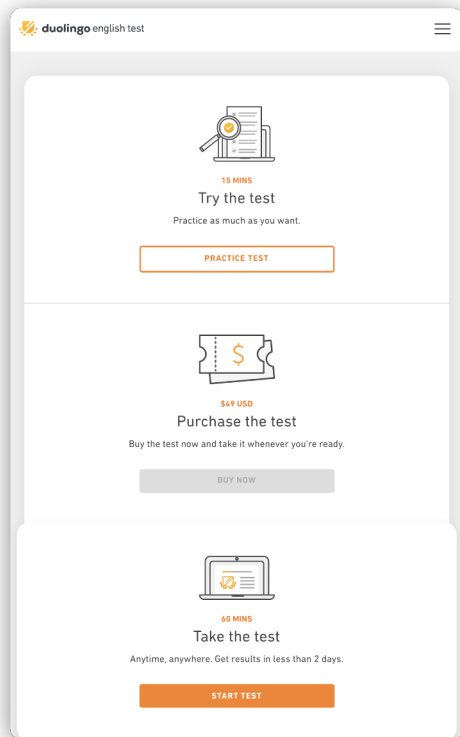
New accepting institutions in 2020. Over **2,500** programs now accept the Duolingo English Test

+1,500%

YTD increase in test volume. Large increases coming from China, Korea, Brazil, Mexico, India, Kazakhstan

How we are responding

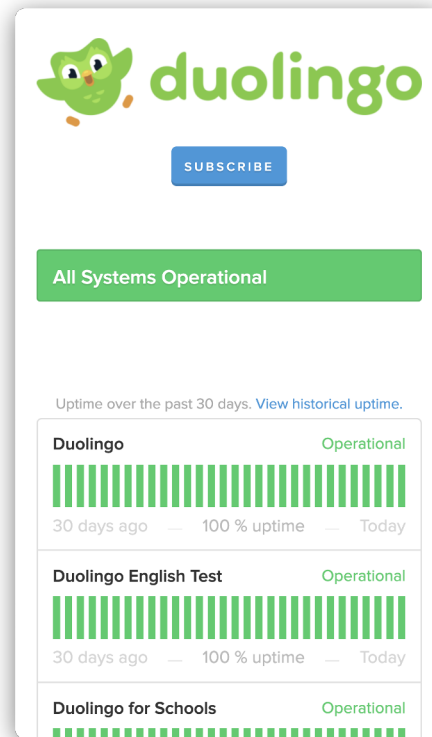
Unlimited test “seats”



Around-the-clock support



100% uptime



Duolingo Background

Duolingo's mission is to make education accessible to all!



Language learning app

- 30+ distinct languages
- iTunes + Google Play “App of the Year”
- Most downloaded education app
- Over 300 million users

Behind the Mission

Luis von Ahn, PhD: Duolingo's co-founder and CEO



- Born and raised in Guatemala
- Inventor of CAPTCHA, reCAPTCHA
- Former CS professor at CMU
- MacArthur Fellow
- Lemelson-MIT Recipient

Accessible

No test centers. No appointments.
Accessible online, on demand
for only \$49 with unlimited sharing,
and completely free for schools.

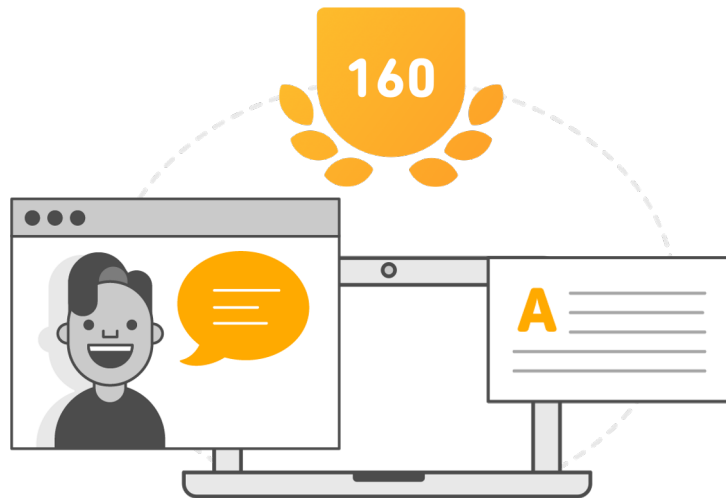


Accessible

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

Comprehensive

Measures speaking, writing,
reading, and listening.
Integrates a proficiency score,
video interview, and writing sample.



