Transforming Language Proficiency Testing:

Exploring the innovations of the Duolingo English Test (DET)

James Holden
August 31, 2023

ENGLISH AUSTRALIA
QUALITY | SUPPORT | ASSURANCE

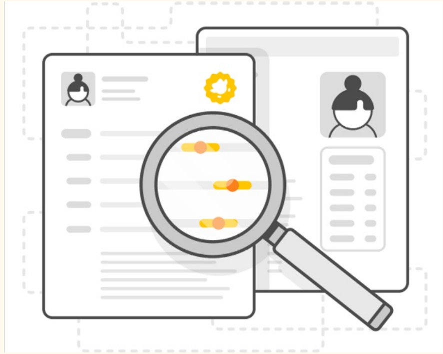
40 YEARS

englishaustralia.com.au

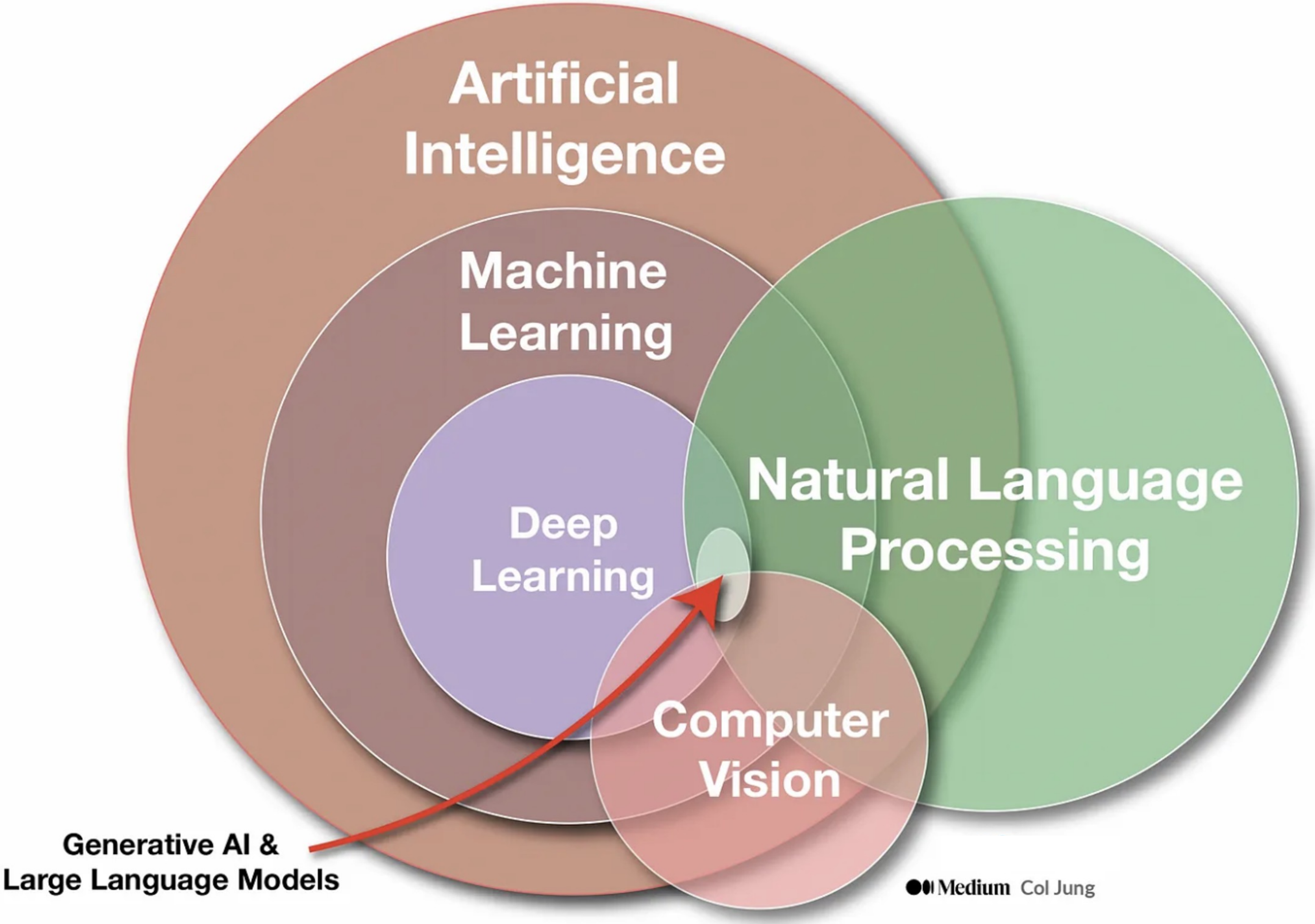
Celebrating 40 years of quality and innovation in ELICOS

Today’s session

1. Duolingo and generative AI
2. AI assisted security/ item creation / test design
3. New concordance / research
4. What does the future hold?



