

February 2024

Peer-reviewed validity + 7 security questions to ask all test providers



Duolingo presenters



Alyson Murray

Sr. Strategic
Engagement Manager,
Canada

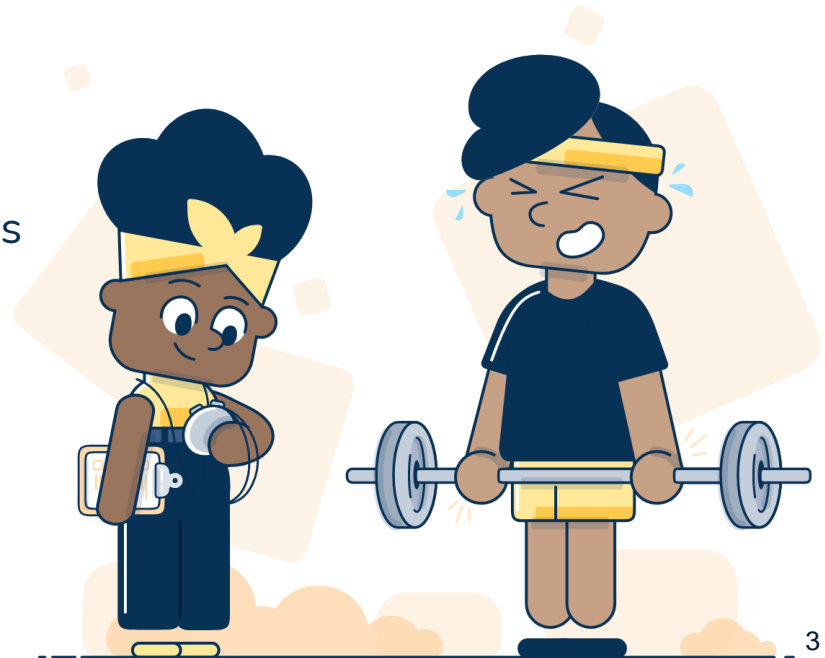


**Masha Kostromitina,
PhD**

Assessment Science
Communications
Manager

Agenda

1. Our mission and overview
1. Validity research overview
2. DET validity research
1. Security questions to ask all test providers
1. Q&A



Our mission

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

It informs everything we do.

The **Duolingo English Test's** mission is to accurately, fairly, and securely assess English language proficiency, while lowering barriers and increasing opportunities for English learners everywhere.

Test Overview



Convenient

Take the test **online anywhere, anytime** — no traveling to a test center or appointment needed.



Affordable

A fraction of the cost of other tests.
**Send results to an unlimited
number of institutions for free!**



Fast

Test takers get **results within 48 hours** of completing the test, and can share it with institutions immediately.

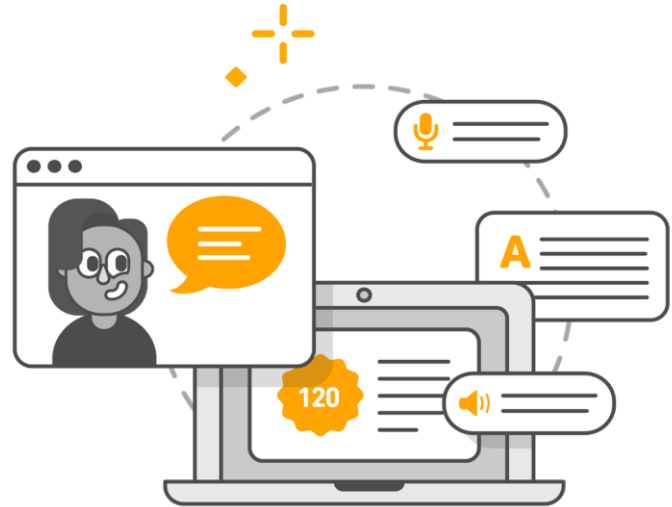
Institutions are automatically notified about shared results, and can review those results immediately.



Comprehensive

Collectively measures **reading, writing, listening, and speaking.**

Integrates proficiency scores, a video interview, and a writing sample.



