duolingo english test

Webinar Roundup:

Research & Security: Demystifying Online Testing

'Mythbusting' misconceptions and showcasing innovation

Click here to watch the June 17, 2021 recording.

Table of contents

Introduction, mission, and test updates (0:00 - 7:03)

Jeremy Matula, Lead Strategic Engagement Manager
Kevin Hostetler, Senior Strategic Engagement Manager

INTRODUCTION

In today's webinar, we'll be 'mythbusting' some misconceptions regarding English proficiency testing. **We'll also showcase how Duolingo's innovative approach continues to set new industry standards and best practices.**

As the world shifts toward more flexible online testing options for students, there's been a lot more conversation about best practices for test design, administration, and security. In the Q&A, we'll be hearing from a select group of experts from the Duolingo English Test on some of the most common and important questions we've received across topics like test construct, validity, item development and quality assurance, scoring, and security and proctoring.

See full biographies in the appendix.

MISSION

English certification should be a bridge, not a barrier. We are here to build a better bridge. We do it by making English proficiency testing accessible and affordable for applicants—and faster and more reliable for institutions. The result: more opportunities for test takers and more diverse talent for institutions.

TEST UPDATES

In 2020, **hundreds of thousands of test takers from 207 countries and related territories and speaking 144 different first languages** took the Duolingo English Test.
Duolingo **offered over 10,000 fee waivers** to support the access initiatives of over **150 counselors and organizations** last cycle, benefiting students in **over 70 countries**. Furthermore, Duolingo is proud to have partnered with the White House and **Vice President Kamala Harris's Call to Action** for the countries of Guatemala, El Salvador, and Honduras.

The Duolingo English Test is turning 5! As we celebrate this anniversary, **we will continue our mission to bring English proficiency testing to your students in a more accessible, cost effective, and secure way.**

Defining and measuring English proficiency (07:35 - 16:49)

Dr. Antony John Kunnan, Principal Assessment Scientist
Dr. Geoff LaFlair Senior, Assessment Scientist

QUESTION

Can you define 'academic' and 'general' English for us? How are they different and how do they overlap? (07:44-11:20)

ANSWER

Dr. Antony Kunnan, Principal Assessment Scientist

When we start learning a language as children we learn a casual or informal language which we refer to as 'general English'. As students start going to school, they are exposed to a more formal variety of English. They have to read informational articles, textbook materials or stories, they have to write and speak in a particular way. **As students progress from primary, through to secondary education, they will be exposed to more and more 'academic English' - English that is used in a certain way.** If you look at the linguistic content you will see the differences - academic English can have a different vocabulary for the same things we say in general English; different sentence patterns, different discourse patterns.

It's best to see the difference as a continuum from general English, into academic English, and then into more specialised forms of English. As we move into even more specialised fields of academic English,for example if you're a scientist, or a legal professional, a religious leader or a business person, then they have their own types of English. This is all a continuum.

This graphic (see slide Registers of Language Use) shows some of the different things that students have to do with language, such as classroom activities, and as they progress, argumentative writing, lab reports etc. We call these different 'registers' - law, business, science, engineering. As we break this down, we see that there are different words, syntax or discourse patterns in our texts or listening materials, in our speech.

Registers of Language Use

- English (language) varies across different registers (or target domains)
- Differences in mode, relationship between interlocutors, purpose of communication, and the topic.
- Creates differences and overlap

QUESTION
How does the Duolingo English Test assess a range of English language proficiency - including 'academic' English? (11:33 - 14:05)

ANSWER
Dr. Geoff LaFlair, Senior Assessment Scientist
I'm going to speak to the variation in linguistic patterns across different registers, and whether the linguistic features are more informational (academic) or colloquial (general); this is something that we can count and quantify. Based on these linguistic features, the texts within a corpus have different dimension scores, 'D1 scores' that represent the aggregation of the linguistic features. The corpus is made up of text passages from fiction, news, and textbook sources as well as our c-test passages. **Items with a positive D1 score are typified by more colloquial features, items with a negative score are more informational, like in a textbook.** In our example (see slide Academic English in C-test Passages), we have a text from an environmental science textbook with a D1 score of -11.80 due to its informational features. This passage contains a lot of domain specific language, e.g. scientific language. We see words like 'molecules', 'organisms', 'environment', and scientific adjectives such as 'chemical' and modifying words like 'bonds'.

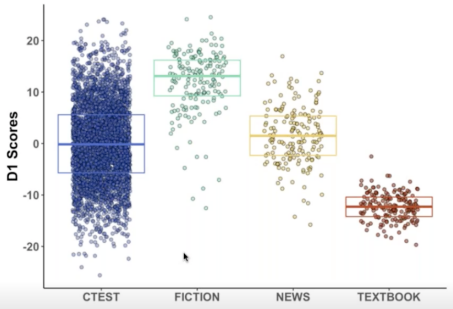
If you see an example from our C-test corpus, you can see a lot of similarities (see slide Academic English in C-test Passages).

When we juxtapose this with the item type difficulty, what we see is a strong correlation between item difficulty and D1 score. As the test gets harder, the passages have an increasingly negative D1 score, the items have more informational and academic-like linguistic features. The easier the item types, the more positive the D1 score, and the passages have more colloquial or narrative linguistic features. **What this means for test takers, is that as they prove their English proficiency is at a higher level, they will receive items with passages that contain more informational and academic language.**

Academic English in C-test Passages

Environmental Science (-11.80)

Atoms and molecules cycle endlessly through organisms and their environment, but energy flows in a one-way path. A constant supply of energy—nearly all of it from the sun—is needed to keep biological processes running. Energy can be used repeatedly as it flows through the system, and it can be stored temporarily in the chemical bonds of organic molecules, but eventually it is released
and dissipated.