Quick AI overview

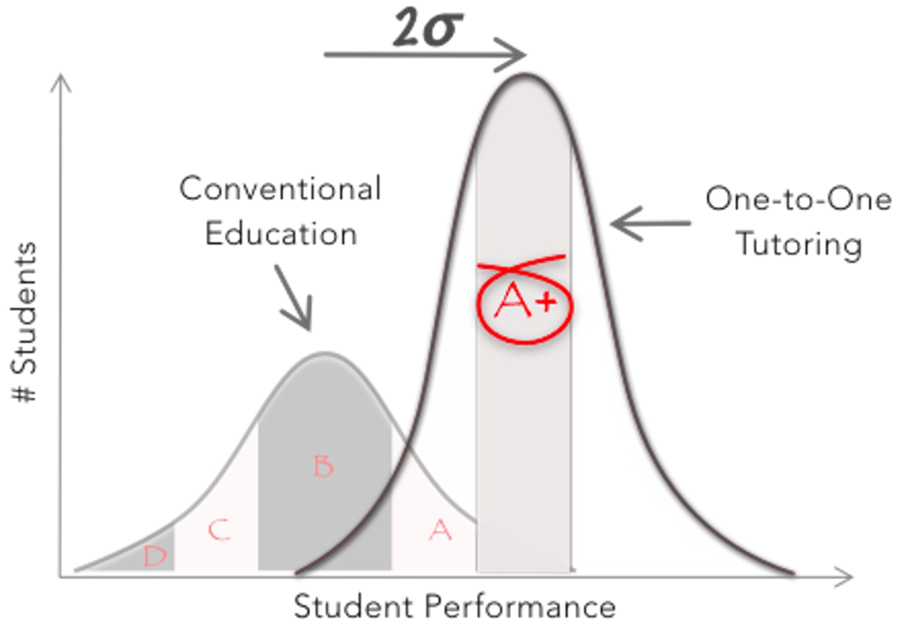


Why is Duolingo excited about AI?

Duolingo's mission is to develop the best education in the world and make it universally available.

- AI offers the best opportunity to achieve this mission and meet students 'where they are'
- AI is solving 'Blooms 2 Sigma Problem'

Blooms 2 Sigma Problem (1984)



- The average student achievement with a 1-1 tutor is better than 98% of students from a traditional class

- How can we deliver 1-1 tutoring against the limitations of time, resources, and the varying needs of students?

Harnessing AI to overcome Blooms 2 Sigma Problem - Testing

Technology enables us to deliver 1-1 English testing against the limitations of time, resources, and the varying needs of students.

In 2016, Duolingo launched the world’s first digital-first high stakes proficiency test, powered by the latest in Artificial Intelligence, assessment science, psychometrics.

The test had a mandate to use technology to achieve 4 key goals:

- Accessible
- Efficient
- Economical
- Psychometric Quality

Harnessing AI to transform language testing

Anonymised, randomised & AI-assisted proctoring

Use AI to build and sustain the largest item bank

Computer Adaptive Test for accuracy and efficiency

duolingo english test

1. AI-assisted proctoring

Problem 1 - The inequity of high-stakes testing

- Traditional testing centre models are not accessible, reliable, affordable or equitable

- English Testing is a high-stakes activity that requires the best security frameworks to uphold the integrity of results.

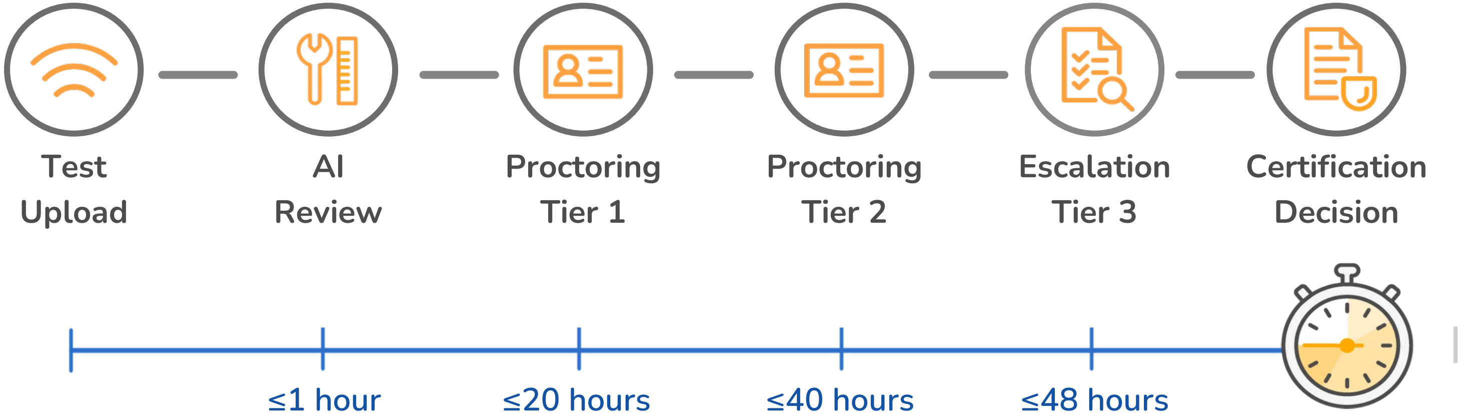
Solution:

- Create a digital test to increase access, while using technology to uphold the necessary high-stakes security requirements.

How to define a digital threat model

Threats	Attacker	Mitigations
Test theft	Scrapers, Cheaters	Large item pool, adaptive engine, Screenshot prevention etc.
Getting third party aid during test	Cheaters, Cheating rings	Monitoring system softwares, preventing window switching, plagiarism etc.
Identity and account abuse	Imposters	ID verification, multiple session and account matching, digital fingerprinting

The certification process



Test Upload

AI Review

Proctoring Tier 1

Proctoring Tier 2

Escalation Tier 3

Certification Decision

≤1 hour

≤20 hours

≤40 hours

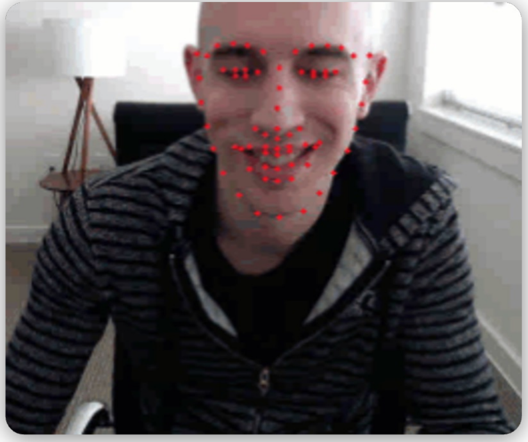
≤48 hours

AI Assistance + Human Proctoring

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

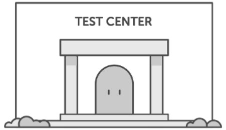
Using the test video, audio, screen recording, keystrokes, mouse movement, and other recorded variables, proctors examine:

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



Digital-first means unique security advantages

- Large item bank overcomes risk of item exposure
- No risk of compromised supply-chains or proctors
- Easily collect test sessions for repeated review
- ML assistance and behaviour database grows with every test taker
- The ratio of proctor to test taker changes from 1-25 > AI + 3-1



vs

duolingo

Confidential

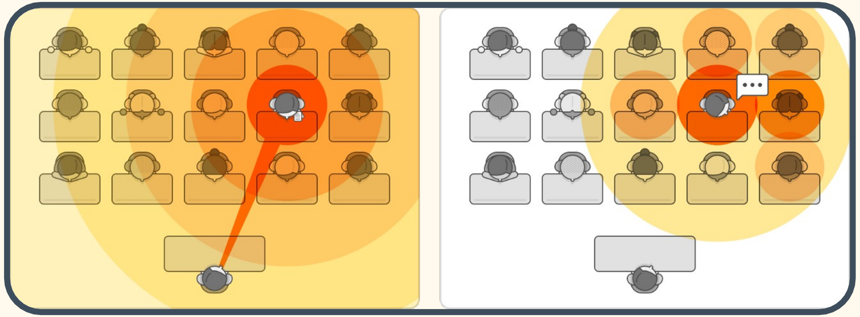
The asynchronous advantage

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

- Synchronous proctoring doesn't allow for multiple rounds of review

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

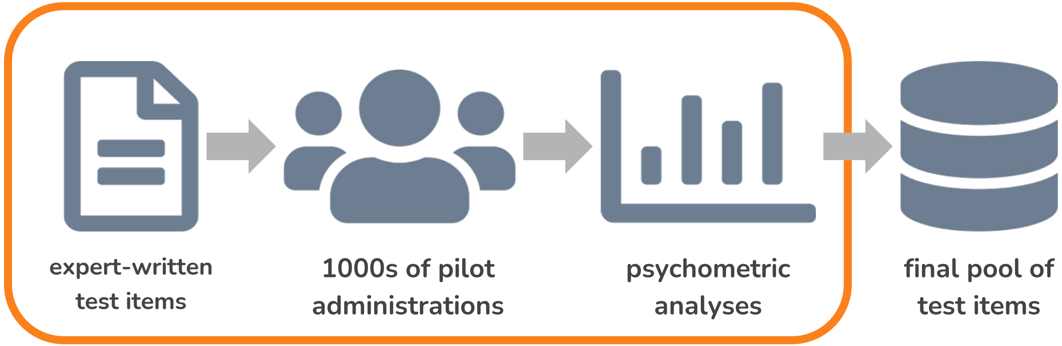
- A single instance of cheating can mean that an entire batch of test takers may be unable to get a valid score.



englishtest.duolingo.com/security

2. Item creation with Human-in-the-loop AI

Problem 2 - Traditional Test Development



expert-written test items

1000s of pilot administrations

psychometric analyses

final pool of test items

Problems with this approach:

- Time-consuming
- Not very secure
- Expensive

Solution:

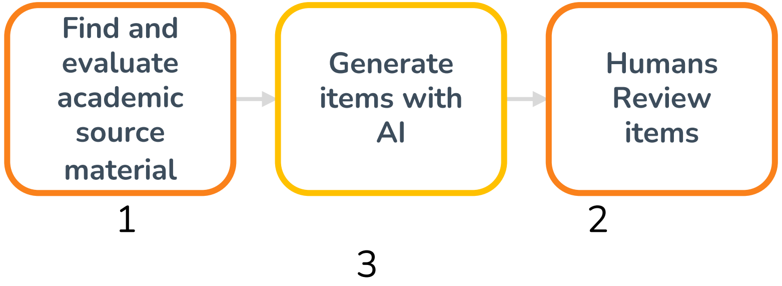
- Item creation with human-in-the-loop AI

Human-In-The-Loop AI

- Goal: Use collaboration of humans and AI to perform a task
    - Maximise the strengths and benefits of **both** humans and AI
    - Have "expert" eyes on all content on the test

- “The Loop”:
    - Use AI systems to make predictions or generate content
    - Have humans review/edit the outputs
    - Use human labels and edits to improve the AI