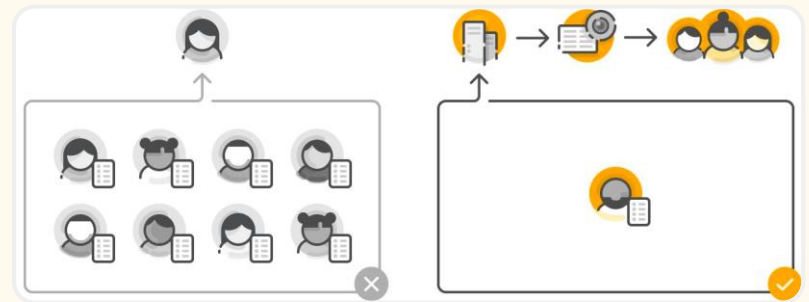
Human-in-the-loop AI security

Expert human proctors, with the help of AI, examine each test session for over **150 different behaviors/environmental factors** over multiple, independent rounds of review to help verify test taker identity, rule adherence, and the legitimacy of certified results from testing.

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

Completed within 48 hours after test submission.



Global acceptance

The Duolingo English Test is accepted by 5,000+ programs around the world, and counting!

US	CA	GB	AU	IE	
3,100+	390+	130+	105+	70+	1,200+

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

2 million+
test takers to date!



duolingo english test

go.duolingo.com/DETxUNHCR

DET Validity Evidence

What is a language assessment?

Assessment

A tool to collect information about language ability according to procedures that are **systematic** and substantively **grounded**

Validity

The extent to which a language assessment **actually measures** what it's supposed to measure (i.e., language ability)

Research to support DET validity argument

- **Content validity**
 - Attali et al. (2022). The interactive reading task: Transformer-based automatic item generation.
- **Concurrent validity**
 - Cardwell et al. (2023). *Considerations when building concordances between English tests.*
- **Predictive validity**
 - Isbell et al. (2023). *Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the DET to the university.*
 - Isaacs et al. (2023). *Examining the predictive validity of the DET: Evidence from a major UK university.*
- **Ecological validity**
 - Kang et al. (2023). *Fairness of using different English accents: The effect of shared L1s in listening tasks of the DET.*

Research to support DET validity argument





How to create a reliable concordance?


- Concordance provides equivalent scores across multiple assessments in order to help in admissions decisions
- Discuss data collection and analysis steps in building the concordance between overall DET scores and TOEFL iBT / IELTS Academic
- Methodological choices in working with data


Brief report

Practical considerations when building concordances between English tests

Ramsey L. Cardwell 
Duolingo, USA

Steven W. Nydick 
Duolingo, USA


J.R. Lockwood 
Duolingo, USA


Alina A. von Davier 
Duolingo, USA

Abstract
Applicants must often demonstrate adequate English proficiency when applying to postsecondary institutions by taking an English language proficiency test, such as the TOEFL iBT, IELTS Academic, or Duolingo English Test (DET). Concordance tables aim to provide equivalent scores across multiple assessments, helping admissions officers to make fair decisions regardless of the test that an applicant took. We present our approaches to addressing practical (i.e., data collection and analysis) challenges in the context of building concordance tables between overall scores from the DET and those from the TOEFL iBT and IELTS Academic tests. We summarize a novel method for combining self-reported and official scores to meet recommended minimum sample sizes for concordance studies. We also evaluate sensitivity of estimated concordances to choices about how to (a) weight the observed data to the target population; (b) define outliers; (c) select appropriate pairs of test scores for repeat test takers; and (d) compute equating functions between pairs of scores. We find that estimated concordance functions are largely robust to different combinations of these choices in the regions of the proficiency distribution most relevant to admissions decisions. We discuss implications of our results for both test users and language testers.

LANGUAGE TESTING

Language Testing
1–11
© The Author(s) 2023

 Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322231195027
journals.sagepub.com/home/ljt



Corresponding author:
Ramsey L. Cardwell, Duolingo, 5900 Penn Ave., Pittsburgh, PA 15206, USA.
Email: ramsey@duolingo.com

How did we build our concordance?

- Combined self-reported and official scores to meet recommended minimum sample sizes for concordance studies.
- Considered several factors in calculating the concordance
 - Weight the observed data to the target population;
 - Outliers;
 - Selection of appropriate pairs of test scores for repeat test takers;
 - Computational aspects of score comparisons (weighting of scores).