Fast



Take a test in ~1 hour, and students receive results in under 2 days.

Secure

Rigorous security protocols,
individual test proctoring,
and computer vision + data forensics
help prevent and detect rule breaking.



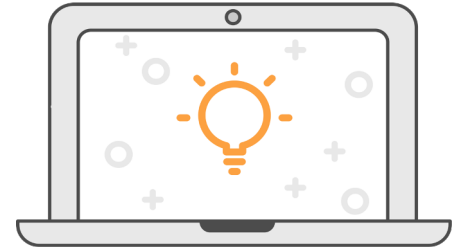
Accurate



Reliability measures and correlation with other tests support the use of test for university admissions.

New model, different format

- Machine-learning supported test item pool
- Administered and scored efficiently online
- Highly predictive of English skills



Efficient and predictive

Item Format	Skills	References
Yes/No (text)	L,R,W	Milton (2010), Staehr (2008) Zimmerman, Broder, Shaughnessy, and Underwood (1977)
Yes/No (audio)	L,S	Milton (2010), Milton, Wade, and Hopkins (2010)
C-Test	R,W	Khodadady (2014), Klein-Braley (1997) Reichert, Keller, and Martin (2010)
Dictation	L,W	Bradlow and Bent (2008, 2002)
Elicited Speech	R,S	Jessop, Suzuki, and Tomita (2007), Van Moere (2012), Vinther (2002)
Open-ended Speaking	S	Luoma (2004)
Open-ended Writing	W	Cushing-Weigle (2002)

5:00

Write for 3 to 5 minutes about the topic below

Agree or disagree with this statement:
"New technologies improve lives." Explain
your reasoning.

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

SUBMIT

3:00

Speak for 1 to 3 minutes about the topic below

Choose a plant or animal that you have learned
about and that you think is interesting. Describe
some of the challenges or dangers the organism
faces and explain how it has adapted for survival.



RECORDING...