Academic English in C-test Passages

Environmental Science (-11.80)

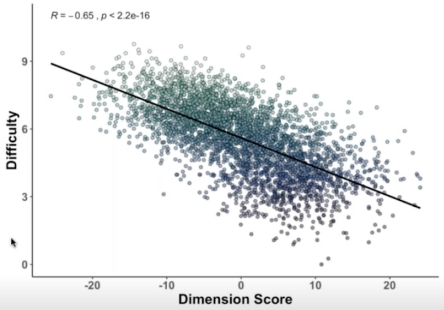
Atoms and molecules cycle endlessly through organisms and their environment, but energy flows in a one-way path. A constant supply of energy—nearly all of it from the sun—is needed to keep biological processes running. Energy can be used repeatedly as it flows through the system, and it can be stored temporarily in the chemical bonds of organic molecules, but eventually it is released
and dissipated.

C-test passage (-9.49)

A cable modem has two connections: one to the cable wall outlet and the other to a PC or to a set-top box for a TV set. Although a cable modem does modulation between analog and digital signals, it is a much more complex device than a telephone modem. It can be an external device or it can be integrated within a computer or set-top box.

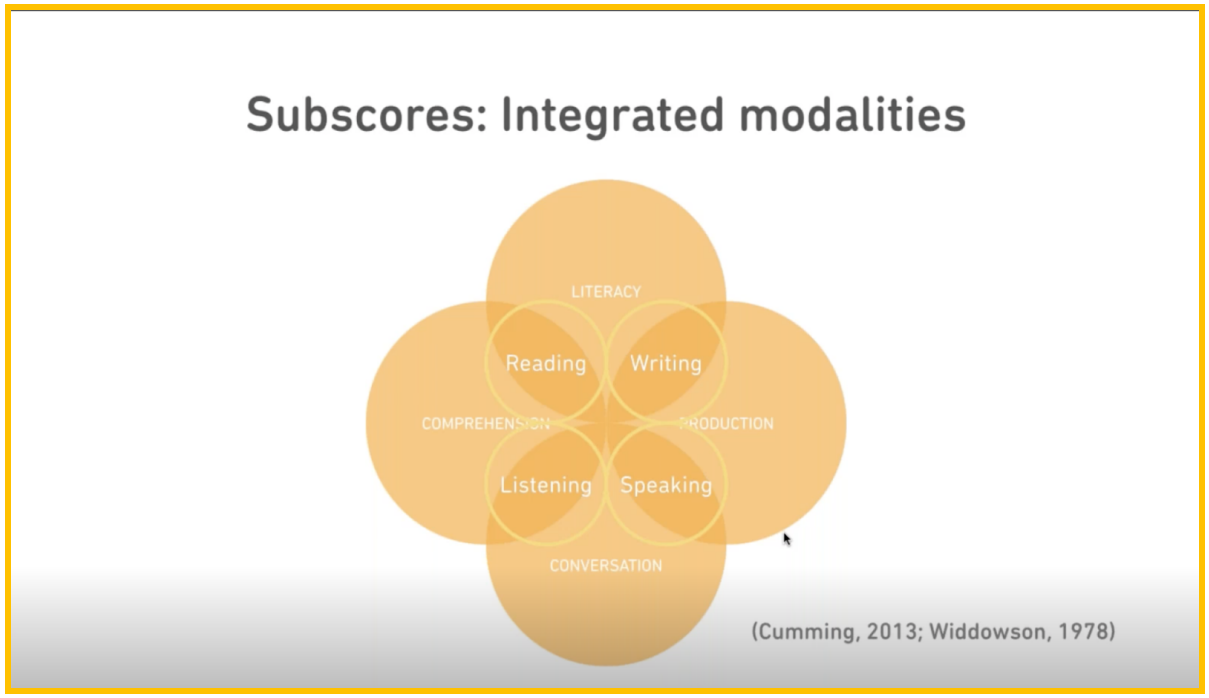
Academic English in C-test Passages

As passages' difficulty estimates increase, the contain more features of academic English



QUESTION
Can you explain the basic concept of the integrated modality subscores used in the Duolingo English Test? (14:08 - 15:05)

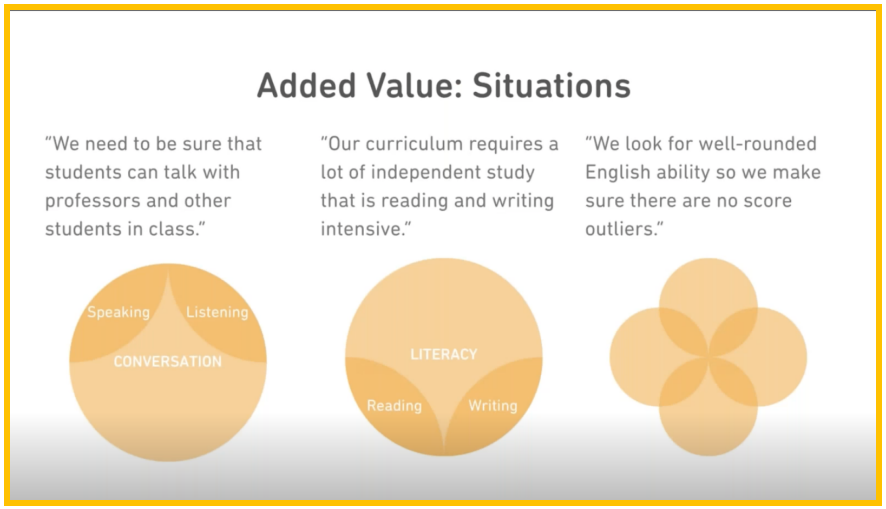
ANSWER
Dr. Geoff LaFlair, Senior Assessment Scientist
The Duolingo English Test provides subscores based on integrated modality scores for Literacy, Conversation, Comprehension and Production. The idea of integrated modalities has been around for a long time. Widdowson first made the case in 1978 in language teaching, and it's been present in the language assessment literature for a long time as well (Cumming, 2013). **The basic premise for their use is that we do not just use language skills to read, write, speak or listen; we use these skills together.** We use them to read and write, a literacy subscore; we use them to listen and speak, which is a conversation subscore etc.