Focus on methodology

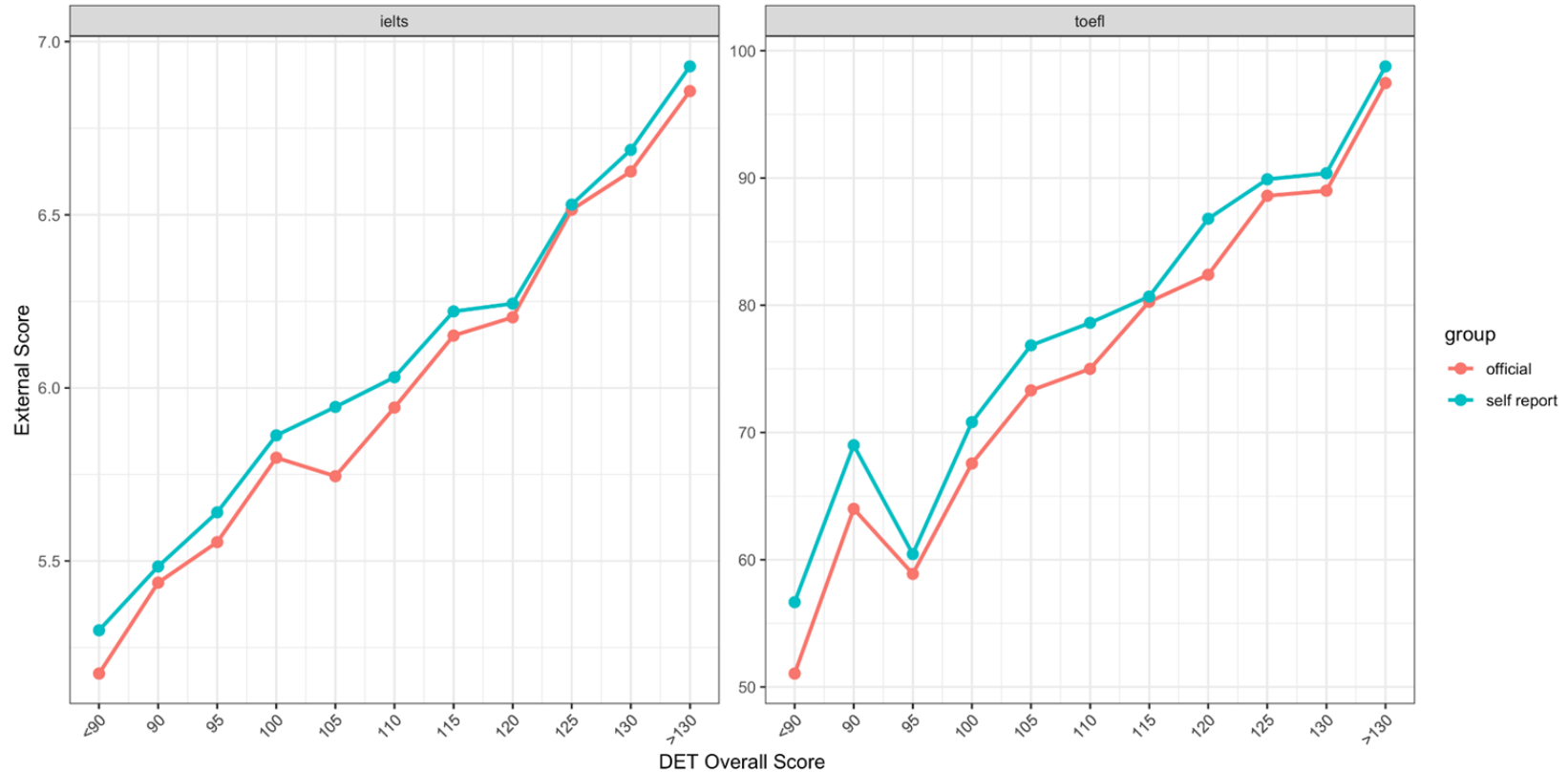
- We made robust methodological choices that were statistically and psychometrically motivated.
- A clear accounting of decisions about data acquisition and analysis, and a demonstration of robustness to such decisions, should be best practice for concordance studies.
- Our methodological transparency allows stakeholders to understand the origin of concordances.