NEXT

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

Research and security



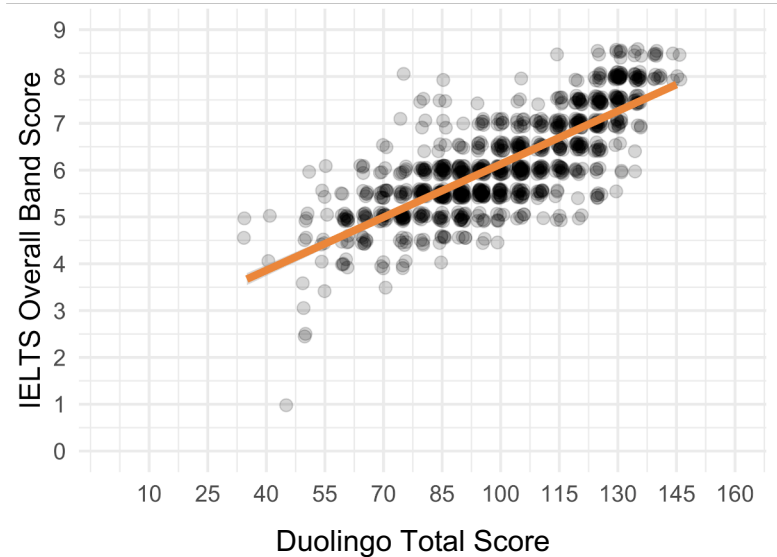
CORRELATION

Relationship to other tests

Relationship	r	n
*TOEFL ↔ IELTS	.73	1,153
Duolingo ↔ IELTS	.78	991

*ETS, 2010

Duolingo ↔ IELTS



CORRELATION

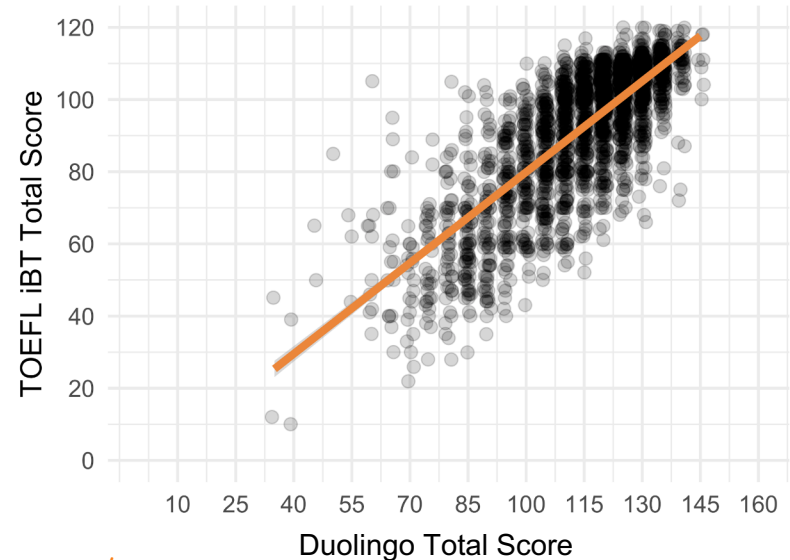
Relationship to other tests

Relationship	r	n
*TOEFL ↔ IELTS	.73	1,153
Duolingo ↔ IELTS	.78	991
Duolingo ↔ TOEFL	.77	2,319

