Hello! 🙋

Ryan Munson  
Strategic Engagement Manager

Geoff Laflair  
Senior Assessment Scientist
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 400 million users
Behind the Mission

Luis von Ahn Ph.D., 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
Test Centers vs Internet Access

1 Test Center per 1,926,260 people  

1 Internet user per 2.2 people
## English Proficiency Options

<table>
<thead>
<tr>
<th></th>
<th>TOEFL</th>
<th>IELTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost</strong></td>
<td>$200</td>
<td>$215</td>
<td>$49</td>
</tr>
<tr>
<td><strong>Locations</strong></td>
<td>4500</td>
<td>1100</td>
<td>Unlimited</td>
</tr>
<tr>
<td><strong>Test Length</strong></td>
<td>4 Hours</td>
<td>3 Hours</td>
<td>&lt; 1 Hour</td>
</tr>
<tr>
<td><strong>Turnaround</strong></td>
<td>14 Days</td>
<td>13 Days</td>
<td>2 Days</td>
</tr>
<tr>
<td><strong>Sharing</strong></td>
<td>5 Included</td>
<td>4 Included</td>
<td>Unlimited</td>
</tr>
</tbody>
</table>
International Education Today

International enrollment in the US declining

Certification is another barrier for students

Traditional testing options have limited value for institutions
Test Overview
Accessible

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

Take a test in under 60 minutes, and students receive results in under 48 hours.
Comprehensive

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

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

Rigorous security protocols, individual test proctoring, and computer vision + biometrics help prevent and detect rule breaking.
## Test Process

<table>
<thead>
<tr>
<th>TESTING</th>
<th>PROCTORING</th>
<th>SHARING</th>
<th>REVIEWING</th>
</tr>
</thead>
<tbody>
<tr>
<td>45-60 minutes</td>
<td>24-48 hours</td>
<td>On Demand</td>
<td>Anytime</td>
</tr>
<tr>
<td><strong>Students</strong></td>
<td><strong>Duolingo</strong></td>
<td><strong>Students</strong></td>
<td><strong>Institutions</strong></td>
</tr>
<tr>
<td>The test taker takes the test online whenever and wherever is convenient.</td>
<td>Test sessions are reviewed via AI and by remote human proctors. Certified results are then sent to the test taker.</td>
<td>Test takers can send certified results wherever and whenever they want, at no extra cost.</td>
<td>Results shared with you show up on the dashboard immediately and in Slate within about 24 hours.</td>
</tr>
</tbody>
</table>
Accepting Institutions - 2 year

- Houston Community College (HCC)
- Sinclair Community College
- Orange Coast College
- Pasadena City College
- Citrus College
- Foothill College
- De Anza College
- El Camino College
- Irvine Valley College
- Miami Dade College
- Mohawk Valley Community College
- Saddleback College
- Kapi‘olani Community College
- Lane Community College
- Mt. San Antonio College
Accepting Institutions - 4 year

Georgia Tech  Bard  Emory University  Duke University

Vanderbilt University  Yale  University of Kentucky  Texas Tech University

Columbia University in the City of New York  University of Oregon  Iowa State University

Barnard  University of Notre Dame  Northeastern University

Over 500 more at englishtest.duolingo.com/institutions
Accepting Institutions - Graduate

200% growth in graduate school/program acceptance YTD
Test format
“Yes/No” Vocabulary Items

- Distinguish English **words** from English-like **pseudowords** (Zimmerman et al., 1977)

- **Written variant** predicts reading, writing, and listening skills (Milton, 2010; Staehr, 2008)

- **Audio variant** predicts listening and speaking skills (Milton, 2010)
C-test

(Klein-Braley, 1997; Khodadady, 2014)
Measure of reading ability, and to some writing and discourse knowledge

Type the missing letters to complete the text below:

Sam bought a new coat yesterday. He is often cold and he chose a coat that is warmer than his other jacket. His coat is grey and it is good with his red hat. He is going to wear it every day this winter.
Dictation

(Bradlow and Bent, 2002; 2008)
Assesses both listening and writing abilities by transcribing an audio recording
Elicited Speech

(Vinther, 2002; Jessop+al., 2007)
Assesses reading and speaking ability by reading a passage out loud.
Scored, Open-ended Speaking and Writing

1:32

Respond to the questions in at least 98 words.

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

Consider the subtext: Are there any terrible creatures beneath the lovely surface?