Getting official score data

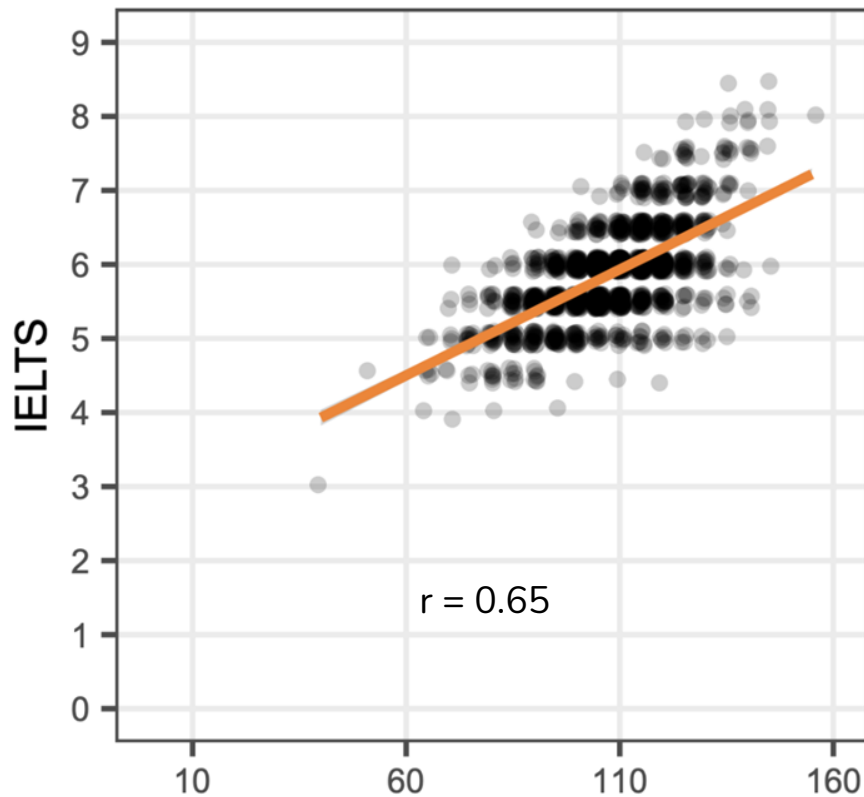
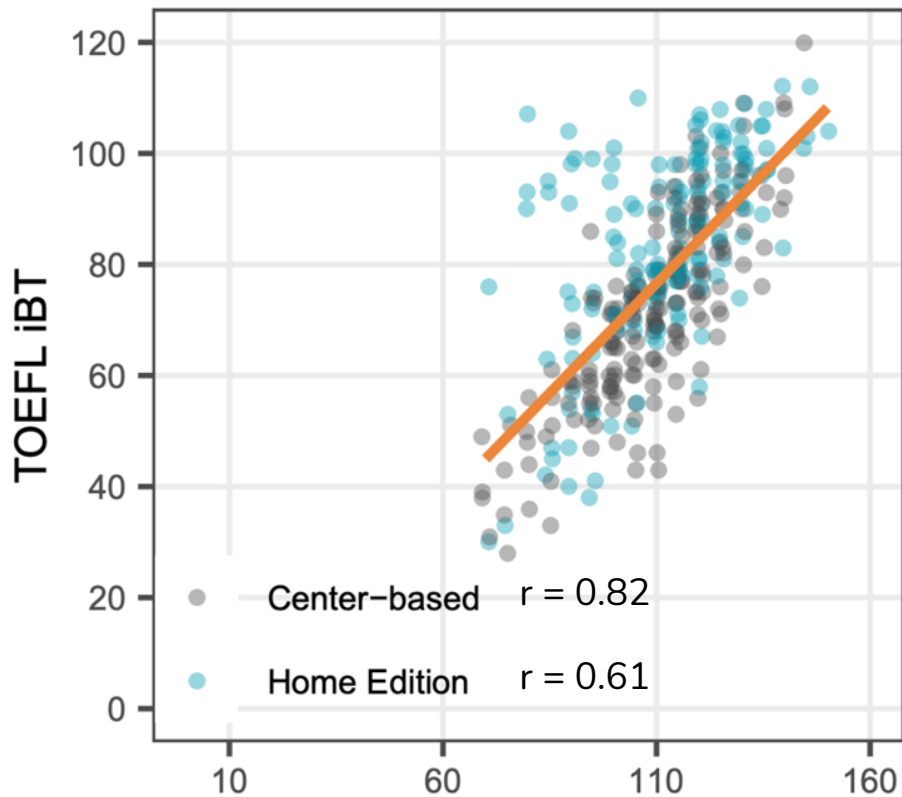
- We set up a process for soliciting official TOEFL/IELTS score reports
- May–July 2022: former DET test takers who met eligibility criteria were invited to submit official TOEFL or IELTS score reports
- Eligible submissions were compensated \$35 or free retest
- Final analytic sample sizes:

	IELTS	TOEFL
Official score reports	1,643	328
Self-reported scores	4,420	1,095
Overall	6,063	1,423

How did we make sure our data were bias-free?