*ETS, 2010

englishtest.duolingo.com/scores

Duolingo ↔ TOEFL



SCORES

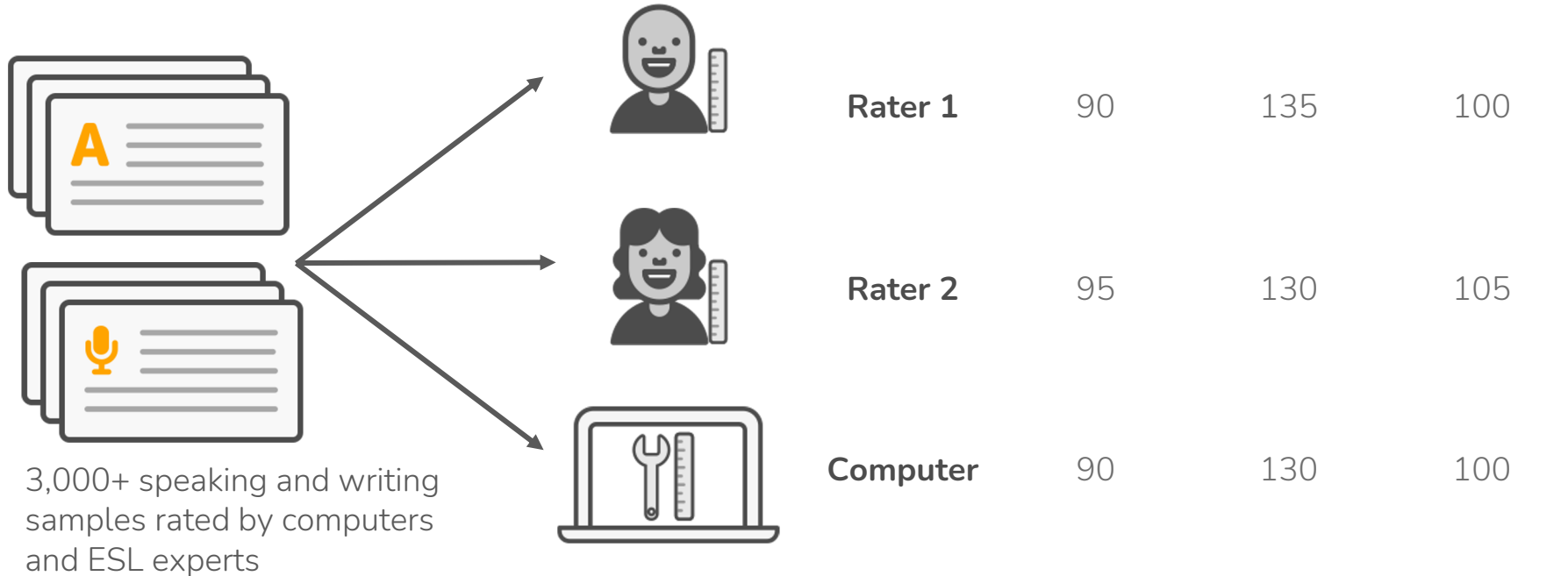
Scale and concordance

Scale	CEFR	Descriptors
10 55	A1 A2	<ul style="list-style-type: none"> Can understand very basic English words and phrases. Can understand straightforward information and express themselves in familiar contexts.
60 85	B1	<ul style="list-style-type: none"> Can understand the main points of concrete speech or writing on routine matters such as work and school. Can describe experiences, ambitions, opinions, and plans, although with some awkwardness or hesitation.
90 115	B2	<ul style="list-style-type: none"> Can fulfill most communication goals, even on unfamiliar topics. Can understand the main ideas of both concrete and abstract writing. Can interact with proficient speakers fairly easily.
120 160	C1 C2	<ul style="list-style-type: none"> Can understand a variety of demanding written and spoken language including some specialized language use situations. Can grasp implicit, figurative, pragmatic, and idiomatic language. Can use language flexibly and effectively for most social, academic, and professional purposes.

Duolingo	TOEFL
90	62-67
95	68-73
100	74-79
105	80-85
110	86-91
115	92-96
120	97-102
125	103-107
130	108-112
135	113-115
140	116-118
145	119

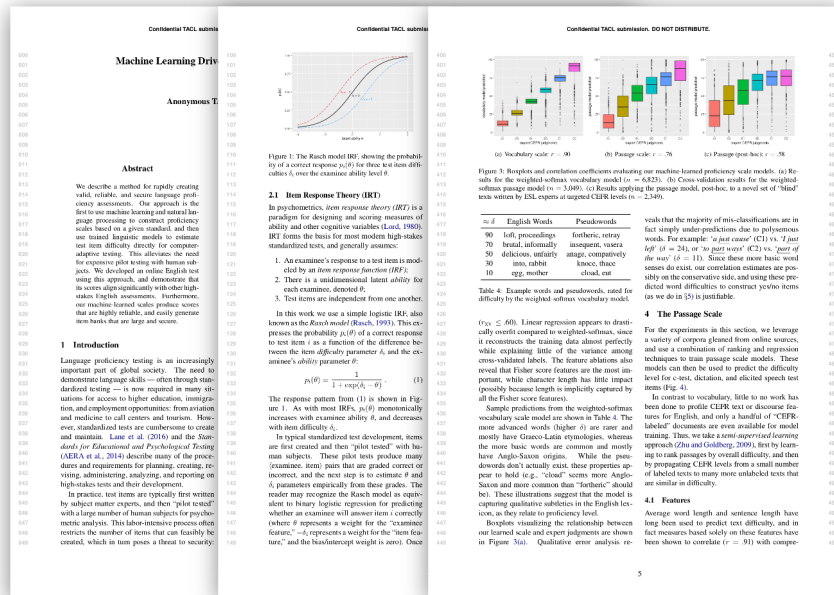
SCORING

Evaluating ML Scoring



- Peer-reviewed paper
Transactions of the Association for Computational Linguistics
Vol. 8, 2020

- Provides evidence supporting the ML-facilitated design of the computer adaptive test (CAT) items



RESEARCH

Psychometric Modeling

- Collaborated with Gunter Maris at ACTNext for external psychometric research
- Evaluated how well our data fit a psychometric model
- Investigated item pool for bias: no substantial bias exists

englishtest.duolingo.com/research

The Duolingo English Test: Psychometric considerations

Gunter Maris*

Abstract

The Duolingo English Test is a computer adaptive test of English language proficiency. The Duolingo English Test is proposed, and its formal characteristics are derived. Statistical properties, its proposed and applied to data from the Duolingo English Test with respect to item and person characteristics. The results indicate that there is no evidence of statistical bias toward different groups of test

This communication is part of the Duolingo English Test external research

Introduction

The Duolingo English Test is a large-scale, high-stakes, online test of English language proficiency that can be delivered anywhere and at any time. This approach to language testing inherently comes with a cold start problem: one needs a large item bank, calibrated to an appropriate item response theory (IRT) model before one can start administering tests. The traditional solution to this problem, large scale pre-testing, not only expensive but suffers from a number of drawbacks. One needs large numbers of test takers to respond to items under realistic testing conditions. Both creating realistic testing conditions and getting test takers from the appropriate populations are challenging.

