NEAS 2024
MANAGEMENT CONFERENCE

QUALITY REVOLUTION:

RETHINKING, RESHAPING, AND REDEFINING EXCELLENCE IN ELT

NEAS
QUALITY ASSURANCE IN EDUCATION AND TRAINING

Assessment in the Digital Age: Innovative Approaches to Testing and Evaluation

Duolingo
Brett Blacker & James Holden

Session Goals

Participants will understand:

- Different ways in humans are collaborating with AI to advance assessment, in particular related to building item banks and security
- How peers are approaching and collaborating with AI tools in the ELT context
- The latest research and development updates from Duolingo

Conference Themes

- *How can we redefine and measure excellence in ELT beyond traditional quality assurance methods, taking into account innovative approaches and revolutionary strategies?*
- *What transformative frameworks and best practices can be explored to reshape the future of ELT quality assurance, ensuring it remains relevant in an evolving educational landscape?*

Overview

1. Introduction

2. Collaborating with AI - Views on the present and future

3. Duolingo's approach to AI
    - Building Assessment using AI
    - Securing Assessment using AI
    - Latest from Duolingo Research / Development

duolingo english test

About Duolingo

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

World's Most Downloaded Education App

Languages
Literacy
Numeracy
Music

World's Leading High-Stakes Digital Language Test

duolingo english test

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

Collaborating with AI - Views on the present and future

Would you prefer to have a human doctor or an AI algorithm diagnose a medical condition?

Would you feel comfortable taking a plane flown by only AI?

Would you trust an AI tool to organise your finances?

If a human was also involved in confirming and validating the AI's decisions and actions, would you feel more comfortable?

Collaborating with AI is about leveraging the best of machines and the best of humans.

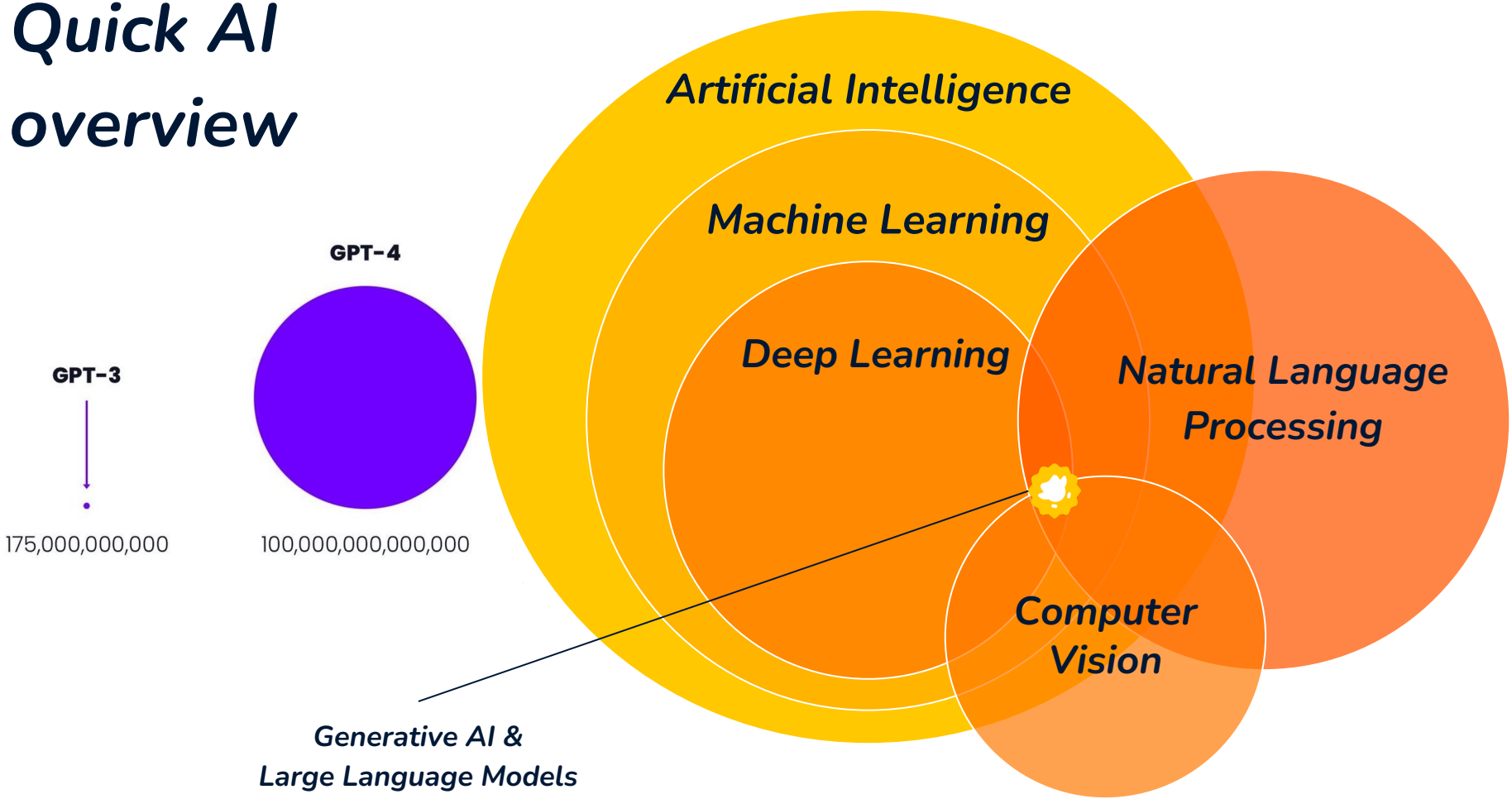
We always want a 'human-in-the-loop'

Best of Machines	Best of Humans
• Data Processing and Analysis	• Creativity and Innovation
• Consistency and Scalability	• Emotional Intelligence
• Predictive Capabilities	• Adaptability and Intuition

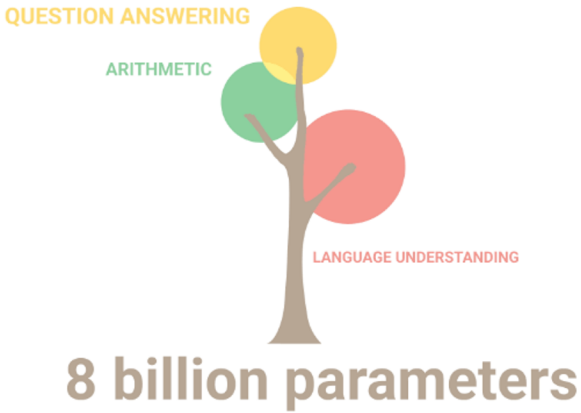
Humans program the AI - AI crunches the data - Humans validate and confirm