Words: 98

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?
Item Development
Generating the test scale

- Common European Framework of Reference
- Duolingo Scale
Item Development

- Expert-annotated CEFR texts & wordlists
- Natural English language sources
- Statistical models for CEFR difficulty estimation
- Item-generation techniques
- Large pool of CEFR-aligned items for CAT administration
## Generating the test scale

<table>
<thead>
<tr>
<th>Value</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>160</td>
<td>“I believe his thoughts were nevertheless more rich and original.”</td>
</tr>
<tr>
<td>130</td>
<td>“What would have happened, had you not helped them?”</td>
</tr>
<tr>
<td>100</td>
<td>“Why have you not published anything yet?”</td>
</tr>
<tr>
<td>70</td>
<td>“We just need a little patience.”</td>
</tr>
<tr>
<td>40</td>
<td>“I was born in 1986.”</td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

- **Common European Framework of Reference**
- **Duolingo Scale**
- **Level estimation via ML/NLP**
Relationship with CEFR

(5-fold cross-validation; wordlists compiled from Capel, 2010/2012; North et al., 2011; other sources)
Test Delivery
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
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)
Security: item exposure rate

- % tests that use a given item (Way, 1998)
- Most high-stakes CATs are significantly higher than ours (e.g., GRE is 14%)

0.10%
(~1 in 1000 tests)
% items shared between any two test administrations (Chen et. al, 2003)

We haven’t found any other high-stakes CATs that disclose their overlap rates publicly

0.43% ( < 1 in 100 items shared)
In Press Peer-Reviewed Publication

Transactions for the Association of Computational Linguistics

April 2020
Validity
Validity

- Appropriate scoring procedures
- Consistent test scores
- Relationships to other measures
- Representative language
- Related to/predictive of other measures
Open-ended Speaking and Writing

- Open-ended, open response speaking and writing in addition to short-form speaking and writing on the computer adaptive testing portion

- Machine-graded speaking and writing
  - Consistent evaluation of linguistic features of the responses
  - High rates of agreement with human raters
Evaluating ML Scoring

How people score

- Content
- Sophistication
- Accuracy | fluency | acoustic

How the machine scores

- Task relevance
- Grammatical complexity
- Lexical sophistication
- Lexical diversity
- Grammatical accuracy
- Length | fluency | acoustic
Evaluating ML Scoring

3,000+ speaking and writing samples rated by computers and ESL experts

<table>
<thead>
<tr>
<th></th>
<th>Student A</th>
<th>Student B</th>
<th>Student C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>90</td>
<td>135</td>
<td>100</td>
</tr>
<tr>
<td>Rater 2</td>
<td>95</td>
<td>130</td>
<td>105</td>
</tr>
<tr>
<td>Computer</td>
<td>90</td>
<td>130</td>
<td>100</td>
</tr>
</tbody>
</table>
Evaluating ML Scoring

Speaking (n = 3,966)

\[ H: H_K = 0.68 \]
\[ H: C_K = 0.82 \]

Writing (n = 3,626)

\[ H: H_K = 0.77 \]
\[ H: C_K = 0.79 \]
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
Internal Consistency

- Split-half
  - $r = 0.91$

- Split-half adjusted
  - $r = 0.95$

($n = 27,707$)
Test-Retest Reliability

- $r = 0.85$
- $p < 0.001$

$(n > 2,000)$
Standard Error of Measure

- Split-half = 6.75
- Test-retest = 8.71
- All within +/- 2 units (previously 5 units)

\[ SEM = SD \times \sqrt{1 - \alpha} \]
Relationship with other tests

<table>
<thead>
<tr>
<th>Relationship</th>
<th>$r$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL ↔ IELTS</td>
<td>.73</td>
<td>1,153</td>
</tr>
<tr>
<td>DET ↔ IELTS</td>
<td></td>
<td>991</td>
</tr>
</tbody>
</table>

*ETS, 2010

Correlation

**Predictive**
## Relationship with other tests

<table>
<thead>
<tr>
<th>Relationship</th>
<th>r</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>*TOEFL ↔ IELTS</td>
<td>.73</td>
<td>1,153</td>
</tr>
<tr>
<td>DET ↔ IELTS</td>
<td></td>
<td>991</td>
</tr>
<tr>
<td>DET ↔ TOEFL</td>
<td>.77</td>
<td>2,319</td>
</tr>
</tbody>
</table>

*ETS, 2010*
Relationship with EAP Faculty

(Ishikawa et al., 2016)

<table>
<thead>
<tr>
<th>Faculty Assessment</th>
<th>2014 Cohort (n = 31)</th>
<th></th>
<th>2015 Cohort (n = 46)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DET</td>
<td>TOEFL</td>
<td>Writing</td>
<td>Speaking</td>
</tr>
<tr>
<td>Written</td>
<td>0.69***</td>
<td>0.12</td>
<td>0.47*</td>
<td>—</td>
</tr>
<tr>
<td>Oral Comprehensibility</td>
<td>0.40*</td>
<td>0.27</td>
<td>—</td>
<td>0.37</td>
</tr>
<tr>
<td>Oral Fluency</td>
<td>0.45*</td>
<td>-0.09</td>
<td>—</td>
<td>0.13</td>
</tr>
<tr>
<td>Oral Pronunciation</td>
<td>0.45*</td>
<td>0.13</td>
<td>—</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Note: *** p < 0.001  ** p < 0.01  * p < 0.05
Internal Research

enlishtest.duolingo.com/research
Technical Advisory Board

Dr. Alina Von Davier
- Senior Vice President of ACTNext
- Pioneer in computational psychometrics
- Formerly, Senior Research Director at ETS
- Responsible for both the psychometrics and scoring of the TOEFL iBT® and TOEIC®