AI Test Development



Find and evaluate academic source material

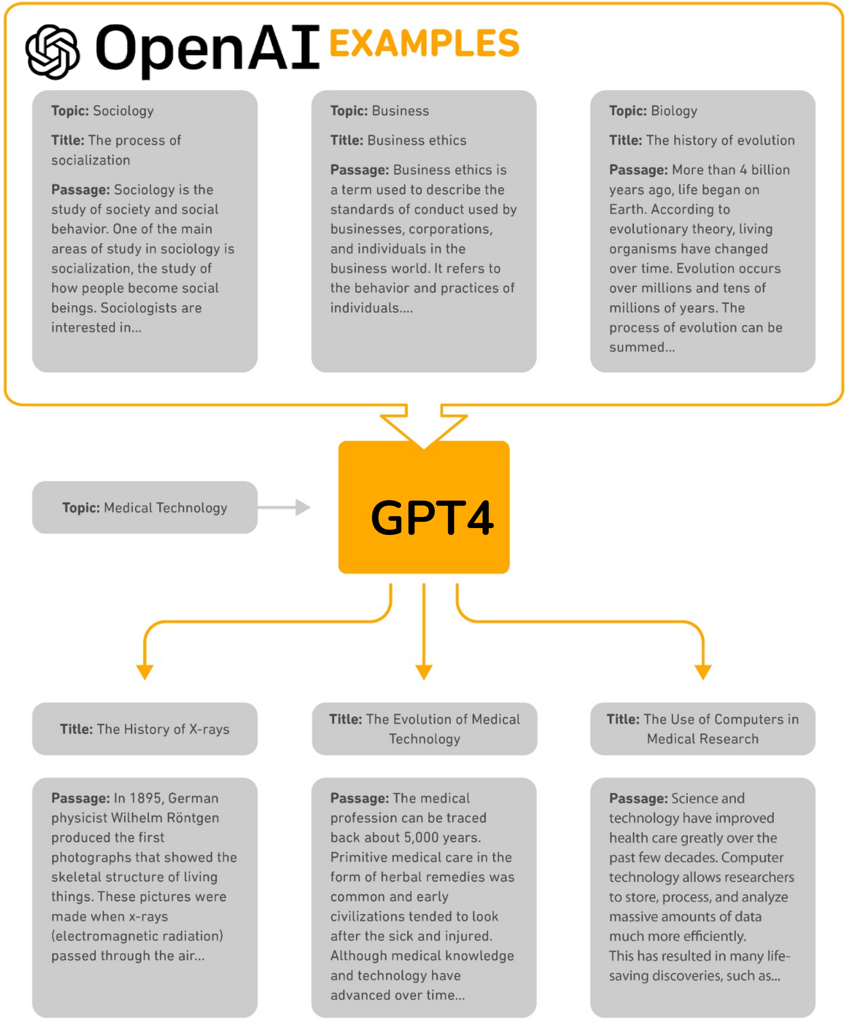
Generate items with AI

Humans Review items

1

3

2



OpenAI EXAMPLES

Topic: Sociology

Title: The process of socialization

Passage: Sociology is the study of society and social behavior. One of the main areas of study in sociology is socialization, the study of how people become social beings. Sociologists are interested in...

Topic: Business

Title: Business ethics

Passage: Business ethics is a term used to describe the standards of conduct used by businesses, corporations, and individuals in the business world. It refers to the behavior and practices of individuals....

Topic: Biology

Title: The history of evolution

Passage: More than 4 billion years ago, life began on Earth. According to evolutionary theory, living organisms have changed over time. Evolution occurs over millions and tens of millions of years. The process of evolution can be summed...

Topic: Medical Technology

GPT4

Title: The History of X-rays

Title: The Evolution of Medical Technology

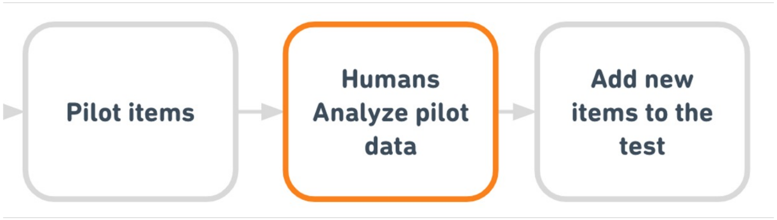
Title: The Use of Computers in Medical Research

Passage: In 1895, German physicist Wilhelm Röntgen produced the first photographs that showed the skeletal structure of living things. These pictures were made when x-rays (electromagnetic radiation) passed through the air...

Passage: The medical profession can be traced back about 5,000 years. Primitive medical care in the form of herbal remedies was common and early civilizations tended to look after the sick and injured. Although medical knowledge and technology have advanced over time...

Passage: Science and technology have improved health care greatly over the past few decades. Computer technology allows researchers to store, process, and analyze massive amounts of data much more efficiently. This has resulted in many life-saving discoveries, such as...

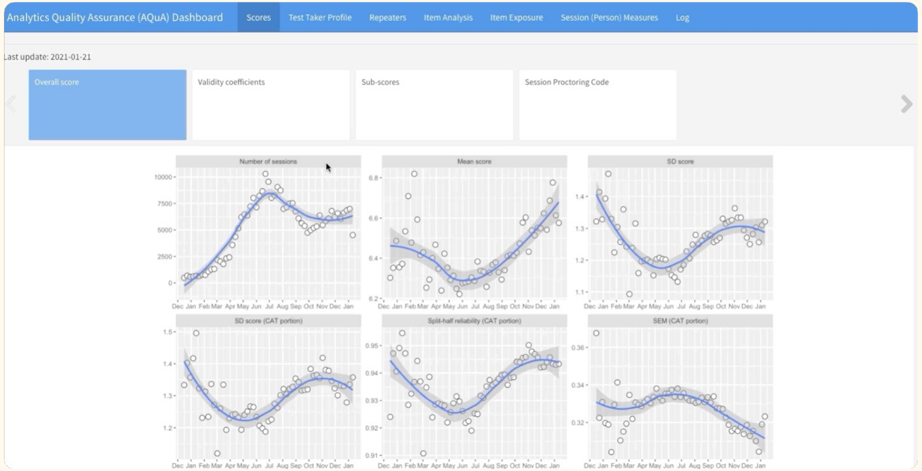
AI Test Development



4

5

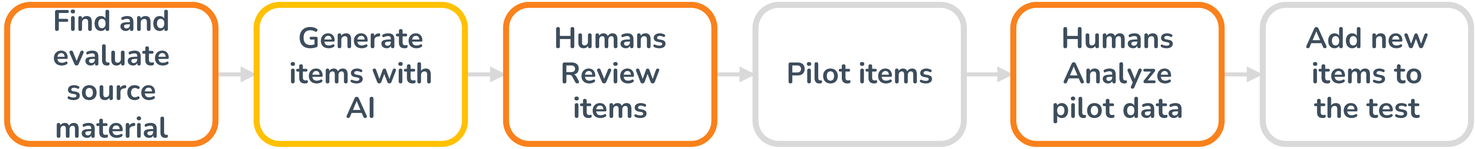
6



Last update: 2021-01-21

We utilize an internally developed dashboard called (AQuAA) that tracks and reports on all validity-related metrics of our test daily across the globe