Human-in-the-loop AI
Duolingo's approach

Quick AI overview



More data = more capabilities



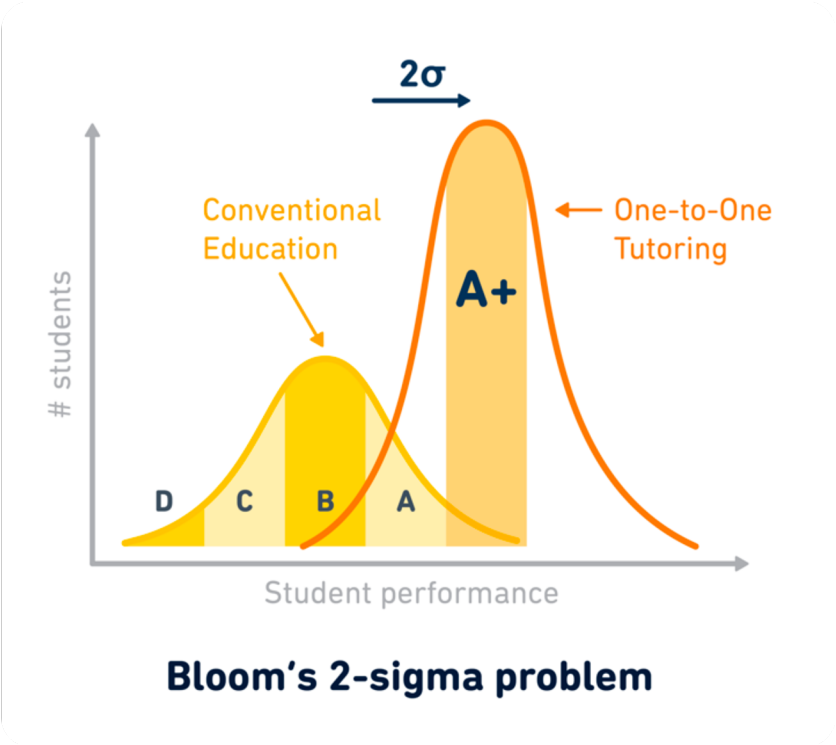
Med-PaLM (Google) was the first LLM to obtain a passing score on U.S. medical licensing questions, and in addition to answering both the multiple choice and open-ended questions accurately, it also asks questions, provides reasoning and is able to evaluate its own responses

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'
- Humans must be in the loop at all times to maximise the strengths of both humans and machines

Blooms 2 Sigma Problem (1984)

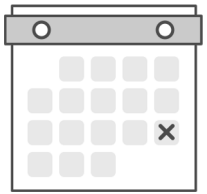


Bloom's 2-sigma problem

- 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?

AI = More inclusive, personalised learning which is no longer constrained by a learners resources

Overcoming traditional testing

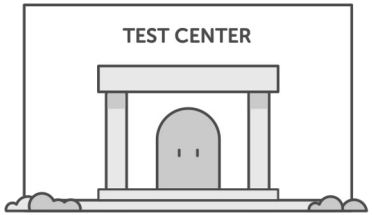


Book a date
Pay >$400

Generative AI item creation

Go to test-center or even travel to a different city

Online on-demand test
Asynchronous AI assisted security

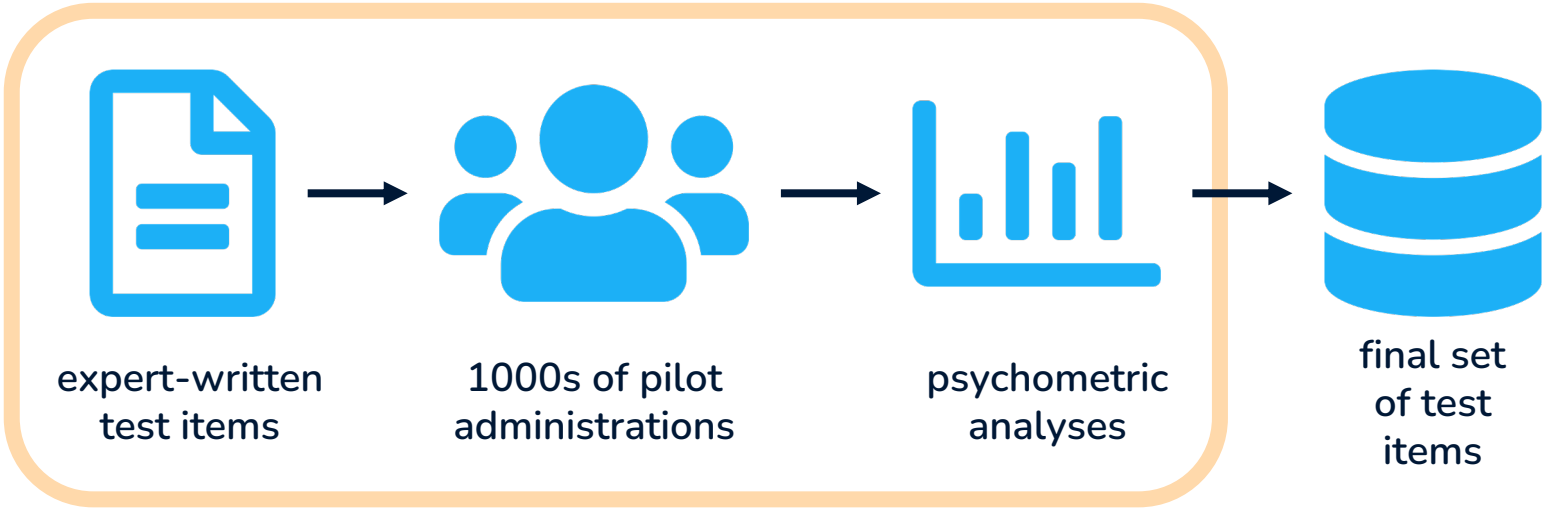


+3 hours to take a test

Computer Adaptive Testing

Building assessment with AI

Conventional test development



Time-consuming, vulnerable & expensive

The Promise of Large Language Models

- Large language models (LLMs) are trained to generate text from billions or trillions of words.
- LLM's excel at few-shot learning - ***can mirror style, format, and content given extremely few examples***
- Allows rapid prototyping of new content and item types