Dr. Micheline Chalhoub-Deville
- Professor of Educational Research Methodology at UNC at Greensboro
- President of International Language Testing Association (ILTA)
- Formerly, served on the TOEFL® Committee of Examiners, the TOEFL® Policy Board, and the British Council Assessment Board

Dr. Bryan Smith
- Associate Professor of English at the Arizona State University
- Researcher in computer-assisted language learning
- Editor of CALICO Journal, an international journal devoted to research and discussion on technology and language learning

Dr. Stephen Sireci
- Distinguished Professor of Psychometrics and Director of the Center for Educational Assessment at UMass Amherst
- Researcher in educational test development particularly issues of validity, standard setting
- Vice President of the National Council of Measurement in Education (NCME).
Test Security
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
Test Rules

Some of the test rules that students need to follow include:

- No headphones or earbuds
- No looking off screen
- No plugins or accessing outside programs
- Nobody else in the test environment
- No outside devices or materials
Onboarding

Computer with internet access, and webcam

Government issued photo-ID

Under an hour in a quiet, well-lit room
Certification Process

1. Test Upload
2. AI Review
3. Proctoring Tier 1
4. Proctoring Tier 2
5. Proctoring Escalation
6. Certification Decision
Data Forensics

- We use **biometrics and computer vision** to help verify examinee identity and ensure test integrity
  - Face detection & ID verification
  - Voice recognition
  - Keystroke/mouse biometrics
  - Eye tracking, lip syncing
Test Centers vs. Remote Proctoring

- While test centers give proctors a 360 degree view of the test experience, remote proctoring uses different sources to understand the same information and more.

- Remote proctoring also holds some other security advantages...
Proctor : Test Taker Ratio

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

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

- Test takers have direct contact with proctors—making them vulnerable to influence  
  ✗
- Proctors are contracted locally so there can be bias or conflict of interest  
  ✗

- Proctors are anonymous  
  ✔
- Test takers have no contact with them  
  ✔
- Test sessions are randomly assigned to proctors  
  ✔
- Proctoring rounds are independent of each other to eliminate bias or influence  
  ✔
Mode of Review

- Proctors only can supervise the test session while each test taker is onsite ❌
- If a test taker breaks a rule, it can invalidate all tests from a test date ❌
- Asynchronous, digital review allows for multiple rounds of test session review ✔
- Rule breaking by an individual test taker does not impact the validity of other tests ✔
- Security algorithms are trained based on thousands of data points from every test session—each test is more secure than the last ✔
Privacy and Data Security

- Data retention is fully GDPR-compliant
- Data access is restricted to those with a clear need within the company
- Strict cyber security protocols are followed
- All data is encrypted
- All user data is deleted after a five year period
- Only test-taker scores and speaking portions are shared with institutions
Working with Results
<table>
<thead>
<tr>
<th>Scale</th>
<th>CEFR</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>A1</td>
<td>- Can understand very basic English words and phrases.</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>- Can understand straightforward information and express themselves in familiar contexts.</td>
</tr>
<tr>
<td>60</td>
<td>B1</td>
<td>- Can understand the main points of concrete speech or writing on routine matters such as work and school.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Can describe experiences, ambitions, opinions, and plans, although with some awkwardness or hesitation.</td>
</tr>
<tr>
<td>90</td>
<td>B2</td>
<td>- Can fulfill most communication goals, even on unfamiliar topics.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Can understand the main ideas of both concrete and abstract writing.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Can interact with proficient speakers fairly easily.</td>
</tr>
<tr>
<td>120</td>
<td>C1</td>
<td>- Can understand a variety of demanding written and spoken language including some specialized language use situations.</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>- Can grasp implicit, figurative, pragmatic, and idiomatic language.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Can use language flexibly and effectively for most social, academic, and professional purposes.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Duolingo</th>
<th>TOEFL</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>62-67</td>
</tr>
<tr>
<td>95</td>
<td>68-73</td>
</tr>
<tr>
<td>100</td>
<td>74-79</td>
</tr>
<tr>
<td>105</td>
<td>80-85</td>
</tr>
<tr>
<td>110</td>
<td>86-91</td>
</tr>
<tr>
<td>115</td>
<td>92-96</td>
</tr>
<tr>
<td>120</td>
<td>97-102</td>
</tr>
<tr>
<td>125</td>
<td>103-107</td>
</tr>
<tr>
<td>130</td>
<td>108-112</td>
</tr>
<tr>
<td>135</td>
<td>113-115</td>
</tr>
<tr>
<td>140</td>
<td>116-118</td>
</tr>
<tr>
<td>145</td>
<td>119</td>
</tr>
</tbody>
</table>
Receiving Results