Human-in-the-loop Item Generation



Find and evaluate source material

Generate items with AI

Humans Review items

Pilot items

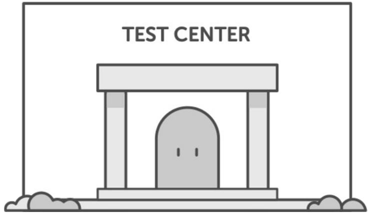
Humans Analyze pilot data

Add new items to the test

- **Result:**
    - Allows us to more efficiently scale our test creation processes
    - Keep development costs low for a cheaper test (**Duolingo English Test costs 80% less than other high stakes tests**)
    - Build a larger item bank to keep the test secure

3. Computer Adaptive Testing

Problem 3 - A Three Hour Test



+3 hours to take a test

Problems this raises:

- Testing concentration rather than language
- Frustrating student experience
- Digital accessibility paradox

Solution:

- Computer Adaptive Testing

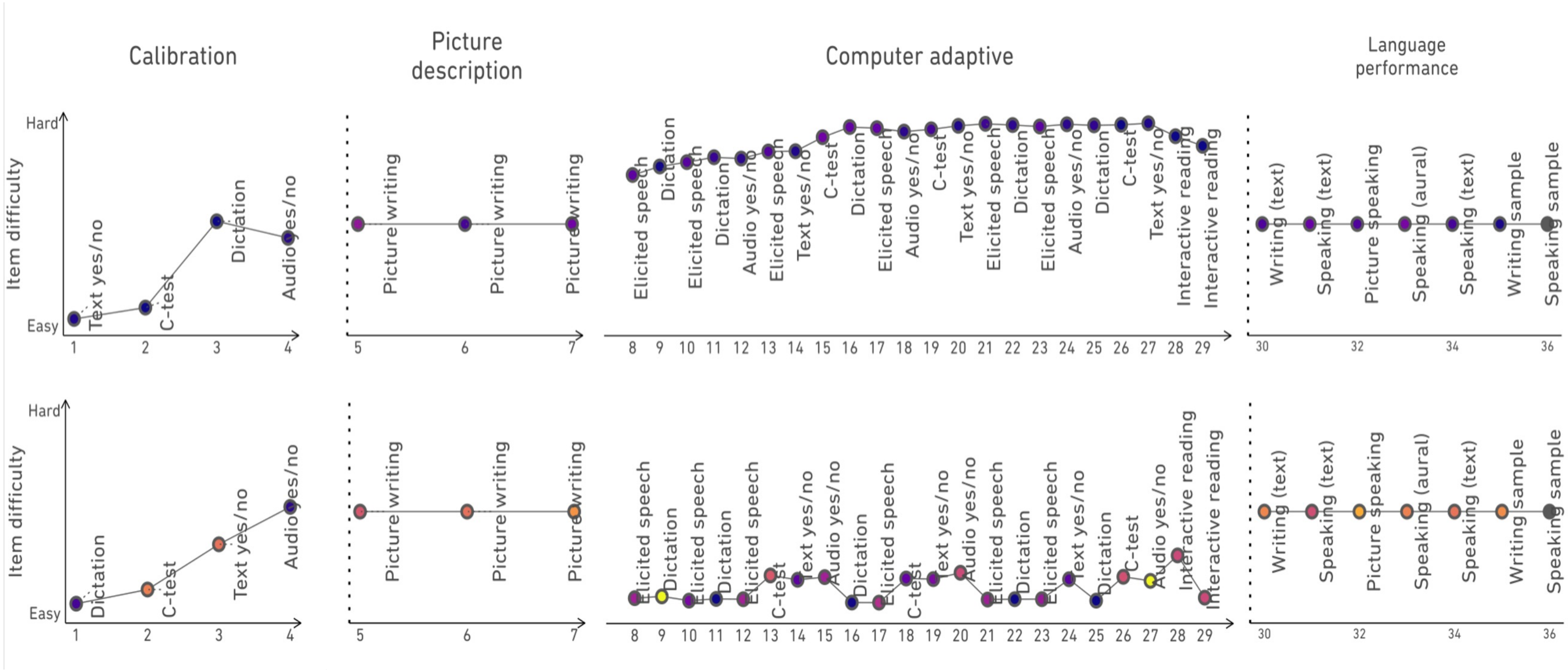
Computer adaptive test design

Adaptive engine draws items as test is being taken

15,000+ items

Number of unique test items
(No two tests are the same)

DET Administration: Four



160

80

Computer Adaptive Testing Benefits

The primary advantage of computer adaptive testing is that it can estimate test-taker ability more precisely with fewer test items and thus create opportunities for a test which is:

- More Efficient
- More Accurate
- More Accessible
- More Robust
- More Enjoyable

New Concordance, Items & Research

2022 update to test concordance

- In 2022, Duolingo updated its IELTS Academic & TOEFL concordance tables.
- Data set for concordance was extensive (More than 5000 IELTS scores)
- The correlation coefficients show strong, positive relationships of DET scores with TOEFL iBT scores and with IELTS scores.
- Concordance will be published in 2023 in the Journal: Language Testing