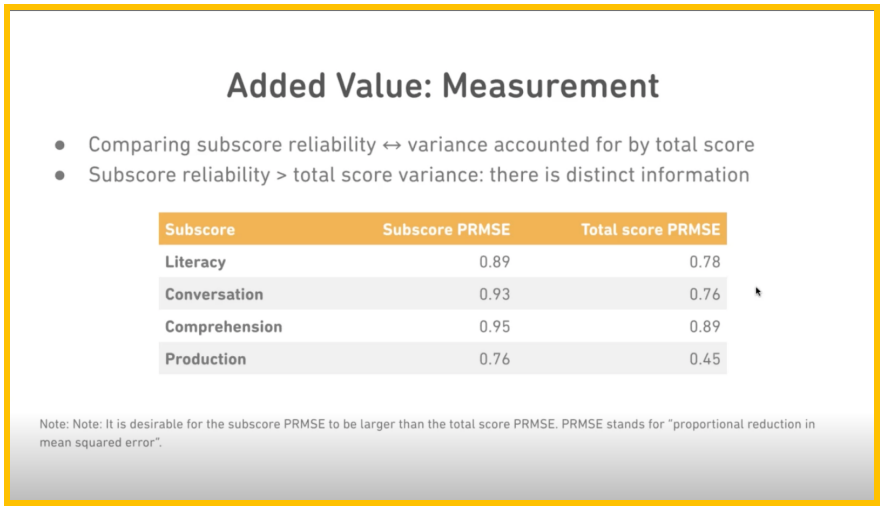


QUESTION
What added value do these subscores provide beyond a composite score or the traditional (SWiRL) subscores? (15:06 - 16:49)

ANSWER
Dr. Geoff LaFlair, Senior Assessment Scientist
There are two ways to answer this. The first is the inferences you can make about test takers based on their subscores. With a conversation subscore you are able to make inferences about how well a student can **both** listen and speak in real scenarios - e.g. with a professor or with other students in class. With the literacy subscore, about how they can read **and** write, for example, on their assignments. **The integrated subscores allow you to better understand how well rounded a test taker's English language abilities are.**

Another way is from a measurement perspective, e.g. do the subscores provide added measurement value above and beyond the total score? As part of our research for the subscores, we used the PRMSE method to look at this. This compares the reliability of the subscore with the amount of variance of the subscore that is explained by the total score. If the Subscore PRMSE is larger than your total, then you know that the subscore is providing added value.





Test Item Development (17:02 - 24:52)

Dr. André Horie, Head of Engineering
Dr. Mancy Liao, Assessment Scientist

QUESTION
Can you provide a summary of how items are generated for the Duolingo English Test? (17:22-20:53)

ANSWER
Dr. André Horie, Head of Engineering
Using the example of our C-test items, **the Duolingo English Test item generation pipeline uses both advanced AI and human expertise to generate test items** (see slide Our item generation pipeline). The utilization of AI in our process represents the main innovation in our approach to testing, but the Duolingo English test is also dependent on expert 'human' intervention. **AI is essential for generating the exceptionally large item bank we need to provide an online, at-home, on-demand test at the quality point we need**, in order to remove the high barrier to entry of traveling to a test centre.

We have a team which sources materials from several suppliers which we can use under license. We check if the material has the linguistic qualities we are looking for (e.g. is appropriate for the domain of English we are looking to assess), whether the source is of a good quality, whether it has typos or grammatical errors, etc. These checks are made by human experts with the help of signals that are automatically extracted from the text.

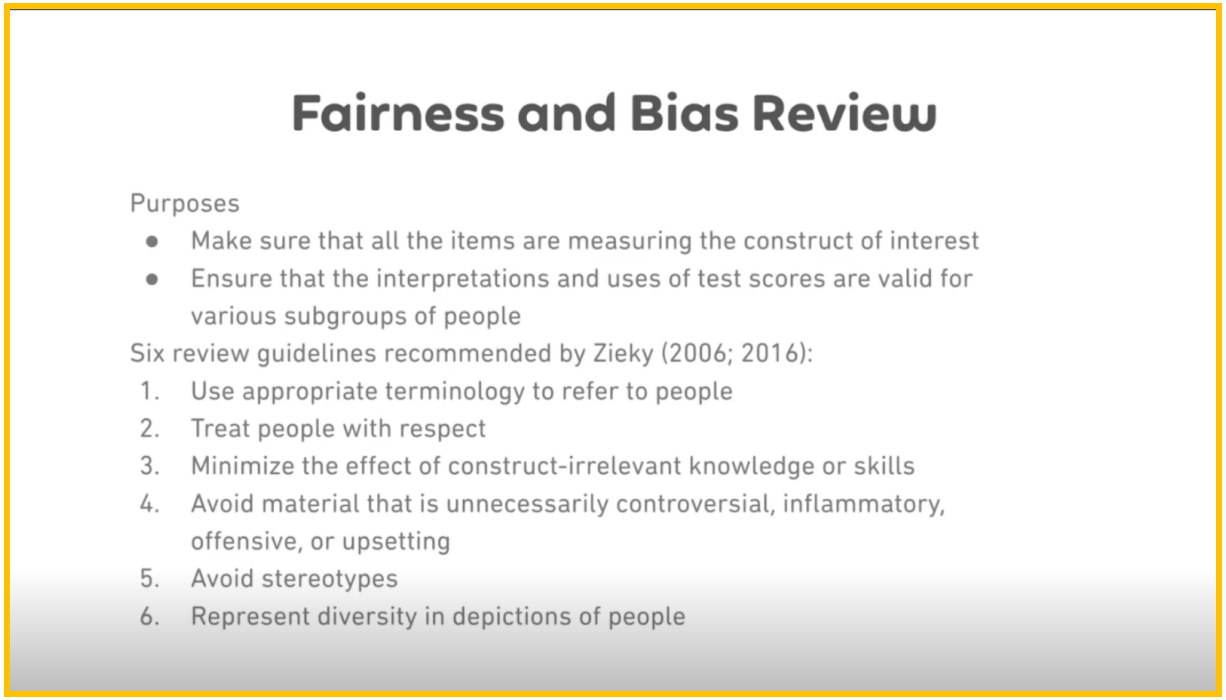
Our experts also review the AI-generated items before they go into the test, they check that the items are valid and that they contain no offensive or exclusionary content. **Our team pilots items and analyzes performance data to ensure that items have the psychometric properties we expect and are working as intended before they are incorporated into the test.**