GPT-4 + DET

Item Generation

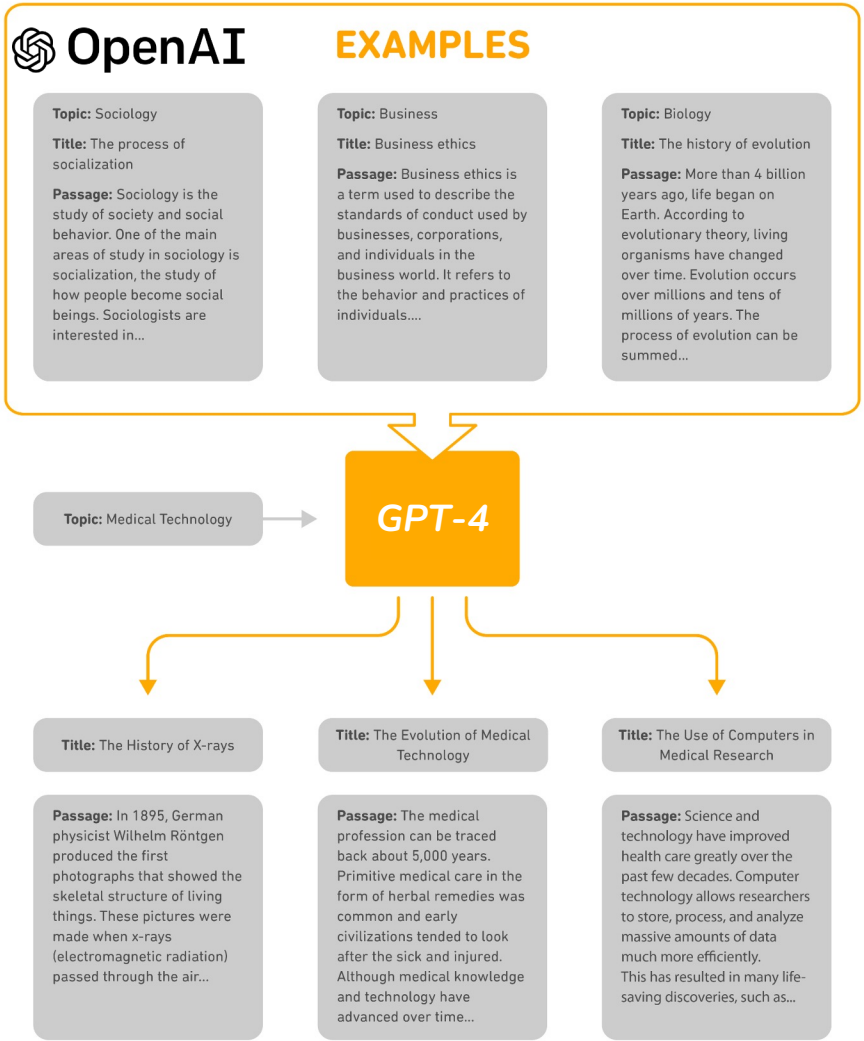
Goal

Create texts and associated materials, such as:

- Reading passages
- Conversations
- Comprehension tasks, correct answers and distractors
- Other information necessary for automated scoring

Method

Generate passages and titles
conditioned by other academic content
and a specified topic using prototypical examples



Example: Interactive Listening

Goal

Generate short conversations oriented around academic scenarios.