To support any test anytime testing, the Duolingo English Test is assembled uniquely for every test taker, in the form of a computer adaptive test (CAT). For the Duolingo English Test an alternative solution to the cold start problem is pursued: items are automatically generated, using machine learning and natural language processing algorithms, to match the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2018), with item difficulty parameters that are produced using machine learning. Item responses are automatically scored on a continuous zero to one scale. The development and scoring of the Duolingo English Test are documented in Serfaty, Laflair, and Haghighi (2019) and Laflair and Serfaty (2019). The Duolingo English Test contains 10 item types, five of which are delivered in the CAT administration. These five item types include a short, audio-voice vocabulary test, a picture description, a picture and audio description, a text-based prompt, and an audio prompt. The writing

tasks in the Duolingo English Test are four types: short-answer, multiple-choice, fill-in-the-blank, and short-answer. A CAT model is used to estimate item difficulty parameters. A CAT model is used to estimate item difficulty parameters. A CAT model is used to estimate item difficulty parameters.

A Model

The Duolingo English Test is a computer adaptive test of English language proficiency. The Duolingo English Test is proposed, and its formal characteristics are derived. Statistical properties, its proposed and applied to data from the Duolingo English Test with respect to item and person characteristics. The results indicate that there is no evidence of statistical bias toward different groups of test

Conclusion

The Duolingo English Test is a computer adaptive test of English language proficiency. The Duolingo English Test is proposed, and its formal characteristics are derived. Statistical properties, its proposed and applied to data from the Duolingo English Test with respect to item and person characteristics. The results indicate that there is no evidence of statistical bias toward different groups of test

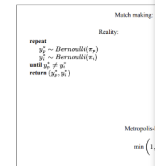


Figure 2. First two steps of the dynamic expansion of continuous responses into conditionally independent binary response variables.

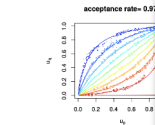


Figure 3. Urnings rating system.

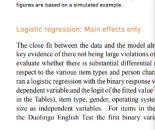


Figure 4. The left panel shows contours for the probability of a correct response, with person abilities on the horizontal axis and the number of correct responses on the vertical axis. The right panel shows the 95% coverage ellipse for the probability of a correct response, with person abilities on the horizontal axis and the number of correct responses on the vertical axis.

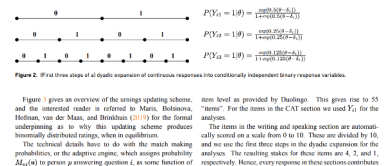


Figure 5. Overview of the urnings updating scheme, and the item-level model.

The urnings rating system is a computer adaptive test of English language proficiency. The urnings rating system is proposed, and its formal characteristics are derived. Statistical properties, its proposed and applied to data from the urnings rating system with respect to item and person characteristics. The results indicate that there is no evidence of statistical bias toward different groups of test takers. The urnings rating system is a computer adaptive test of English language proficiency. The urnings rating system is proposed, and its formal characteristics are derived. Statistical properties, its proposed and applied to data from the urnings rating system with respect to item and person characteristics. The results indicate that there is no evidence of statistical bias toward different groups of test takers.

The urnings rating system is a computer adaptive test of English language proficiency. The urnings rating system is proposed, and its formal characteristics are derived. Statistical properties, its proposed and applied to data from the urnings rating system with respect to item and person characteristics. The results indicate that there is no evidence of statistical bias toward different groups of test takers. The urnings rating system is a computer adaptive test of English language proficiency. The urnings rating system is proposed, and its formal characteristics are derived. Statistical properties, its proposed and applied to data from the urnings rating system with respect to item and person characteristics. The results indicate that there is no evidence of statistical bias toward different groups of test takers.