Our item generation pipeline

Example of pipeline for c-test items

QUESTION
How does Duolingo utilize human experts in processes to create test content and evaluate it for fairness? (21:08 - 22:55)

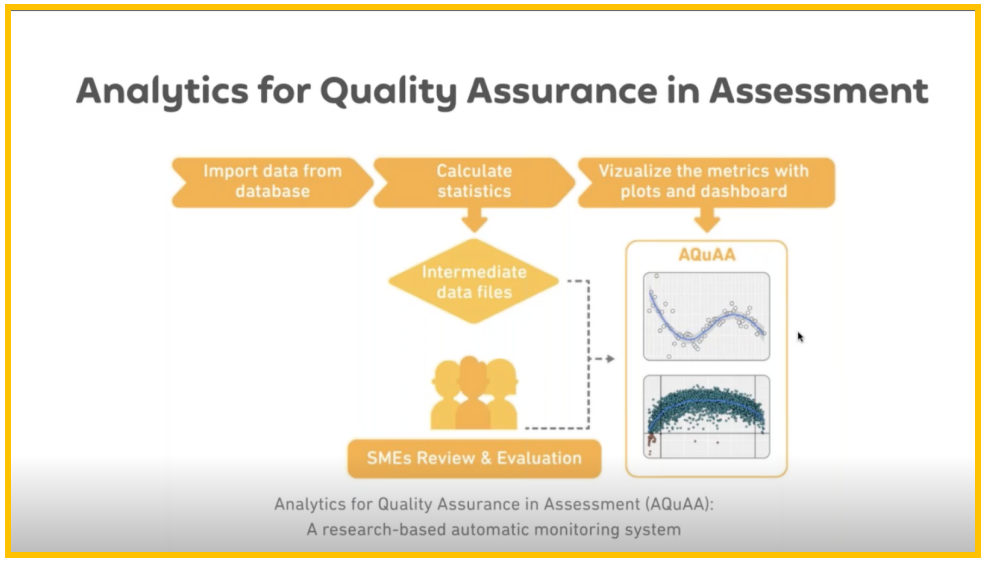
ANSWER
Dr. Mancy Liao, Assessment Scientist
Human review plays an important role in our process, and particularly in ensuring that all of the AI generated items are reviewed by a panel subject matter expert for bias and fairness before they are launched. The review ensures that items are measuring the Construct of Interest, which in the case of the Duolingo English Test is English proficiency. Through the review we ensure that the interpretation of subscores is valid across various subgroups of test takers. The item content is evaluated to ensure that **people are represented in a respectful way, that appropriate terminology is used to refer to people, and that different groups of test takers can see themselves reflected in the test**. We ensure that certain groups are not over-represented in negative situations and exclude sources which contain overly-specialised jargon, controversial or stereotypical source content.



QUESTION
Does Duolingo actively monitor test item quality after launching items within the test item bank? (23:01- 24:52)

ANSWER
Dr. Mancy Liao, Assessment Scientist
After items are launched and are live in the test, we start to accumulate data on how test takers are performing on the items. We analyse the item response data to ensure the items give everyone a fair opportunity to prove their English proficiency and to measure it accurately. As we expect, the psychometric properties and overall quality of an item tends to diminish over time as items are exposed to more and more test takers. Because of this we continually monitor item difficulty so we can manage the quality of both the item and the item pool over time. This includes continuing to monitor the score based metrics of an item, such as the reliability and overall score and subscores, to ensure that newly launched items impact proficiency scores as intended.

We use an interactive dashboard to track our assessment quality assurance processes (see slide Analytics for Quality Assurance in Assessment). The metrics in this dashboard are updated automatically either daily or weekly depending on the nature of the metrics. Subject matter experts and researchers will review metrics immediately if there are any indicators of an anomaly. Otherwise metrics are reviewed weekly to ensure we can react promptly to any validity issues that may emerge.