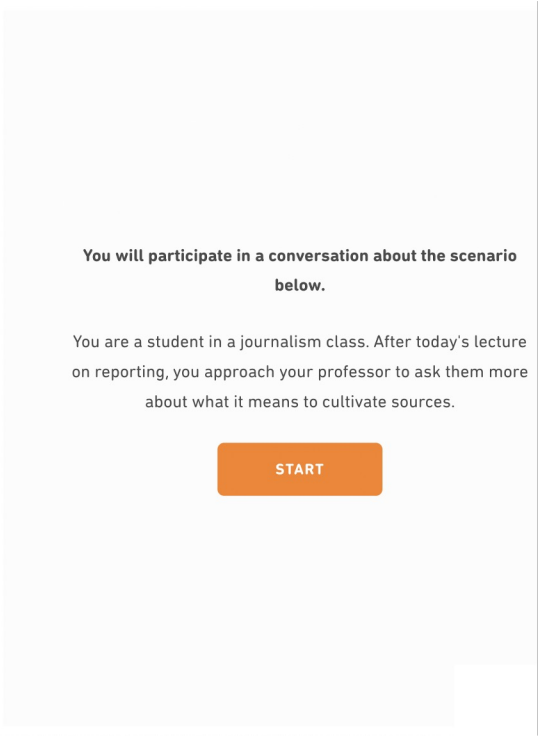
Challenges

1. Scenario selection - what types of situations to cover?

   → Human expert

1. Scenario expansion - adding details to create interesting & diverse conversations

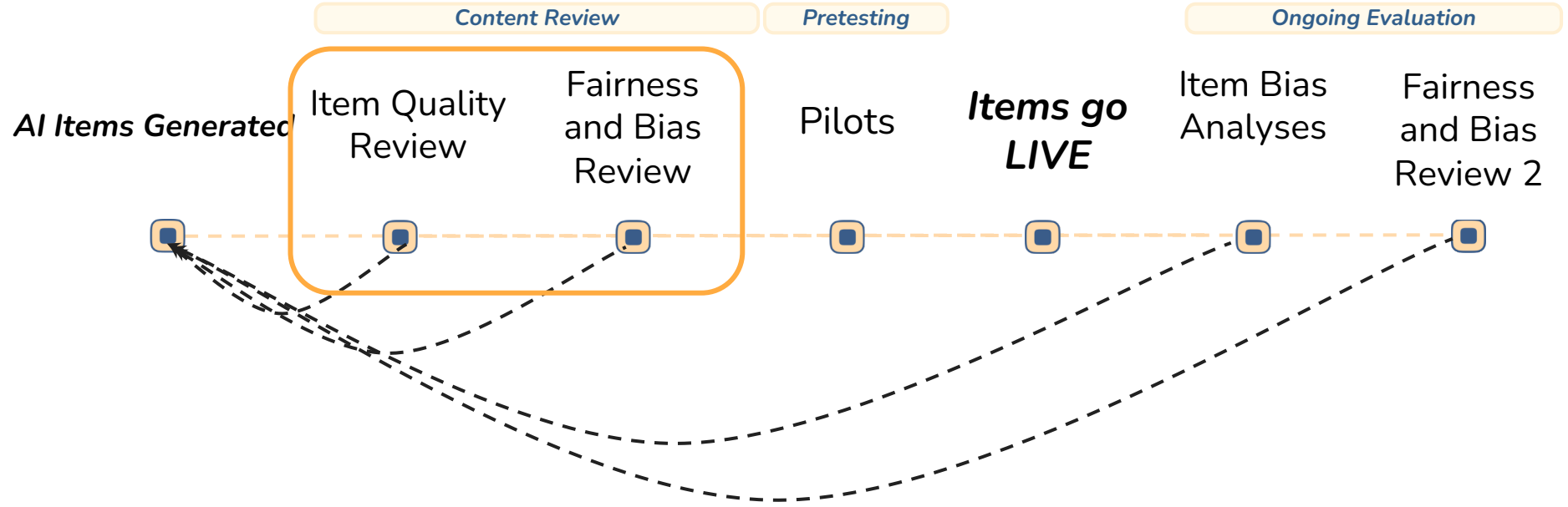
   → LLMs



AI item banks require extensive human review

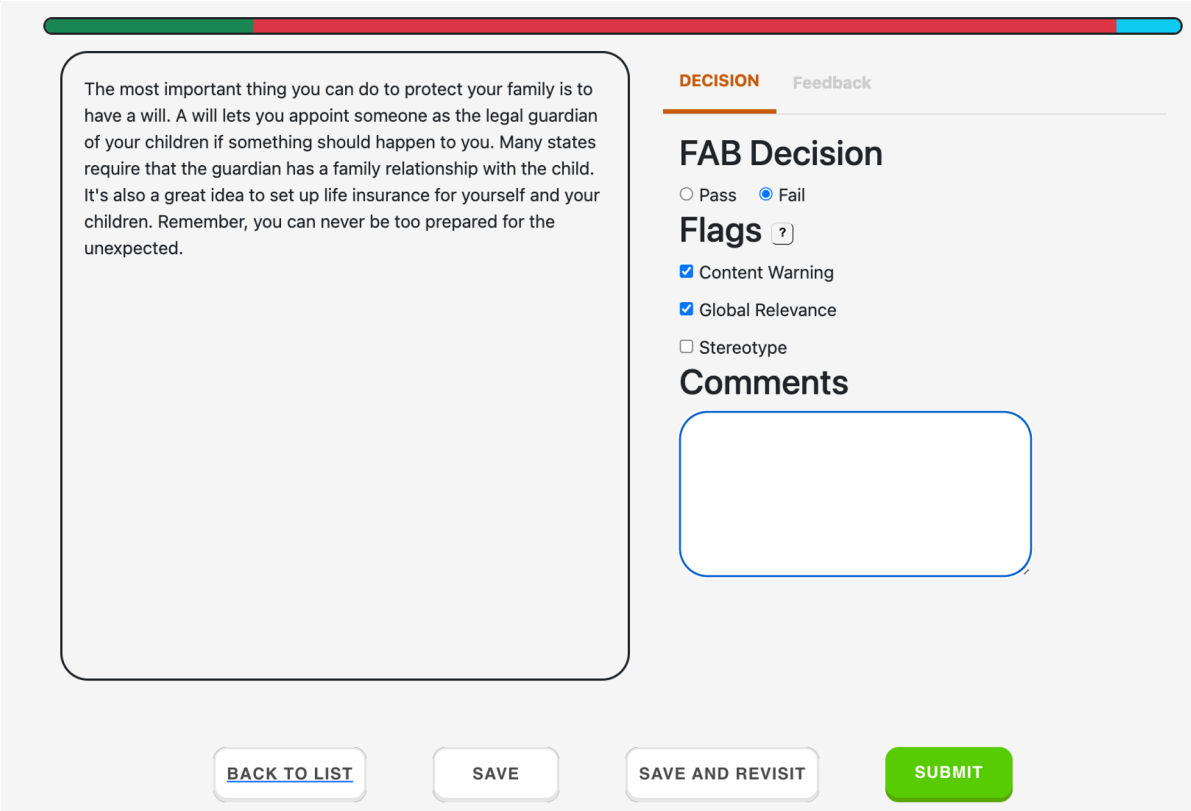
Human-In-The-Loop AI

Goal: Use collaboration of humans and AI to perform a task that is difficult for both



Content Review

FAB review process



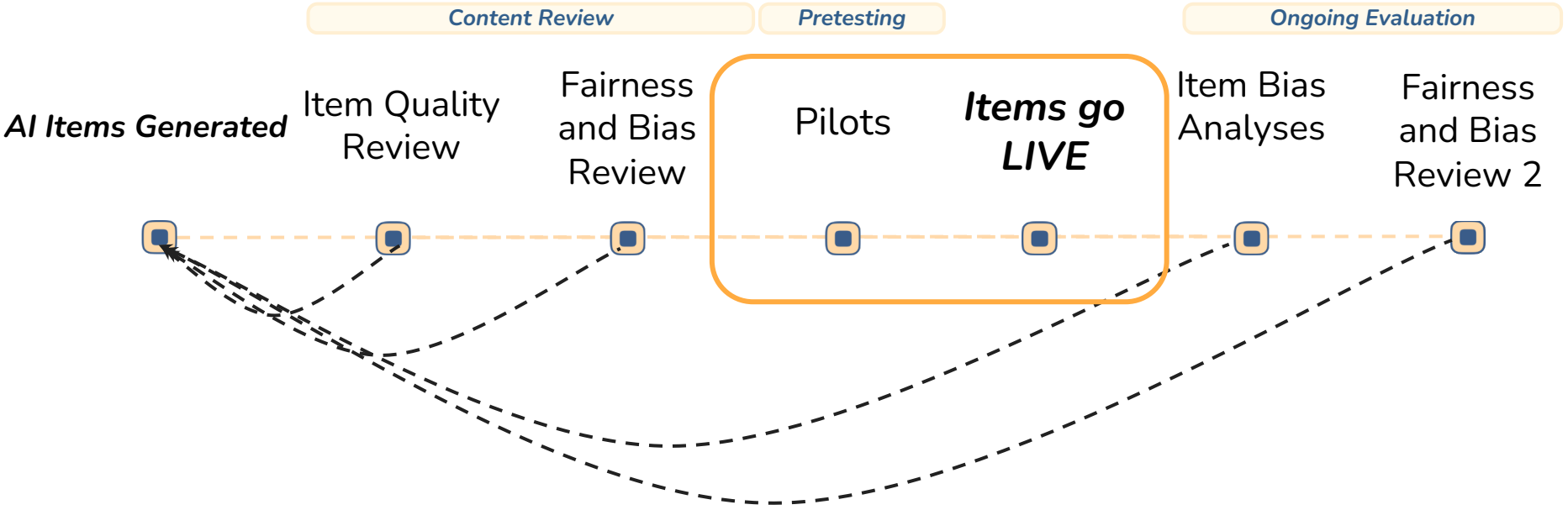
Content Review

Calibration items

Text	Your Label	Our Label	Your Codes	Our Codes	Your Comment	Explanation
As the general American public becomes informed about the cost of a university education, they are becoming more reluctant to pay for it. The rising costs of health care, rent, food, and commodities overall have made people take a closer look at their finances. Some people can't afford to pay to go to college. The fear of student loan debt has driven young people away from a college education.	Fail	Fail	Content Warning	Content Warning	this topic might be sensitive to some, particularly those struggling with financial challenges related to education, although it is described in a general way without targeting or blaming specific groups. Since many of the TT demographic are planning to go to college it's best to fail this item	The passage focuses on income disparity and poverty.

Human-In-The-Loop AI

Goal: Use collaboration of humans and AI to perform a task that is difficult for both