DELIVERING

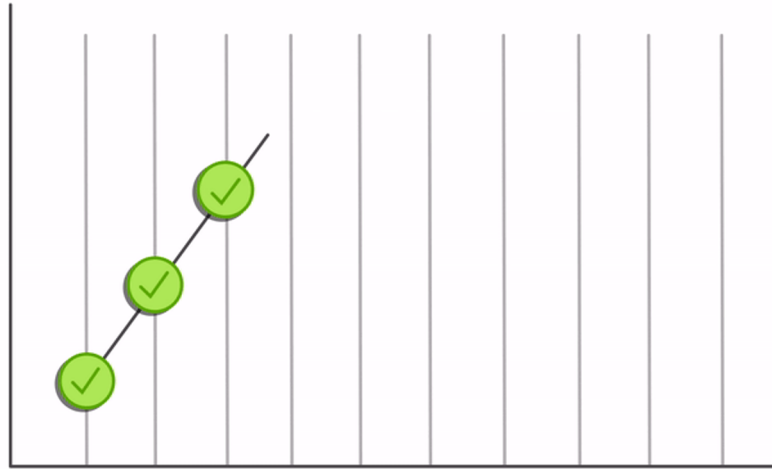
Test Administration

- Computer adaptive test ensures virtually unique test experience
- Draws from an extremely large question pool
 - Questions generated by language assessment experts and machine learning
- Monitor test item exposure and retire questions as necessary



DELIVERING

Computer Adaptive Testing



Note: item types are equally represented in the test, to represent all skills
Each format is shown 3–7 times (6 on average)

DELIVERING

Security: item exposure rate

- % tests that use a given item

(Way, 1998)

0.10%

(~1 in 1000 tests)

- % items shared between any two test administrations

(Chen et. al, 2003)

0.43%

(< 1 in 100 items shared)

Most high-stakes CATs are significantly higher
(e.g., GRE is 14%)

DELIVERING

Test Item Security

1000+

Number of tests a student would
need to take in order to see a single
repeated question

~41 Years

Time for a test taker to see a single
repeated question
...by then, it would be
long retired

SECURING

Certification Process



Test Upload

We record every test experience via:
Webcam
Microphone
Screen
Keyboard
Mouse



AI Review

Machine learning, biometrics authentication, and computer vision analyze the test for irregular response patterns and suspicious behavior to help identify rule breaking and generate a “risk index”



Proctoring Round 1

Independent, anonymous proctors are assigned to the test session and are trained to look for 75 different behaviors such as eye movement and background noise



Proctoring Round 2

A second independent, anonymous proctor reviews the same session for over 75 different behaviors. Tier 2 is comprised of professional, certified ESL or English language educators



Proctoring Escalation

As needed, an additional round of review is added for tests where there is disagreement between the earlier rounds or further examination is needed



Certification Decision

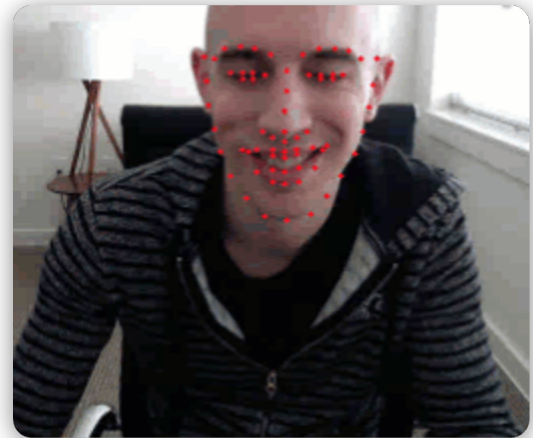
Based on this process, the results are certified or not certified

SECURING

Computer Vision and Biometrics

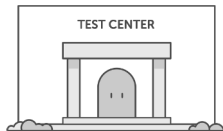
We use computer vision and biometrics to help verify examinee identity and ensure test integrity

- Face detection & ID verification
- Voice recognition
- Keystroke/mouse biometrics
- Eye tracking, lip syncing



SECURING

Proctor : Test Taker Ratio



1:25

- Varies depending on the test and center, but on average, 1 proctor supervises over 25 test takers

X

- Proctors are often contracted locally so they are vulnerable to bias or conflict of interest

X



2+:1

- Since there are multiple independent rounds of proctoring, each test session is individually reviewed by two, and sometimes three, different proctors