DET to IELTS & TOEFL concordance



Existing test concordance tables

DET	TOEFL iBT	DET	TOEFL iBT
160	120	105	70—75
155	119	100	65—69
150	117—118	95	59—64
145	113—116	90	53—58
140	109—112	85	47—52
135	104—108	80	41—46
130	98—103	75	35—40
125	93—97	70	30—34
120	87—92	65	24—29
115	82—86	10—60	0—23
110	76—81		

DET	IELTS Academic
160	8.5–9
150–155	8
140–145	7.5
130–135	7
120–125	6.5
105–115	6
95–100	5.5
80–90	5
65–75	4.5
10–60	0–4


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

Speaking and writing validity research

Check for updates

Article

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 ($r \geq .74-.89$), 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: dlsbell@hawaii.edu

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



DET speaking scores & academic preparedness

204 university stakeholders rated speech samples from 100 DET sessions for:

- Comprehensibility (ease of understanding)
- Academic acceptability
 - Undergraduate study
 - Graduate study
 - Teaching (TAing)
 - Group work

Method

Participants

- 100 speaking samples from DET test-takers
 - Mandarin Chinese (30), Arabic (21), Spanish (20), French (14), Persian (12), English (3)
 - Undergraduate (51), Graduate (49)
 - DET scores between 70-145
- 204 stakeholders to listen to the samples
 - Graduate students (58), undergraduate students (58), faculty members (47), administrative staff (41)

Key findings

- Strong relationship of comprehensibility/acceptability with DET scores
- Faculty were most lenient raters, administrative staff were strictest

	Comprehensibility (r)	Acceptability (r)
DET Overall	0.81	0.84
DET Conversation	0.86	0.89
DET Production	0.74	0.77
DET Speaking Portfolio	0.77	0.80

Implications



- Empirical evidence that DET speaking tasks elicit language samples relevant to the university context
- DET scores allow stakeholders to make accurate judgements about test-takers' future performance in the academic domain

The Impact of Task Duration on the Scoring of Independent Writing Responses

Naismith, Attali, & LaFlair (2023)

Available in the Social Science Research Network repository



Download This Paper

Open PDF in Browser



Add Paper to My Library

The Impact of Task Duration on the Scoring of Independent Writing Responses

44 Pages · Posted: 18 Oct 2023 · Last revised: 19 Oct 2023

Ben Naismith

Duolingo, Inc.

Yigal Attali

Duolingo, Inc.

Geoffrey T. LaFlair

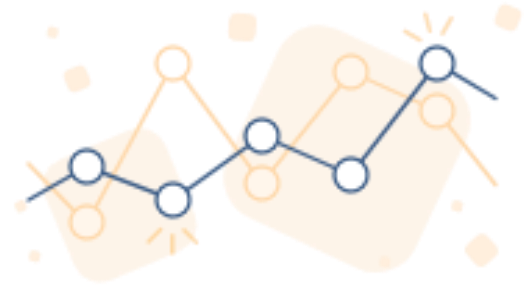
Duolingo, Inc.

Abstract

In writing assessment, there is inherently a tension between authenticity and practicality: tasks with longer durations may more closely reflect real-life writing processes but are less feasible to administer and score. What is more, given total testing time, there is necessarily a trade-off between task duration and number of possible total tasks. Traditionally, high-stakes assessments have managed this trade-off by administering one or two writing tasks each test, allowing 20 to 40 minutes to complete each one. However, research on second language (L2) English writing has not found longer task durations to significantly improve score validity or reliability. Importantly, very few studies have compared much shorter durations for writing tasks (e.g., five minutes or less) to more traditional allotments. To explore this issue, we asked L2-English test takers to respond to two writing prompts (for the same task type) with either 5-minute or 20-minute time limits. Responses were then evaluated by expert human raters and an automated writing evaluation (AWE) tool. With both methods, the shorter task duration yielded scores that were significantly lower but that evidenced equally high test-retest reliability and criterion validity. Implications for writing assessment are discussed in relation to scoring practices and writing task design.