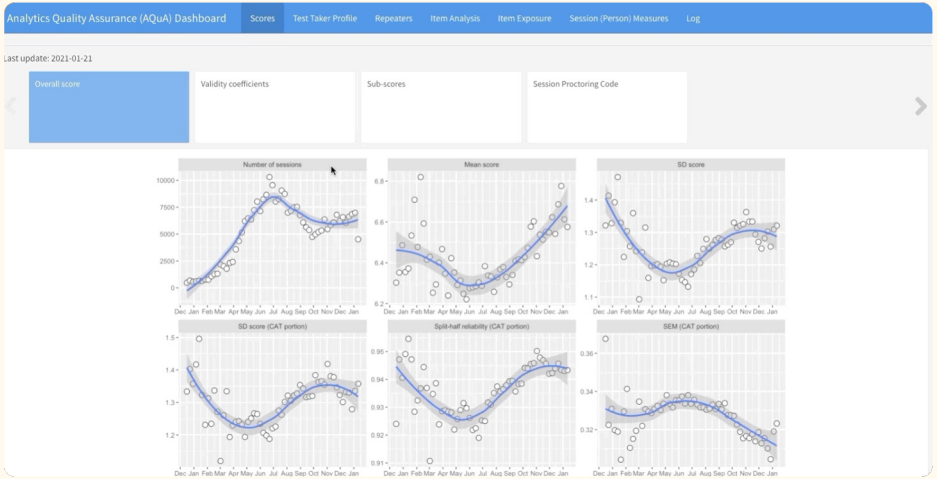


AQuAA

Analytics for Quality Assurance in Assessment

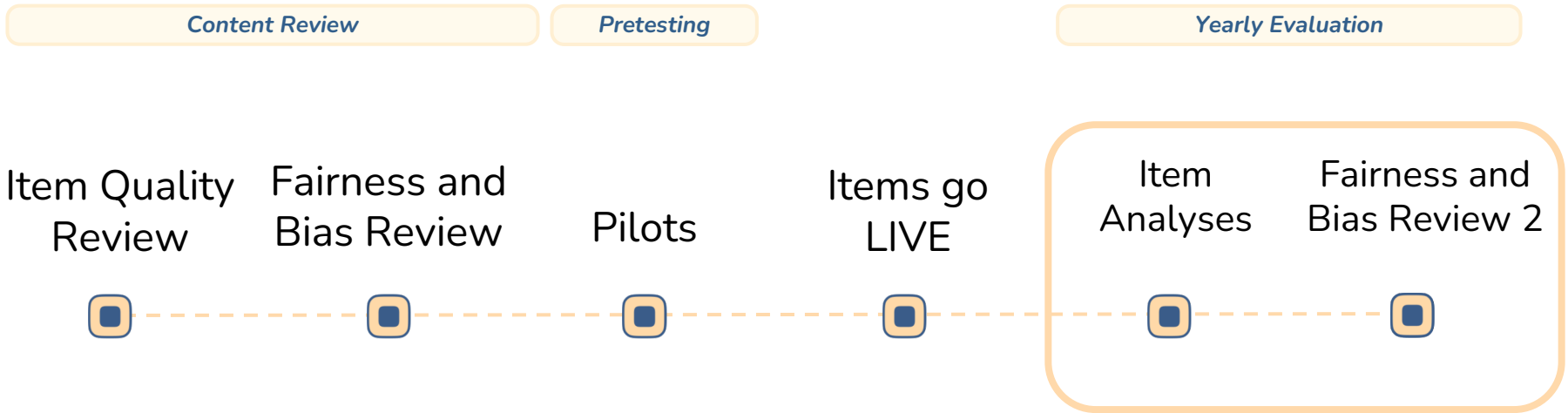
We utilise an internally developed dashboard that tracks and reports on all validity-related metrics of our test daily across the globe

englishtest.duolingo.com/research



Content Review

Item Development Timeline



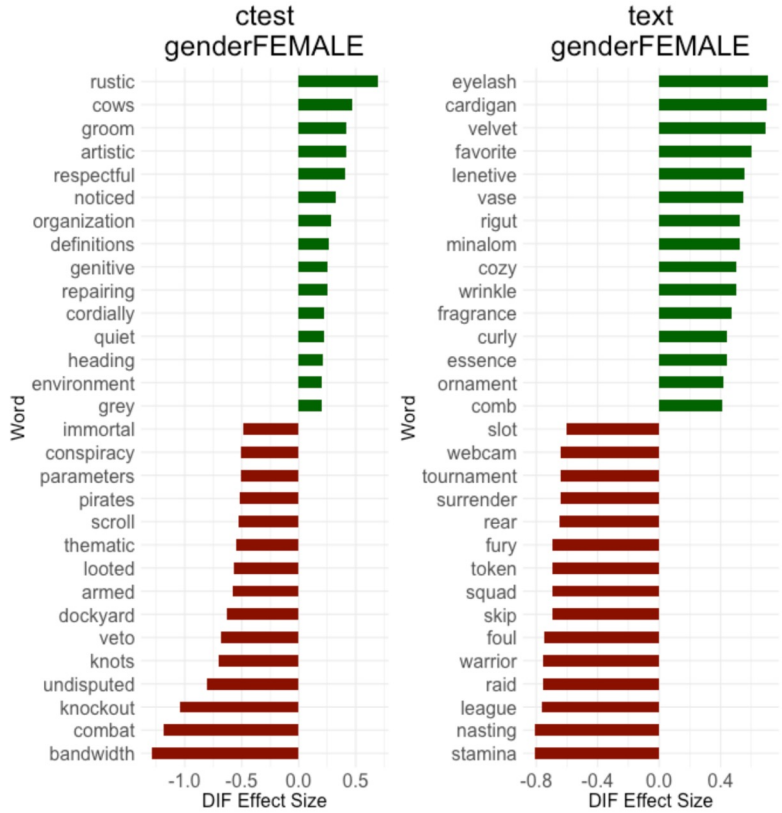
Item Bias

Differential Item Functioning (DIF)

A statistical method for testing whether items are “biased”

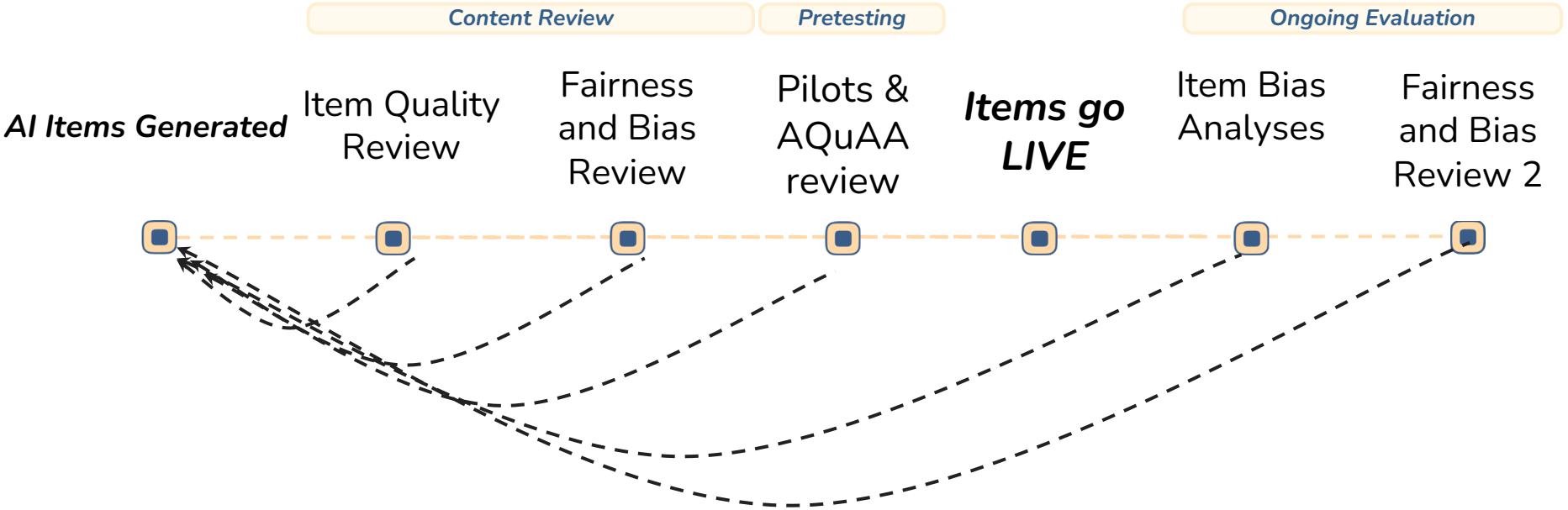
Item “bias”:

Individuals with the ***same level of English proficiency*** but ***different backgrounds*** (e.g., male and female) have ***different probabilities*** of answering an item correctly.



Human-In-The-Loop AI

Goal: Use collaboration of humans and AI to perform a task that is difficult for both



Benefits of AI item creation

Item banks which are more:

- Secure
- Aware and adaptable to test-taker population trends
- Efficient and affordable to create
- Scalable
- Customised

Our Own Contexts - Assessment Creation

Have you attempted to create your own assessments / item banks using LLMs?

What review process or human involvement do you use in your own assessment development?