Updated concordance: common entry points

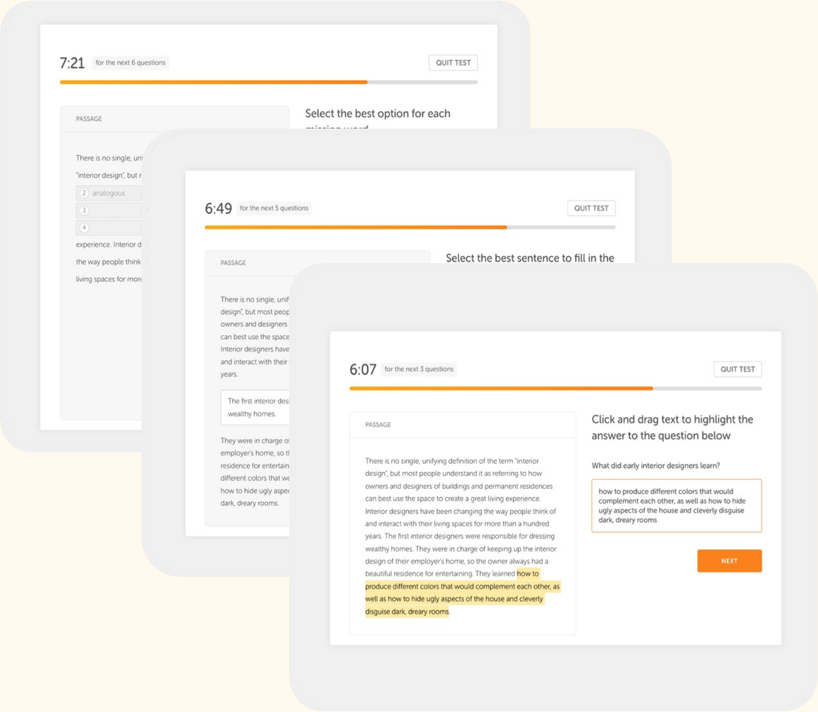
IELTS Academic	2019 DET Concordance	2022 DET Concordance
7	115–120	**130–135**
6.5	105–110	**120–125**
6	95-100	**105-115**
5.5	85-90	**95-100**
5	75-80	**80-90**

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

New items:

Interactive reading

- Increased the amount of scored reading, academic content, and score reliability
- Cutting-edge item generation techniques, uniquely linked tasks and passages, and innovative response formats
- Includes 5 different tasks all stemming from the same passage of text.



6:07 for the next 3 questions

QUIT TEST

PASSAGE

There is no single, unifying definition of the term "interior design", but most people understand it as referring to how owners and designers of buildings and permanent residences can best use the space to create a great living experience. Interior designers have been changing the way people think of and interact with their living spaces for more than a hundred years. The first interior designers were responsible for dressing wealthy homes. They were in charge of keeping up the interior design of their employer's home, so the owner always had a beautiful residence for entertaining. They learned how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms.

Click and drag text to highlight the answer to the question below

What did early interior designers learn?

how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms

NEXT

New items:

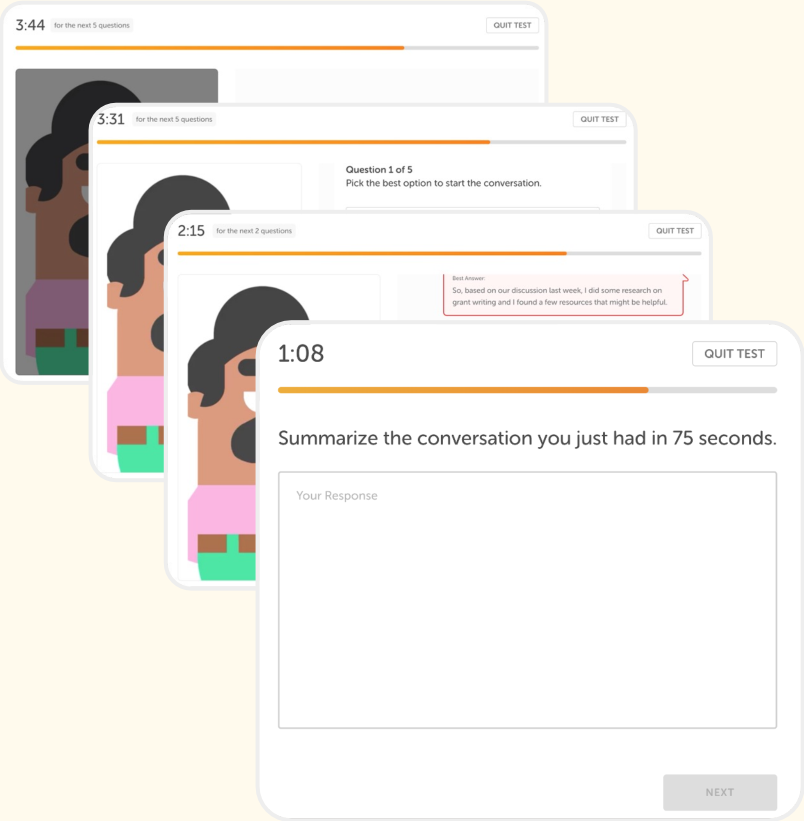
Interactive listening

Helps assess skills related to:

- Conversing in academic settings
- Starting and participating in a conversation
- Summarising the conversation based on their understanding of audio input

Test takers interact with the conversation in several ways:

- Listen to a conversation turn and select the best response from several options (immediate feedback is presented)
- Summarise the conversation in writing



Question 1 of 5
Pick the best option to start the conversation.

Best Answer:
So, based on our discussion last week, I did some research on grant writing and I found a few resources that might be helpful.

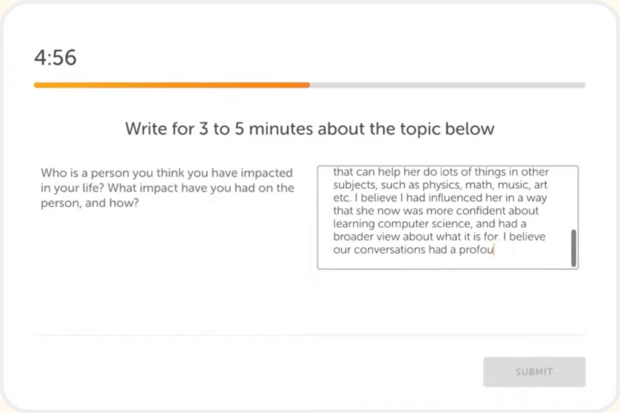
Summarize the conversation you just had in 75 seconds.