✓

- Proctors are anonymous and have no interaction with test takers

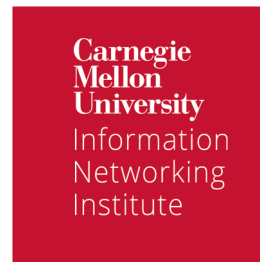
✓

In practice



Graduate Education Use Case: CMU Information Networking Institute

- Networking, Security and Mobile & IoT Engineering
- Slate CRM
- Holistic, Faculty-led Application Review



1400 apps

85% International
Fall 2021 Cycle

Graduate Education Use Case:

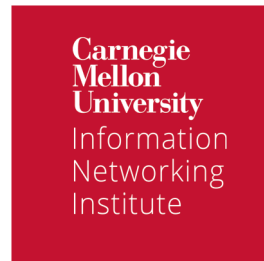
CMU Information Networking Institute

- Improve Accessibility
- Increase Flexibility
- Easy to Use

“Duolingo was one of those capabilities I didn't know I wanted until I had it.

As soon as I'd used it for a few students I found myself asking for it time and again for students who did not have Duolingo scores available but showed possible English language challenges in their essay and free text responses.”

– Faculty Reviewer



College of Engineering
ADVANCED COLLABORATION®

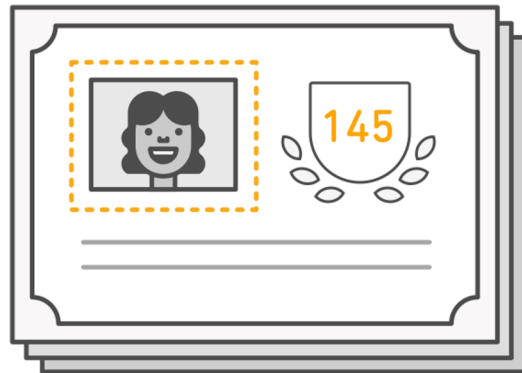


505

Submissions in Fall 2021 Cycle

Receiving Results

- **Dashboard:** Log in to your Duolingo account to see results immediately.
englishtest.duolingo.com/dashboard
- **Slate Integration:** Plug-and-play integration that pulls results daily.
- **Other CRMs:** Public API can be configured by your CRM administrator




DASHBOARD

Receiving Results

RESULTS PRIOR TEST INVITES FILTER EXPORT

Surname(s)	Given Name(s)	Test Taken	Score Sent	Score	Interview	Certificate
Kachmar	Pavlo	Jul 15, 2019, 12:24 PM	Jul 17, 2019, 10:29 AM	125	VIEW	VIEW ▼
Ogita	Malachi	Jul 15, 2019, 12:14 PM	Jul 17, 2019, 10:29 AM	125	VIEW	VIEW ▼
mandanka	tanvi	Jul 15, 2019, 12:28 PM	Jul 17, 2019, 10:29 AM	100	VIEW	VIEW ▼
Gito	Christian	Jul 15, 2019, 11:57 AM	Jul 17, 2019, 10:29 AM	125	VIEW	VIEW ▲



Email: christiangito.freelancer@gmail.com **DOB:** 1989-10-28 **Coupon:** VERSIONTHREE

Add notes... (Max 255 characters.)

DASHBOARD

Receiving Results



Li, David
September 15, 2019

VIDEO INTERVIEW

Describe the last place you visited with your family or close friends. What did you like most about this place? How could the place be improved for the people who visit there?

< 1 of 2 >



< 2 of 2 >

RESPONSE

Money isn't always the only motivation for having a job. Other important reasons include the desire to be productive and to contribute to society. Many people truly enjoy utilizing their skills and serving people who need their assistance. I would keep a job even if I didn't need one to live so that I could remain connected to the world.

SLATE

Receiving Results

The screenshot displays the SLATE application interface. On the left is a sidebar with navigation links: Dashboard, Application, Transcript, Recommend..., Essay, Portfolio, and Duolingo. The main content area shows a user profile for '123456 Loh Herng, Kai' with a search bar and a menu icon. Below this is a video interview window titled 'Kai Loh Herng' dated 'July 16, 2019'. The window contains a 'VIDEO INTERVIEW' section with the prompt: 'Describe some good opportunities for volunteering in your community. Do you volunteer in your community? Why or why not?'. To the right of the text is a video player showing a man speaking, with a play button and a 10-second rewind icon. At the bottom of the window is a pagination indicator '< 1 of 2 >'. A blue circular badge with the text 'slate SILVER PARTNER' is overlaid on the right side of the video player.