Securing assessment with AI

Our Own Contexts - Security

How are digital innovations being used in security (personal, home, finance)?

Is digital security more or less trusted?

The live proctoring disadvantage

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

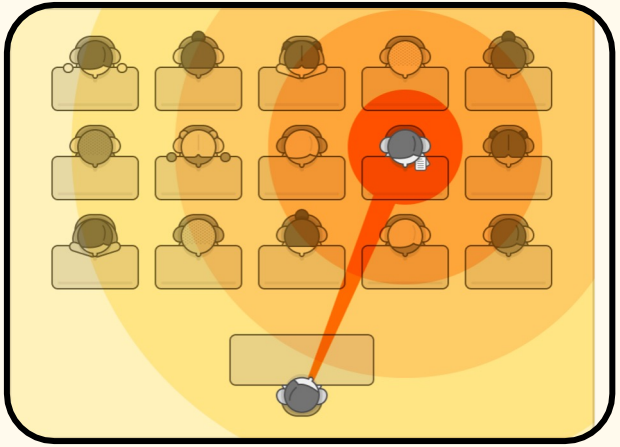
- A single point of failure can disrupt the entire chain.

Live 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 compromises the entire testing cohort.



Los Angeles MAGAZINE

CRIME EDUCATION

A Chinese Cheating Ring at UCLA Reveals an Industry Devoted to Helping International Students Scam Grades

Operation [redacted] was overshadowed by Operation Varsity Blues, but it's just as scandalous

By Christopher Beam - April 26, 2019

Share

On paper, Liu Cai was a model student. After moving to the United States from Beijing, he majored in biology at UCLA and volunteered at the Boys &

Defining a digital threat model

Digital Threats	***Attacker***	***Mitigations***
Test theft	Scrapers	Large item pool, adaptive engine, Screenshot prevention etc.
Getting third party aid during test	Cheaters, Cheating rings	Monitoring system softwares, preventing window switching, multi-layered evidence collection, no human<>test taker contact
Identity and account abuse	Imposters	ID verification, previous session and account matching, ID history, shared video response etc.

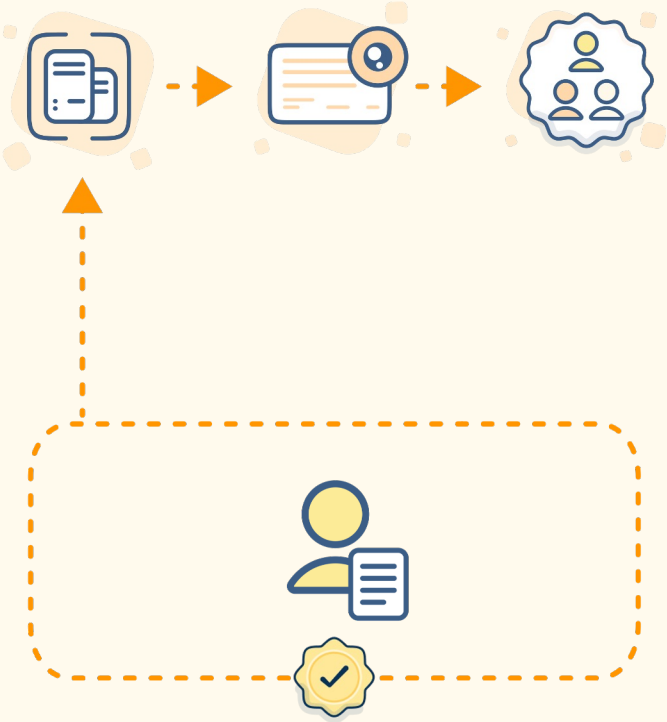
“Human-in-the-loop” AI

Remote

- Testers & proctors located all over the world who have no interaction with each other

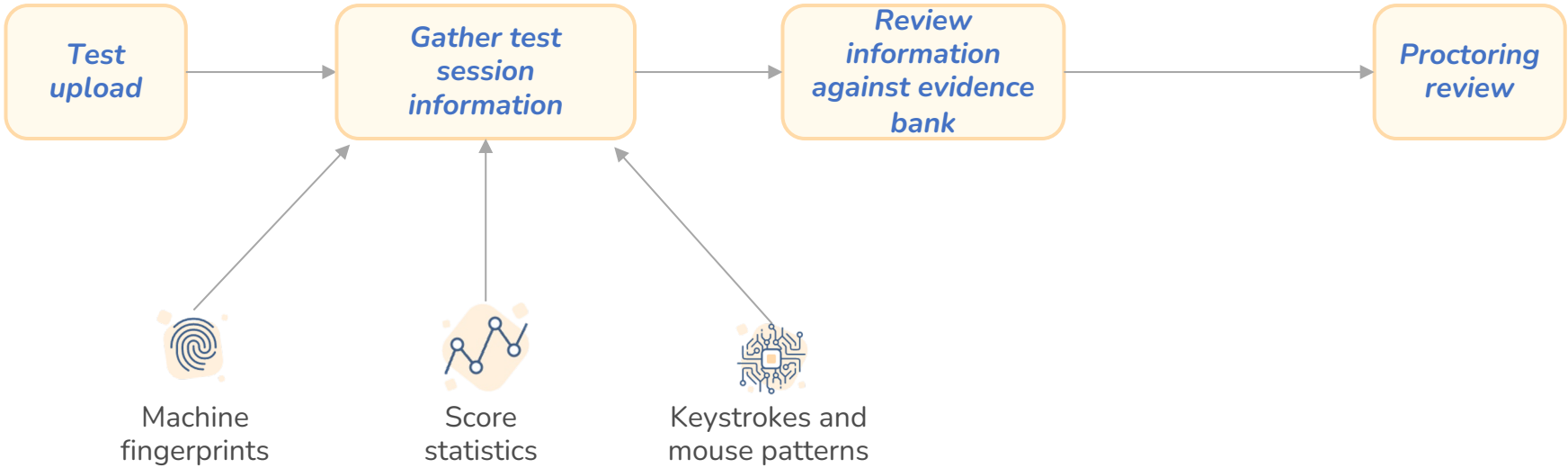
Record-and-review

- Test video is recorded and uploaded
- AI scans 150+ evidence categories
- AI flags and test video is reviewed by human proctors

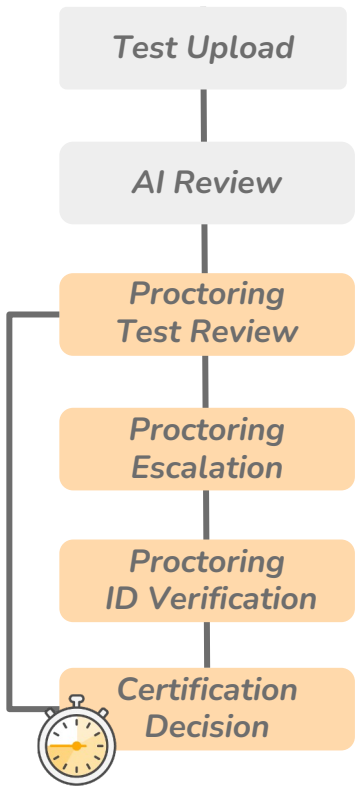


Asynchronous proctoring approach

How Do We Detect Cheating?



The proctoring process



Offboarding begins as the test taker waits for the video to upload.