Your Response

Item Updates: Scored speaking and writing sample

- Increased scored speaking and writing content while maintaining test accessibility
- Expanded academic content on the test
- Increase test score reliability

Rubrics for scored speaking and writing:

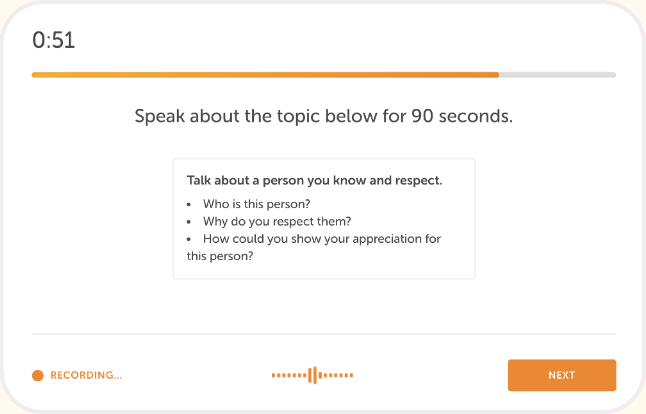


4:56

Write for 3 to 5 minutes about the topic below

Who is a person you think you have impacted in your life? What impact have you had on the person, and how?

that can help her do lots of things in other subjects, such as physics, math, music, art etc. I believe I had influenced her in a way that she now was more confident about learning computer science, and had a broader view about what it is for. I believe our conversations had a profou



0:51

Speak about the topic below for 90 seconds.

Talk about a person you know and respect.

- Who is this person?
- Why do you respect them?
- How could you show your appreciation for this person?

Latest research

Isbell et al. (2023):

Examines relationship between test-takers' performance on the DET and university stakeholders' evaluations of test-takers' speech (comprehensibility and academic acceptability).

Test-takers' comprehensibility and acceptability of speech showed a strong relationship with their official DET scores and subscores ($r \geqslant .74$–$.89$).

Research Page:

Check for updates

Article

LANGUAGE TESTING

Language Testing
1–30
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322231165984
journals.sagepub.com/home/ltj
SAGE

Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the Duolingo English test to the university

Daniel R. Isbell
University of Hawai'i at Mānoa, USA

Dustin Crowther
University of Hawai'i at Mānoa, USA

Hitoshi Nishizawa
University of Hawai'i at Mānoa, USA

Abstract

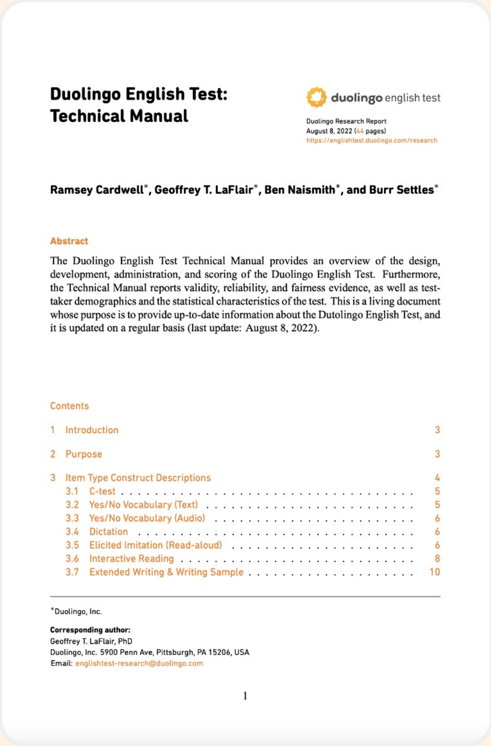
The extrapolation of test scores to a target domain—that is, association between test performances and relevant real-world outcomes—is critical to valid score interpretation and use. This study examined the relationship between Duolingo English Test (DET) speaking scores and university stakeholders' evaluation of DET speaking performances. A total of 190 university stakeholders (45 faculty members, 39 administrative staff, 53 graduate students, 53 undergraduate students) evaluated the comprehensibility (ease of understanding) and academic acceptability of 100 DET test-takers' speaking performances. Academic acceptability was judged based on speakers' suitability for communicative roles in the university context including undergraduate study, group work in courses, graduate study, and teaching. Analyses indicated a large correlation between aggregate measures of comprehensibility and acceptability ($r = .98$). Acceptability ratings varied according to role: acceptability for teaching was held to a notably higher standard than acceptability for undergraduate study. Stakeholder groups also differed in their ratings, with faculty tending to be more lenient in their ratings of comprehensibility and acceptability than undergraduate students and staff. Finally, both comprehensibility and acceptability measures correlated strongly with speakers' official DET scores and subscores ($r \geqslant .74$–$.89$), providing some support for the extrapolation of DET scores to academic contexts.

Corresponding author:
Daniel R. Isbell, Department of Second Language Studies, University of Hawai'i at Mānoa, Honolulu, HI 96822-2217, USA.
Email: disbell@hawaii.edu

englishtest.duolingo.com/research

Technical manual

- Introduction to the test
- Item type & construct coverage
- Test development and scoring
- Test administration & security
- Test taker demographics and performance statistics
- Accessibility, fairness & bias
- Quality assurance
- Concordance



Duolingo English Test: Technical Manual

duolingo english test

Duolingo Research Report
August 8, 2022 (44 pages)
https://englishtest.duolingo.com/research

Ramsey Cardwell*, Geoffrey T. LaFlair*, Ben Naismith*, 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, the Technical Manual reports validity, reliability, and fairness evidence, as well as test-taker demographics and the statistical characteristics of the test. This is a living document whose purpose is to provide up-to-date information about the Dutolingo English Test, and it is updated on a regular basis (last update: August 8, 2022).