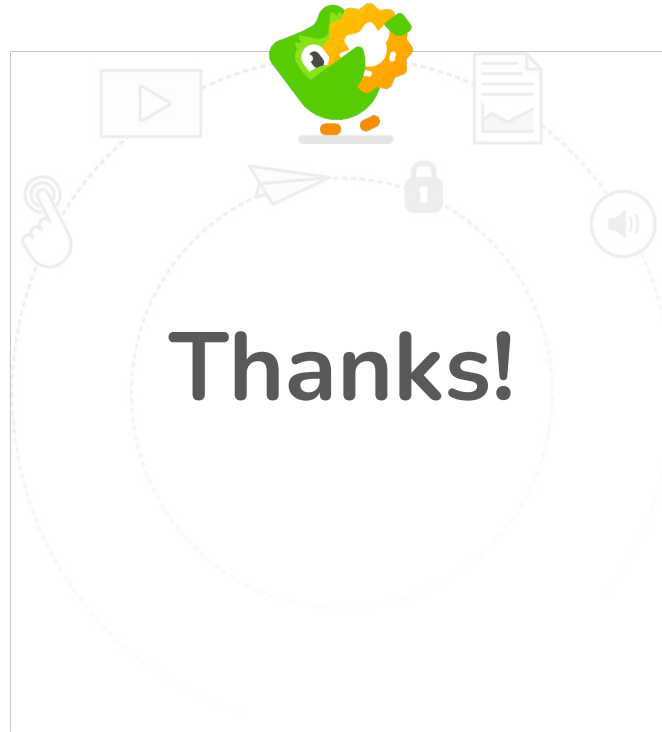
Start receiving results!

- Becoming an accepting institution is easy, fast, and free
- Take 2 minutes: go.duolingo.com/receiverresults
- Receive dashboard access within 1 business day
- Getting started guide, email templates, social media assets

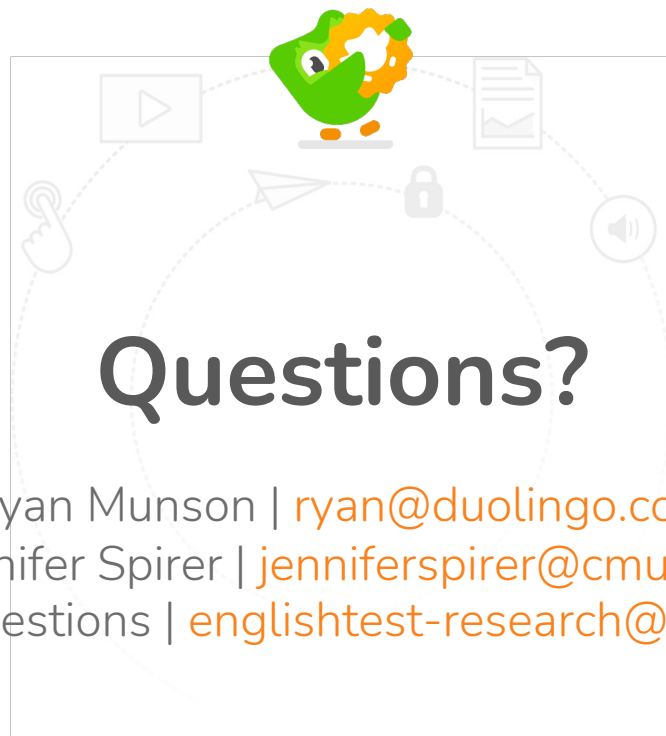




duolingo
english test



**Carnegie
Mellon
University**
Information
Networking
Institute



Ryan Munson | ryan@duolingo.com

Jennifer Spirer | jenniferspirer@cmu.edu

Research Questions | englishtest-research@duolingo.com

Upcoming Events

Supporting Graduate Student Mental Health and Wellbeing Report Release

- April 29, 2021
- Webinar from 2:00-3:00 PM ET

For other upcoming CGS webinars as well as the links to recent webinar recordings visit <https://cgsnet.org/cgs-webinars>