Test Scoring (25:04 - 33:30)

Kevin Yancey, Senior Machine Learning Engineer
Ramsey Cardwell, Assessment Scientist

QUESTION

Can the item types on the Duolingo English Test and other English proficiency tests be scored validly with AI? What research supports the validity of scoring English proficiency with AI? (25:04-27:54)

ANSWER

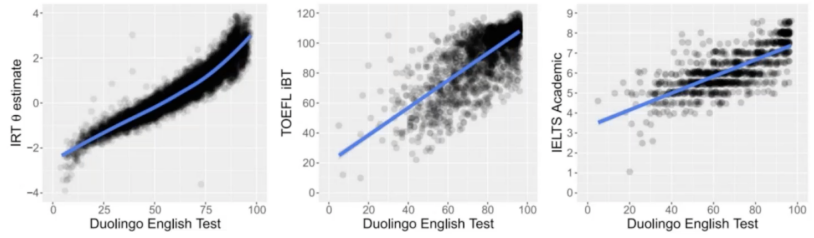
Kevin Yancey, Senior Machine Learning Engineer

First of all, it's important to know that roughly 60% of the Duolingo English Test score is based on objective response items. **These are questions that have an objectively right or wrong answer and are very easy to score by a computer.** These include things like the Yes/No vocabulary, C-test and dictation items that occur on the Duolingo English Test. For these types of items the only thing we need to use AI for is the estimation of the overall item difficulty. Like all Computer Adaptive Tests, we use item response theory to make stronger inferences about a test taker's ability based on their responses to different items. As an example, **we can make stronger inferences about your English language ability if you get a harder item correct than if you got an easier item correct**. We use a combination of Machine Learning and traditional approaches to estimate the difficulties of these items. This process is elaborated on in a TACL paper in 2020 by Settles et al. which shows that this method correlates very strongly with the traditional approach (0.96) while enabling the much larger item banks which enable on-demand testing and are important for the security of the Duolingo English Test.

Roughly another 30% of the DET is based upon what we call constructive responses, which includes the essay writing and free response speaking sections. **These are harder to grade but it's definitely doable by computer.** In fact, automatic essay scoring has been around for decades now, and there is a paper published as far back as 1966 that showed that automatic rater grades could be indistinguishable from those of human grades on a set of essays. We use a method that is similar to ETS's e-rater discussed in Attali & Burnstein's 2006 paper 'Automated Essay Scoring With e-rater® V.2'.

We take the grading rubric or the construct and we break it down into various sub-constructs that we want to base the grading upon. **We look to Natural Language Processing for features that, based on decades of research, are known to correlate strongly with these sub-constructs.** These features work as proxies for these sub-constructs and we can compute a final essay score based on a weighted average of these features.

Settles, et al. 2020, "Machine Learning–Driven Language Assessment," TACL



(a) IRT θ estimates: $\rho = .96$ (b) TOEFL iBT: $r = .74$ (c) IELTS: $r = .75$

Following methodology similar to "Automated Essay Scoring With e-rater® V.2" (Attali 2006)...

Sub-construct	**Example Features**
Relevance - Is the content of the user submission relevant to the prompt?	• Cosine similarity between the response and reference responses defined per prompt (Higgins 2006). • Log-probability of the text as estimated by an n-gram language model trained on a large bank of responses to the item (Attali 2011).
Accuracy - Is the answer free of mechanical / lexical / grammatical errors?	• Number of spelling errors detected via spelling correction normalized by length (Attali 2006). • Number of grammatical errors detected via grammatical error correction (Leacock 2010).
Sophistication - Is the use of words and sentence structure sophisticated and varied?	• Length statistics (e.g., mean word character length, mean sentence token length, number of sentences) (Dong 2016) • Token-type ratio (Attali 2006). • Proportion of A1, A2, … C2, and out-of-vocabulary words as looked up in a CEFR-labeled dictionary.
Organization - Is the organization logical and coherent?	• Coherence - Cosine similarity between sentences (Somasundaran 2014, Foltz 1998). • Detection of introduction and conclusion sentences (Burnstein 2003).

QUESTION
Can AI be as consistent as humans in scoring? Is there any evidence to support that AI can be superior to human scoring in some cases? (27:58 - 30:47)

ANSWER
Ramsey Cardwell, Assessment Scientist
Absolutely. Human raters, the people who assign a numerical score to speaking and writing samples, are 'only human'. **When you do a repetitive task over and over again, the quality can start to shift and there's research on this phenomenon, which is known as 'rater fatigue'**. The people who are scoring essays are doing so for several hours at a time, often day after day, and it is difficult to maintain the same consistency in grading over those long periods of time. **Even when humans are performing at their optimum capacity, we are all subject to subconscious biases which can affect the consistency with which we apply rating criteria**.

A study by Powers et. al in 1992 looked at the scores professional raters awarded essays that were handwritten versus those that were typed. The content of the essays was the same, but raters awarded different scores based on the two modalities; they found that the handwritten essays received significantly higher scores. **Even after training, differences in scoring persisted.**

One way that we evaluate scoring consistency is to look at rating agreement. In traditional cases this looks at how two humans agree with each other when they rate (score) the same essays. We've done similar research comparing how our automatic rater agrees with human scores. In the slide Human vs Machine Consistency you can see that **there is a strong correlation between the human scores and the machine scores**. An interesting finding was that the rate of agreement between two human scores was actually slightly lower than the machine–human rate of agreement.