≤1 hour Analysis of biometrics, hardware, software, and PII data takes place.

≤20 hours A proctor reviews the entire test, looking for anything that would prevent a fair certification.

≤45 hours Some tests are escalated to Senior Proctors for another review, while some tests receive a certification decision after 1 review.

≤48 hours ID and face images are reviewed by our exclusive ID Verification Team.

≤48 hours All test takers receive a certification decision via email and their DET homepage within 48 hours of upload.

Example - Plagiarism Detection

Why is plagiarism detection necessary?

- Some test takers could cheat by searching the Internet for answers or use external resources

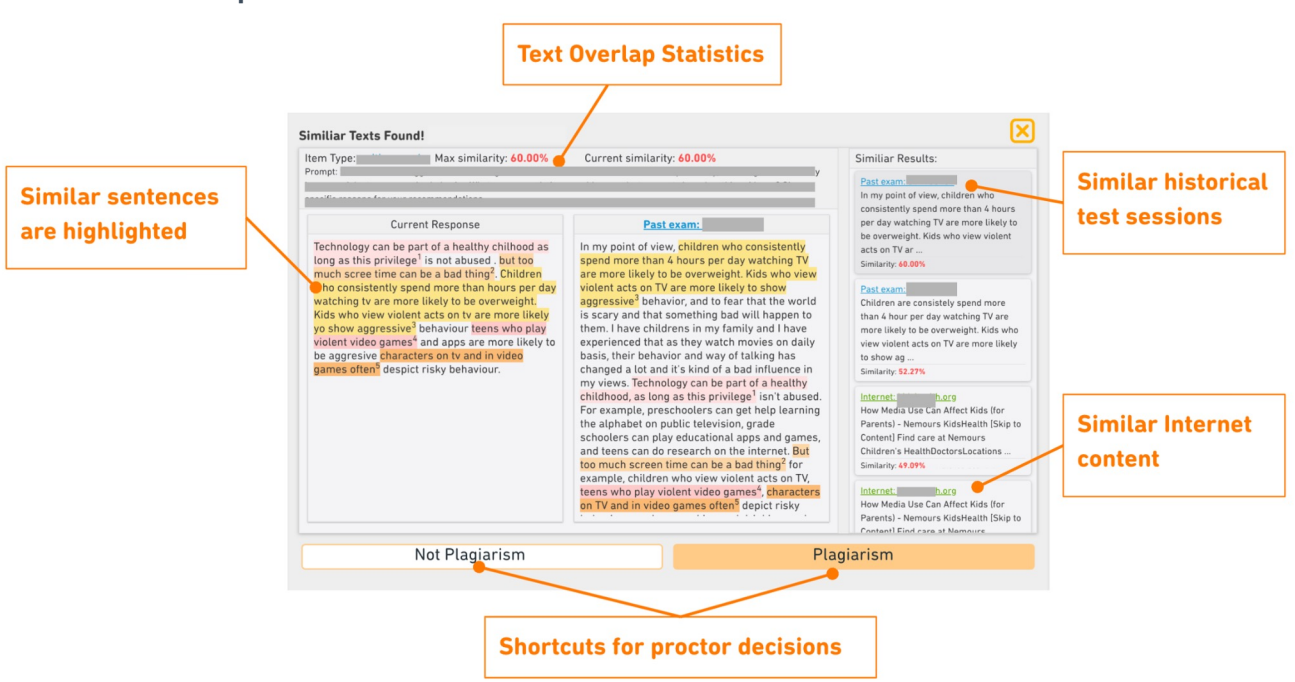
- Identifying and preventing such malicious behaviors is necessary to maintain the integrity of the test

- Plagiarism detection can:
  - Catch these malicious behaviors
  - Have a deterrent effect on these malicious behaviors

SECURITY AND PROCTORING

Plagiarism detection

The similarity of text responses are evaluated against external resources AND historical DET responses.



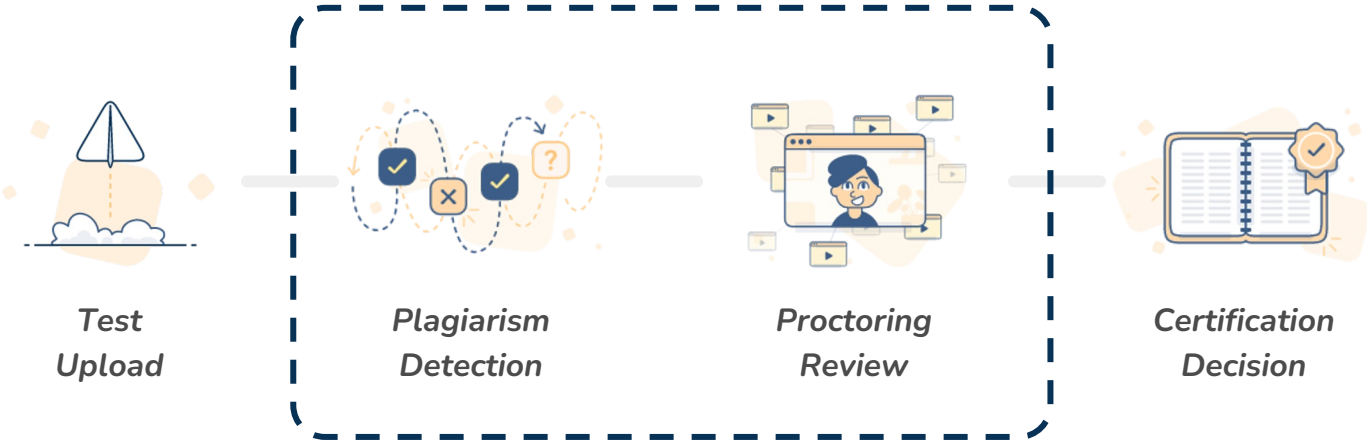
SECURITY AND PROCTORING

Plagiarism detection

- Expanded to spoken responses in 2023
- Allows our human proctors to focus on observing behavior

SECURITY AND PROCTORING

Plagiarism detection



Among all the sessions that are flagged by the plagiarism detection algorithm, 93.6% are determined as having malicious behaviors by proctors.

Humans always make the final certification decision.

Digital security options for the ELT sector

Safe Exam Browser

- Accessible (free and open source - Works on computers & tablets)
- Security Strengths (Website blocking, Virtual Machine Detection, Windows Switching prevention)
- No registration system, ID monitoring, nor comprehensive cheating detection

Safe Exam Browser
SEB

Exam.net

- Secure browsers and screen lockdown, as well as ‘Background cheat detection’
- 3 security modes, tailored to the testing environment
- Comprehensive ability to create and mark assessments, however $$$

exam.net

Our Own Contexts - Security in ELT

How are you currently managing security and proctoring in your own assessments?

Have you used any digital security tools or software?

Do you plan to / are under pressure to do this?