Contents

*Duolingo, Inc.

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

go.duolingo.com/dettechnicalmanual

The Future - Duolingo

English Test

- Better mimic interactional competence in authentic academic settings
- New items based on stakeholder survey responses (E.g Can use references appropriately to support ideas, analyze, and refine arguments.)
- Test focus on specific industries, skillsets and language needs
- Champion responsible AI standards

Learn More

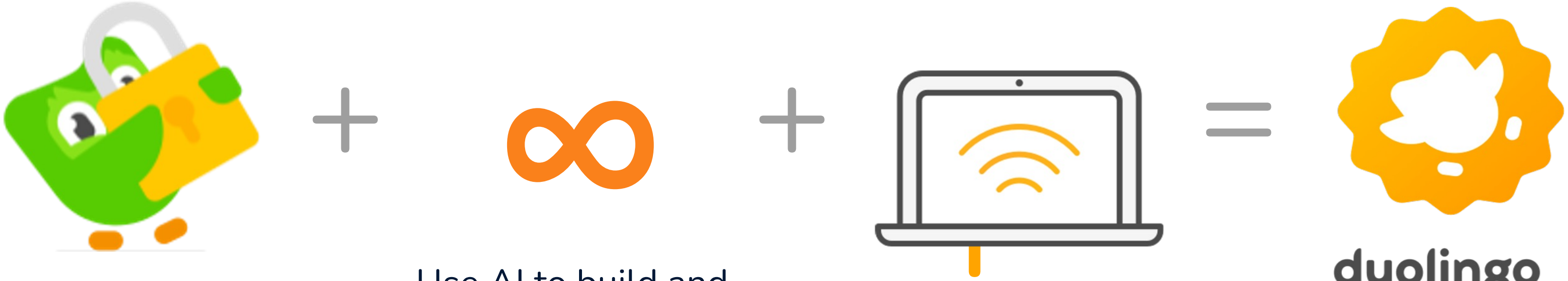
Australia Open House Events

Join Duolingo’s **Head of security and Assessment research leads** for a series of Afternoon PD events in:

1. Melbourne (October 02)
2. Sydney (October 04)
3. Brisbane October 06)

AI-POWERED DET

Digital-first innovation



Anonymised, randomised & AI-assisted proctoring

+

Use AI to build and sustain the largest item bank

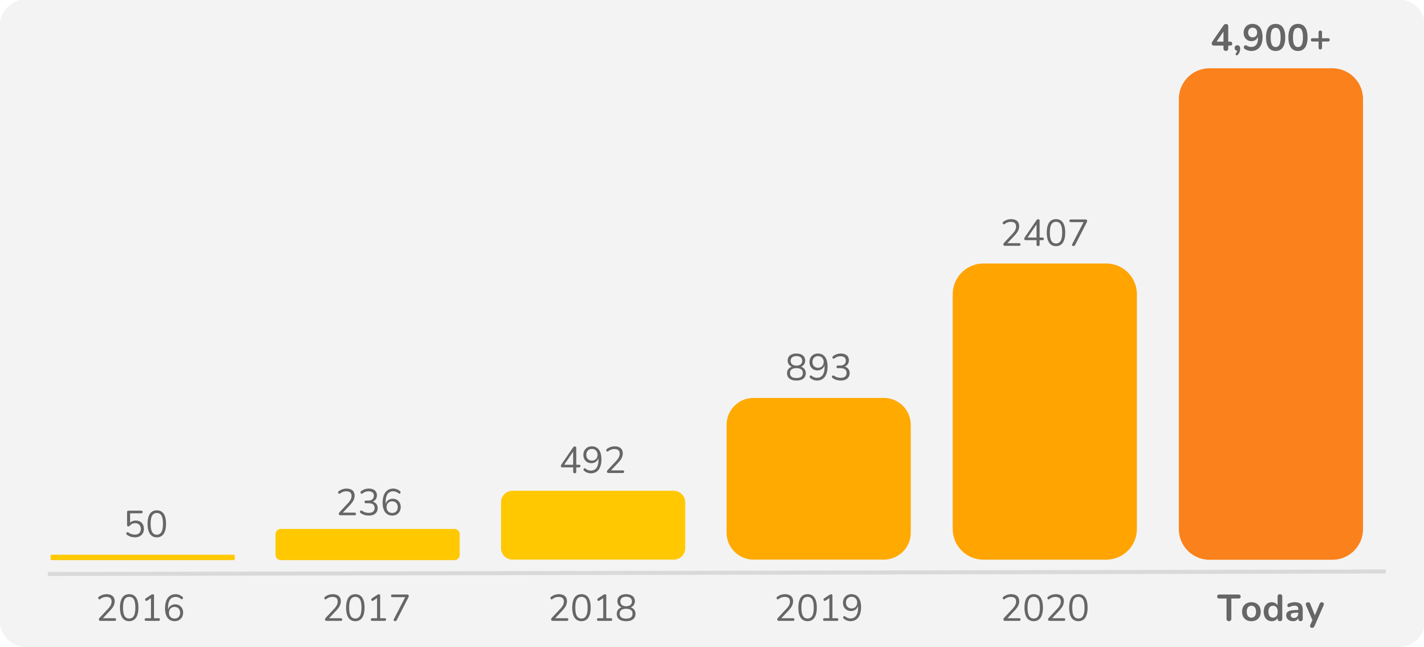
+

Computer Adaptive Test for accuracy and efficiency

=

duolingo english test

Thanks! Questions?



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

Email Enquiry

Browse Research