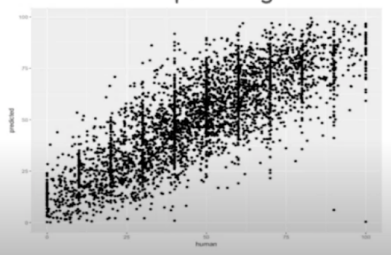
Human vs. Machine Consistency

Human raters are only human

- Rater fatigue (Ling, Mollaun, & Xi, 2014)
- Subconscious biases (e.g., Powers, Fowles, Farnum, & Ramsey, 1992)

Human–Machine rating agreement

Speaking



Agreement

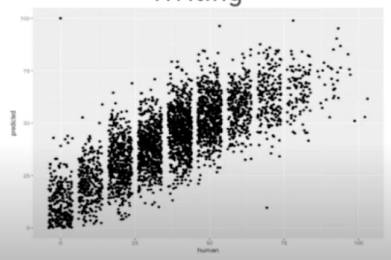
$H{:}H_{\kappa} = 0.77$

$H{:}C_{\kappa} = 0.79$

$H{:}C_{r} = 0.81$

n = 3,966

Writing



Agreement

$H{:}H_{\kappa} = 0.68$

$H{:}C_{\kappa} = 0.82$

$H{:}C_{r} = 0.80$

n = 3,626

QUESTION
Beyond accuracy, reliability, and consistency, are there any other important benefits to AI scoring? (30:51-33:12)

ANSWER
Ramsey Cardwell, Assessment Scientist
Humans are good at certain things, but they're not as good at others - one of those things is synthesizing large amounts of data in order to make a decision. There's a lot of research literature on this across various domains including medicine, hiring, and assessment (e.g., Dawes, Faust, & Meehl, 1989). This research shows that machines outperform humans at what they call 'actuarial' or 'statistical' decision making vs 'clinical' decision making, or rather, decisions requiring human intuition or judgement. **A machine can take into account things like the exact number of unique words in an essay or the relative frequency of every word and incorporate that into a score.** Human raters simply are not able to do that, at least not in an efficient way.

AI scoring is also fully consistent; you program an algorithm to do something and it will just do that - time after time. Any biases that are in that algorithm are therefore easier to detect and rectify. There's a growing public discourse about how there can be biases in AI, which is true, but that bias will be consistent. **We can look at the data and more easily detect those biases and figure out solutions which can be immediately released and implemented consistently.**

Finally, it is very costly to train and employ human raters, and so using machine scoring allows for a much lower cost, and that reduction in cost can be passed on to the test takers, greatly lowering the financial burden on test takers.

Advantages of Machine Scoring

- Synthesizing all available data (Dawes, Faust, & Meehl, 1989; Clauser, 2000)
- Fully consistent, so any biases are easier to detect and rectify
- Much lower cost → less financial burden on test takers

Test Security and Proctoring (33:34 - 43:05)

Basim Baig, Head of Security
Bobby Finnegan, Senior Operations Manager
Kim Snyder, User Operations Manager

QUESTION
Can you explain the concept of 'human in the loop' AI? Why does Duolingo use this kind of approach for test security? (33:34 - 35:58)

ANSWER
Basim Baig, Head of Security
Let's start with the bottom line - all certification decisions that are made on the Duolingo English Test are made by humans. The term 'human-in-the-loop AI', in some ways undersells the role of humans in our processes, really it should be AI-assisted human decision making. Our proctors leverage the output of rule breaking detection algorithms so they can offload the tedious tasks to the machine, where it goes through 80+ signals. They can then focus on the most important signals in the test video itself.

At Duolingo we have a strong belief that optimol test security lies in seamlessly merging both technological and human expertise. As a technology company with a digital test, we sometimes overemphasize the technological aspects of our security, so let's take a moment to recognise that humans are the dominant force in our test security.

Every test is proctored by multiple independent proctors who are not colocated with test takers or each other. So traditional test security problems such as collusion or bribery are instantly mitigated. Each individual proctor leverages technology according to the task that they have to do, but ultimately makes the decision by watching the video of the test. Our team proctor each individual test, and will often slow down or replay videos. Finally, we use multiple rounds of proctoring to eliminate some of those human biases which Ramsey was discussing earlier.

When you think about the security of the test, this is the image I want you to have in your mind. There is a human proctor working on an actual test session. They have a host of digital signals at their disposal in their user interface, as well as the wisdom of the other proctors who looked at the session before them, obviously anonymised. **They take their time, they focus on this individual test taker, not distracted by other test takers and they make a certification decision.** This evidence is ready for any future reviews that need to happen. This is something that is really lacking in the traditional test centre model.

Human-in-the-loop AI

- Bottom line: Humans make all certification decisions
- We seamlessly merge human and machine learning together for optimal decision making.
- AI plays the part of 'clue finder', and humans are the judges.
- Humans do the bulk of the work.

QUESTION
What is Duolingo's philosophy on improving test security over time? (35:59-38:28)

ANSWER
Basim Baig, Head of Security
Security starts with defining policy. If we take the example of test takers trying to pass off as someone else, the 'imposter problem', we first need to define our policy approach to imposters. We can set a policy which says 'test takers must not fake their identities' and 'if they do they will be denied from taking the test again'.

From this policy we then define a view of the world, a threat model, for tackling the imposter problem. It is a framework which describes all of the stakeholders who are involved in this policy: institutions, we don't want to send any fake scores; the test takers themselves, we want to make sure that only people who are imposters get accused; as well as a list of all the other potential issues that come with the imposter problem. **With this understanding of the nature of the threat, in its real world context, and a policy we want to implement, we can develop the security protocols and mechanisms we want to employ to mitigate the threat.**

With our example of imposters, we might want to implement protocols such as identity and document verification. **We will monitor the impact of these mechanisms to see how effective each specific mitigation strategy is, and how they evolve over time with changes to the threat landscape**.

The process for continuously safeguarding and improving security over time is to apply the policy, adjust the threat model and determine the mix of security mechanisms you will need to meet the threat. And then, rinse and repeat. We have a laundry list of threats and imposters are just one. Each of them has an associated policy and security mechanisms that we've either built or are building to safeguard the test. As an agile organisation, we are constantly looking to improve our existing mechanisms or put in place new ones.