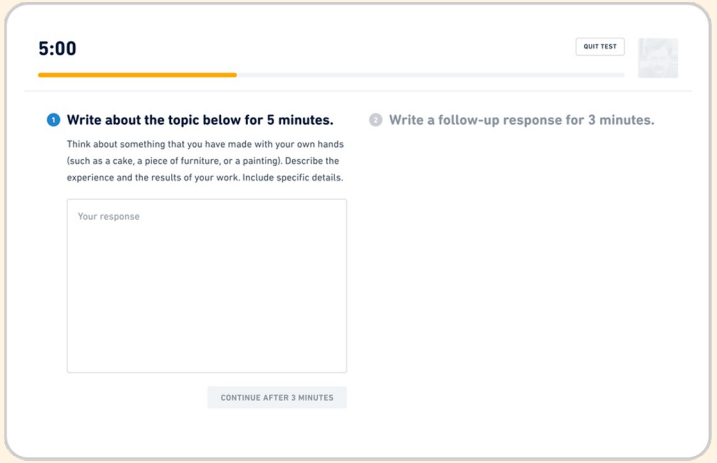
What are the potential risks?

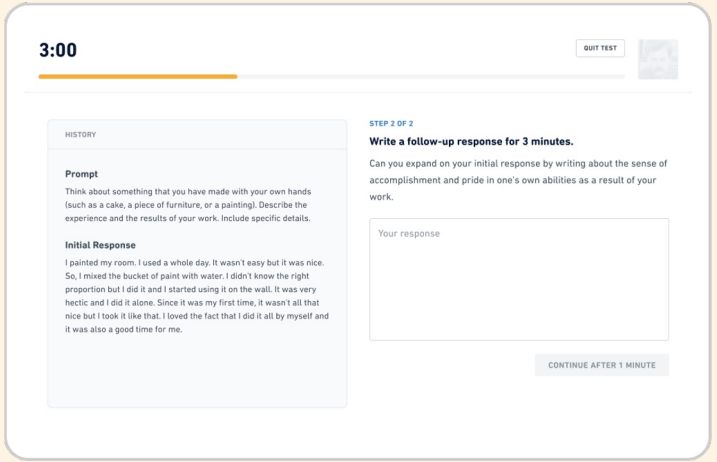
Research & Development Update

TEST UPDATES

24’V8 new items:
Interactive Writing

- New task taps into ***higher level writing skills (e.g., elaboration, revision)***, allowing test takers to demonstrate their academic writing proficiency better.





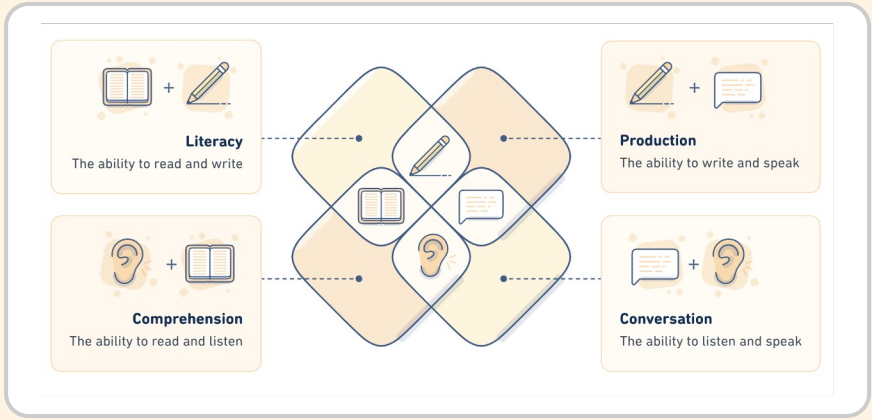
TEST UPDATES

DET 2024

New Scoring Coming in July

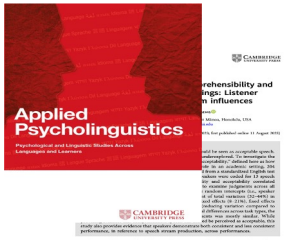
- All tests will now include new ***Speaking, Writing, Reading, and Listening subscores*** alongside our existing integrated modality subscores.

“DET requires a score of 115, with no subscore below 105” etc.



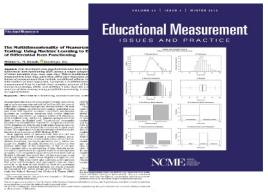
DET validity & reliability research

2023 - Predictive and Academic Validity: DET scores display a strong correlation with academic evaluations of students' English proficiency and subsequent academic performance (Isbell et al. (2023)



2023 - Concurrent Validity: In a 2023 study of over 5000 students, DET scores display significant, strong correlations with scores from TOEFL and IELTS English language tests taken by the same individuals. Cardwell et al. (2023)

2023 - Test-Retest Reliability: Overall test-retest reliability of **0.93**, along with the reliability of subscores ranging from 0.90 to 0.92, underscores the DET's scoring consistency. In March 24 test-retest reliability has reached **0.95** (Journal of *applied psychological measurement - Will be published in 2024*)



2024 - Task Duration: The Impact of Task Duration on the Scoring of Independent Writing Responses. Equally high test reliability and criterion validity (writing) was demonstrated by a 5 minute task, compared to a 20 minute task. (Naismith et al. 2023)

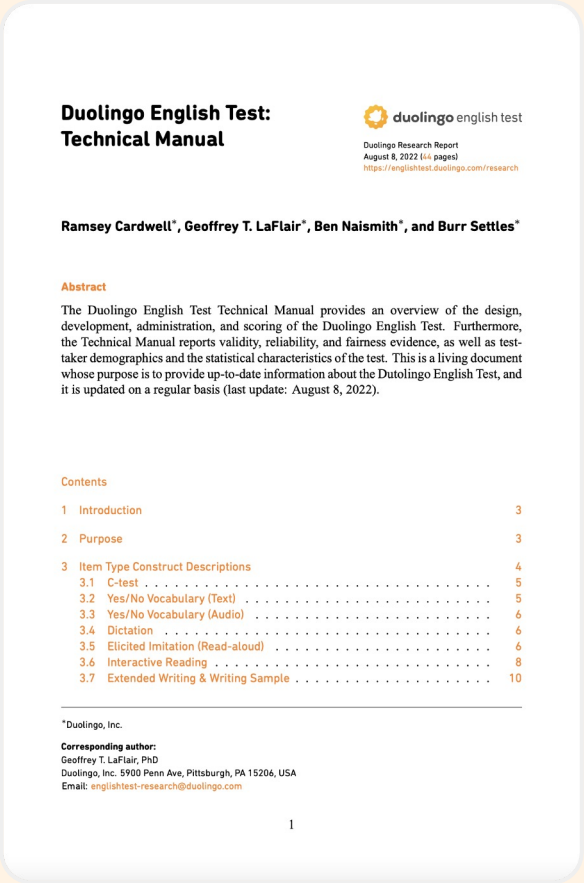
Ongoing - External Validity:
Endorsed a high-stakes English test by Australian English Quality Assurance organisation NEAS.
Endorsed with full CEFR validity alignment by the UK Chartered Institute of Linguists
Successful tracer studies from the world's most elite universities

- englishtest.duolingo.com/scores
englishtest.duolingo.com/research

TEST DESIGN

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



Duolingo English Test: Technical Manual

duolingo english test

Duolingo Research Report
August 8, 2022 (## 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

Collaborating with AI is about leveraging the best of machines and the best of humans.

We always want a 'human-in-the-loop'

Best of Machines	Best of Humans
• Data Processing and Analysis	• Creativity and Innovation
• Consistency and Scalability	• Emotional Intelligence
• Predictive Capabilities	• Adaptability and Intuition

Humans program the AI - AI crunches the data - Humans validate and confirm