- Dashboard: Log in to your Duolingo account to see results immediately.
- Slate Integration: Plug-and-play integration that pulls results daily.
- API Integration: Flexible way to integrate with tools other than Slate, like TargetX, Salesforce, Banner, Peoplesoft, and more.
## Receiving Results - Dashboard

### Current Results

<table>
<thead>
<tr>
<th>Surname</th>
<th>Given Name</th>
<th>Date Sent</th>
<th>Score</th>
<th>Country</th>
<th>Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kamath</td>
<td>Salum</td>
<td>Jan 7</td>
<td>125</td>
<td>India</td>
<td>VIEW</td>
</tr>
<tr>
<td>Horng Loh</td>
<td>Kai</td>
<td>Jan 6</td>
<td>140</td>
<td>China</td>
<td>VIEW</td>
</tr>
<tr>
<td>Sundar</td>
<td>Andrew</td>
<td>Jan 6</td>
<td>60</td>
<td>India</td>
<td>VIEW</td>
</tr>
<tr>
<td>Torrez</td>
<td>Jose</td>
<td>Jan 6</td>
<td>95</td>
<td>Guatemala</td>
<td>VIEW</td>
</tr>
<tr>
<td>Kim</td>
<td>James</td>
<td>Dec 12, 2018</td>
<td>130</td>
<td>South Korea</td>
<td>VIEW</td>
</tr>
</tbody>
</table>

*These results are from an older version of the test that uses a 100-point scale and does not include long-form speaking/writing. Learn more.*

### Prior Tests

<table>
<thead>
<tr>
<th>Surname</th>
<th>Given Name</th>
<th>Date Sent</th>
<th>Score</th>
<th>Country</th>
<th>Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kamath</td>
<td>Salum</td>
<td>Jan 7</td>
<td>99</td>
<td>India</td>
<td>VIEW</td>
</tr>
<tr>
<td>Horng Loh</td>
<td>Kai</td>
<td>Jan 6</td>
<td>89</td>
<td>China</td>
<td>VIEW</td>
</tr>
</tbody>
</table>

*Email: kaihenglong@yahoo.com  DOB: Sep 12, 1995  Date Taken: Dec 11, 2018*

Add notes...
CRM Integration

- Automatically receive results within admissions CRM via a public API
- Allow results to sync with existing applications
Questions?

Geoff LaFlair | geoff@duolingo.com
Ryan Munson | ryan@duolingo.com
Appendix
Accepting Institutions

- Babson College
- Bard College
- Barnard College
- Baylor University
- Boise State University
- Bowdoin College
- Clark University
- Clarkson University
- Colby College
- Columbia University
- Creighton University
- Dallas Baptist University
- Dartmouth College
- Duke University
- Elon University
- Emerson College
- Emory University
- Fairfield University
- Florida Institute of Technology
- Furman University
- Georgia Institute of Technology
- Goldey-Beacom College
- Houston Community College
- Iowa State University
- Johns Hopkins University
- Jacobs University
- Lafayette College
- Lindenwood University
- Loyola Marymount University
- Macalester College
- Miami Dade College
- Middle Tennessee State University
- Middlebury College
- Milwaukee School of Engineering
- Montana State University
- The New School
- Notre Dame
- Northeastern University
- Northern Arizona University
- Occidental College
- Orange Coast College
- Pasadena City College
- Pepperdine University
- Rhode Island School of Design
- Robert Morris University
- Rollins College
- Santa Clara University
- Sewanee - University of the South
- Stanford University
- Suffolk University
- Tennessee Technological University
- Texas Christian University
- Texas Tech University International
- The University of Chicago
- Trinity University
- Troy University
- Tufts University
- Union College
- University of Akron
- University of Alabama
- University of Alabama at Birmingham
- University of Arizona
- University of Evansville
- University of Kentucky
- University of La Verne
- University of New Hampshire
- University of North Texas
- University of Notre Dame
- University of Oregon
- University of Rhode Island
- University of Tampa
- University of Tulsa
- University of Vermont
- University of Wyoming
- Valparaiso University
- Vanderbilt University
- Washington and Lee University
- Washington University in St. Louis
- Wellesley College
- Wesleyan University
- Whitman College
- Widener University
- Yale University

A comprehensive list available at englishtest.duolingo.com/institutions