QUESTION
Why does Duolingo utilize asynchronous proctoring after test sessions have been completed? What impact does it have on test takers? (38:33-41:35)

ANSWER
Bobby Finnegan, Senior Operations Manager
Multiple rounds of human expert proctors review tests after a test is uploaded and graded. All proctoring is handled within 48 hours of a test being completed and uploaded. We review post-test for a variety of reasons.

Our approach to asynchronous post-test proctoring allows test takers to start a new test at any time that is convenient to them as opposed to waiting for a specific timeframe. We want our test takers to be able to test on their own terms at the time that is best for them.

Additionally, by not utilizing live, real-time proctoring, we reduce the amount of data transferred back and forth between the test taker's computer and Duolingo's servers. This is especially important when it comes to ensuring that the Duolingo English Test is accessible to test takers with slower internet speeds and connections.

By recording a test session and reviewing it post-test, our multiple levels of human expert proctors are able to review specific sections of any given test session recording repeatedly in order to make a precise judgment call. **Proctors focus on one test taker at a time rather than having to keep an eye on a room full of multiple individuals completing assessments concurrently.** Each and every test is reviewed by at least two proctors.

When comparing the Duolingo English Test to a standard test center, there are many advantages to our approach. We allow our test takers to complete the test in the privacy of their own home away from distractions. There are no other individuals testing with you in your space. **With other assessments, mass score cancellation due to the behaviors of other test takers are not uncommon. This type of situation is completely removed from the testing process of the Duolingo English Test.**

Since the test taker is able to test from a private location in their own familiar environment, there is reduced anxiety compared to traveling to a test center, being placed in an unfamiliar space, and completing an assessment in a room with a variety of external distractions. If there is a rule violation, the test taker is also given a clear email noting the specific reason and what needs to be done in order to receive a valid test result.

All in all, we believe that using a post-test asynchronous review with human expert proctors makes the experience as frictionless as possible for our test takers.

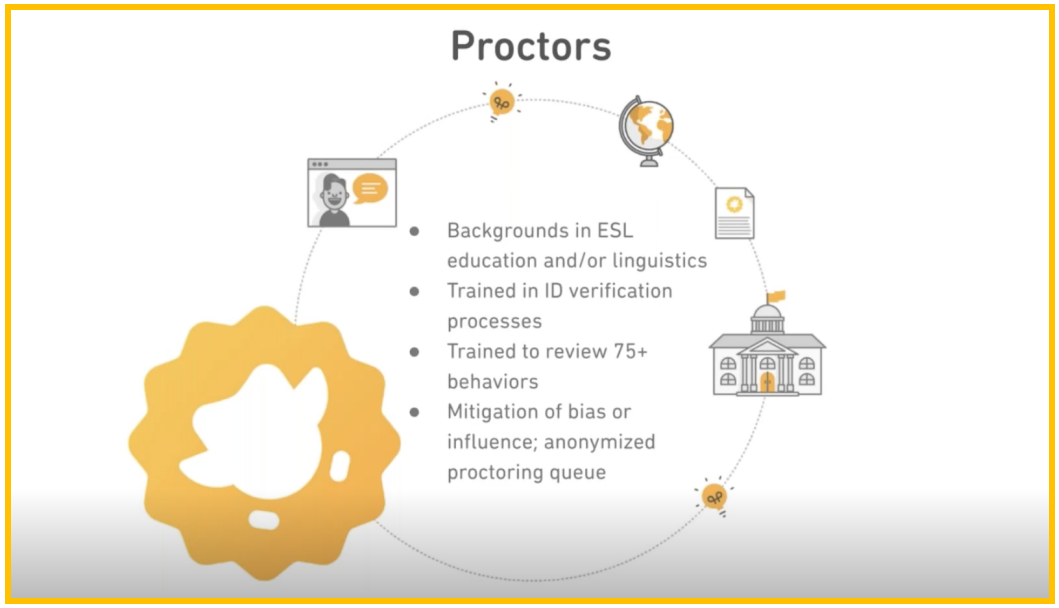
Asynchronous Review

- Accessibility and convenience
- Multi-layer review with the ability to double and triple check specific moments
- A familiar environment with fewer distractions
- Reduced anxiety

QUESTION
Can you describe what Duolingo's proctors are trained to evaluate? What are the benefits of multiple, independent rounds of proctoring? (41:38-43:03)

ANSWER
Kim Snyder, User Operations Manager
Our proctors have backgrounds in ESL education and/or linguistics, and many are current or former ESL teachers. They are trained in ID verification and to look for a variety of suspicious behaviors. Overall, **we train them to look out for anything that would prevent a fair test session or a fair test certification**. Keep in mind, this might not always involve malicious behavior. For example, if a room is too dark to see the test taker's eyes, or the exam audio is not functioning correctly, a proctor will note that the test cannot be certified.

One of the main benefits of multiple, independent rounds of proctoring is that there is the mitigation of bias or influence from external sources, as Basim discussed. The proctoring queue is anonymized and proctors are not permitted or able to search for a specific test taker's test session. In addition, we can train different levels of proctors to focus on observing different parts of the test which leads to specialization and allows us to be more efficient in our proctoring processes.



Additional resources (43:10 - 45:09)

Jeremy Matula, Lead Strategic Engagement Manager

WEBSITE

The Duolingo English Test website offers a range of complimentary resources building on today's session and which dig much deeper into the research underpinning the Duolingo English Test.

Research | Security | Scores | Test readiness

OTHER WEBINAR RECORDINGS

Check out our subject-specific roundtables on research and security.

Research Roundtable | Security Roundtable

SOCIAL MEDIA

Hear the latest from Duolingo English Test by following us on social media.

Facebook | Instagram | Twitter | LinkedIn

COLLABORATIVE STUDIES, RESEARCH GUIDANCE, AND SUPPORT

If your institution is interested, our Assessment Research team can explore predictive validity study opportunities with you. Please contact your Duolingo English Test representative or institutional@duolingo.com for further collaboration, guidance, and support.

Presenters

Jeremy Matula
Lead Strategic Engagement Manager
Jeremy Matula helps lead the team responsible for undergraduate university engagement within the United States. He also works directly with university partners. Jeremy's background includes diverse experience in education technology, university partnerships management, higher education admissions, and sales.

Kevin Hostetler
Senior Strategic Engagement Manager
Kevin Hostetler works directly with our university partners as a strategic engagement manager. Kevin has worked at The George Washington University, University of California, Los Angeles, and University of Southern California, with over a decade's experience coordinating international admission.

Dr. Antony John Kunnan
Principal Assessment Scientist
Antony John Kunnan is a principal assessment scientist in the Duolingo English Test group where he works on research matters. He has published widely; his latest authored book is "Evaluating language assessments" (Routledge, 2018) and edited book is "The Companion to Language Assessment" (Wiley, 2014). His research on fairness, citizenship, and validity have been published in prominent journals and books. He is also co-editor of Language Assessment Quarterly. In past lives, he was a professor at Cal State Los Angeles, and in Hong Kong, Singapore and Macau. More information can be found on his personal website: www.antonykunnan.com

Dr. Geoff LaFlair
Senior Assessment Scientist
Geoff LaFlair is a Senior Assessment Scientist at Duolingo where his primary responsibilities include research and development of the Duolingo English Test. Prior to joining Duolingo, he was an Assistant Professor in the Department of Second Language Studies at the University of Hawai'i at Mānoa and the Director of Assessment in the Center for English as a Second Language at the University of Kentucky. He earned his PhD in applied linguistics, specializing in language assessment, corpus linguistics, and quantitative research methods from Northern Arizona University. His research has been published in Language Testing, Applied Linguistics, The Modern Language Journal, and the Transactions of the Association for Computational Linguistics.

Dr. André Horie
Head of Engineering

André Kenji Horie leads the engineering function at the Duolingo English Test. Prior to that, he played a major role within Duolingo, setting the technical direction for the system architecture and infrastructure of the Duolingo language learning app. He has a computer engineering degree from the University of São Paulo, and a PhD in Natural Language Processing from the University of Tokyo. André has lived in three of the four quadrants of the Earth, and is now in Pittsburgh with his lovely wife and goofy lemon beagle.

Dr. Mancy Liao
Assessment Scientist
Mancy Liao is a psychometrician at Duolingo where she conducts validity research on the Duolingo English Test. Before joining Duolingo, she obtained her PhD in Educational Measurement at the University of Maryland, College Park and her bachelor's degree in Applied Psychology in Sun Yat-sen University. She is passionate about utilizing quantitative methods to facilitate and maintain the validity and fairness of educational assessments. Her research has been published in Educational Psychology Review and Applied Psychological Measurement and she has presented at the NCME and the IMPS conferences among others.

Kevin Yancey
Senior Machine Learning Engineer
Kevin Yancey is a Senior Machine Learning Engineer with 20+ years of software engineering experience. As a former expat and English language teacher, he has a special interest in the applications of AI to human language learning. He got his master's degree in Natural Language Processing (NLP) at Waseda University in Japan. His work in NLP includes automatic L2 vocabulary difficulty estimation and readability estimation.

Ramsey Cardwell
Assessment Scientist
Ramsey Cardwell has over four years of experience working on validity theory and the communication of psychometric research. Prior to joining Duolingo, he interned at the College Board and the Medical Council of Canada where he worked on test accommodations and score reporting. Ramsey is a doctoral candidate in Educational Research Methodology at the University of North Carolina at Greensboro and has also earned a masters in Quantitative Psychology from McGill University and a bachelors in Psychology, Linguistics and Chinese from the University of North Carolina. Between undergraduate and graduate school he taught English as a foreign language for three years in South Korea.

Basim Baig
Head of Security
Basim oversees the Duolingo English Test's security features and has 6+ years of industry experience building engineering solutions at previous online providers like Yelp. He holds a master's degree in cyber security from Stony Brook University, and is a proud immigrant who deeply cares about access within the assessment industry as he needed to complete English assessments himself.

Bobby Finnegan
Senior Operations Manager
Bobby oversees proctoring and support operations for the Duolingo English Test and has been with the team since February of 2017. He implemented the initial processes for multi-layer proctoring of the Duolingo English Test, while serving as the lead of test taker and institutional support as the team was growing. Bobby manages three Operations Managers who specialize in specific areas or tiers of Duolingo English Test operations. Before joining Duolingo, Bobby led Pittsburgh delivery operations for Groupon, managed driver operations for Uber, and served as an extern K-12 school counselor in Philadelphia. He obtained a master's degree in School Counseling from West Chester University, and a bachelor's degree in Public Relations from Point Park University. He lives in Pittsburgh with his husband, two dogs, and four cats.

Kim Snyder
User Operations Manager
Kim Snyder co-manages the international proctoring team for the Duolingo English Test. She also works closely with the engineering team on improving test security. With a background in ESL education, she first became familiar with English language assessments when she was teaching ESL and TOEFL prep classes in Seoul, Korea. In addition, Kim worked as a Duolingo English Test proctor for several years prior to becoming a manager, and she played a large role in the creation and implementation of current online proctoring and ID verification practices. She has also pioneered the proctor hiring and training programs across three tiers of proctoring for the Duolingo English Test.