We prioritize efficiency and reliability

- This is especially true for our writing assessment
- DET independent writing tasks are 1-5 minutes
- They target four purposes for writing
 - Description (3 x 1 minute)
 - Narration (5 minutes)
 - Persuasion (5 minutes)
 - Summarization (2 x 75 seconds)
- Other tests usually have two 20-40 minute writing tasks
- Critics question the validity of DET writing tasks: **Do they have a point?**



**How does task duration impact
writing performance and
scoring?**

Effects of longer task duration are puzzling

- Writing assessments have to balance authenticity and efficiency
- Research usually compares time limits between 30-60 minutes
- Longer duration **sometimes...**
 - leads to higher scores (Hale, 1992; Lee et al., 2021)
 - doesn't impact scores (Knoch & Elder, 2010; Livingston, 1987)
- Longer task duration **does not** lead to...
 - more valid scoring (Knoch & Elder, 2010; Power & Fowles, 1996)
 - more reliable scoring (Hale, 1992; Klein, 1981)



Study design:

5-min vs. 20-min writing

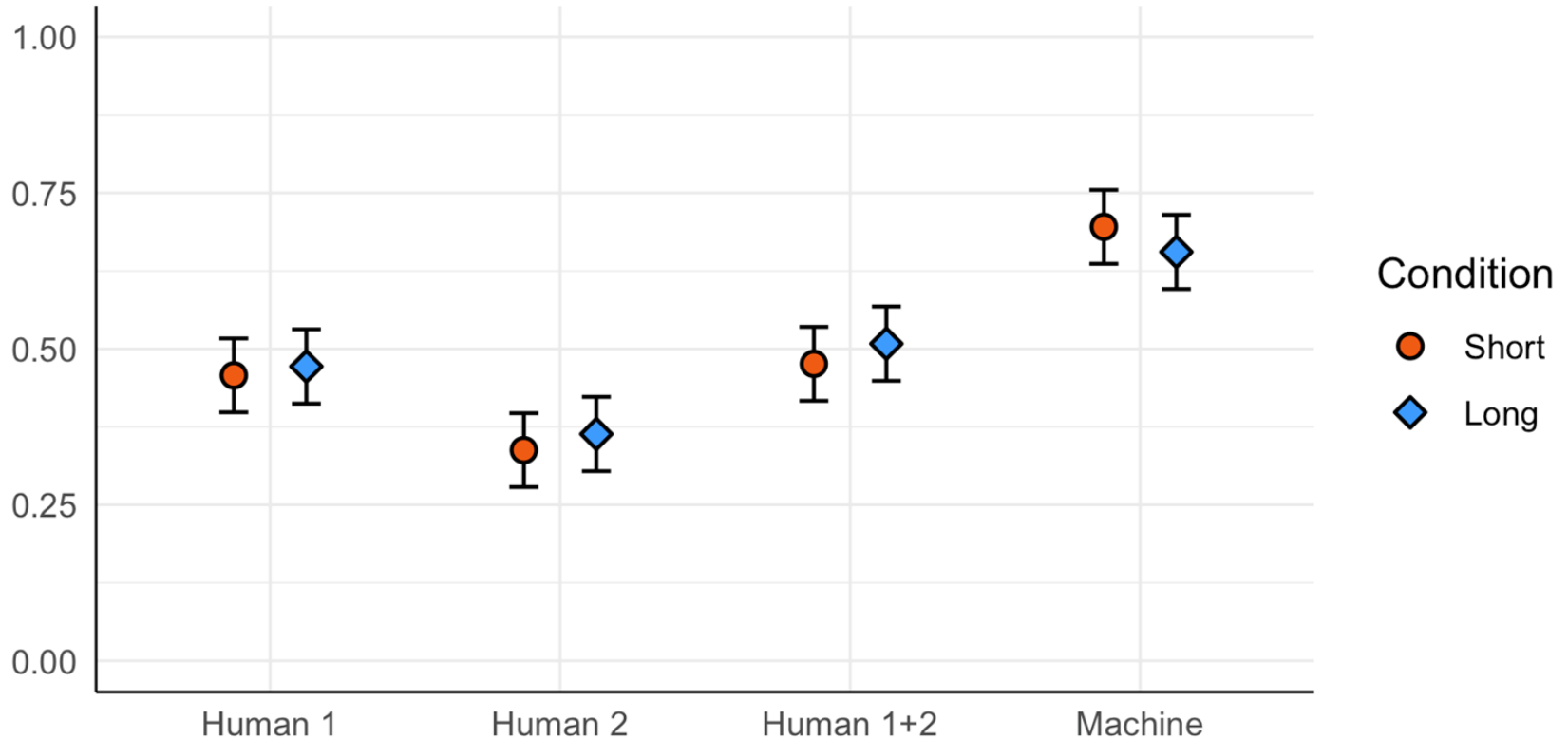
- DET practice test volunteers
- 3 persuasive prompts
- Collected data from **repeaters**
- Machine and human scoring

Total dataset = 1200 responses

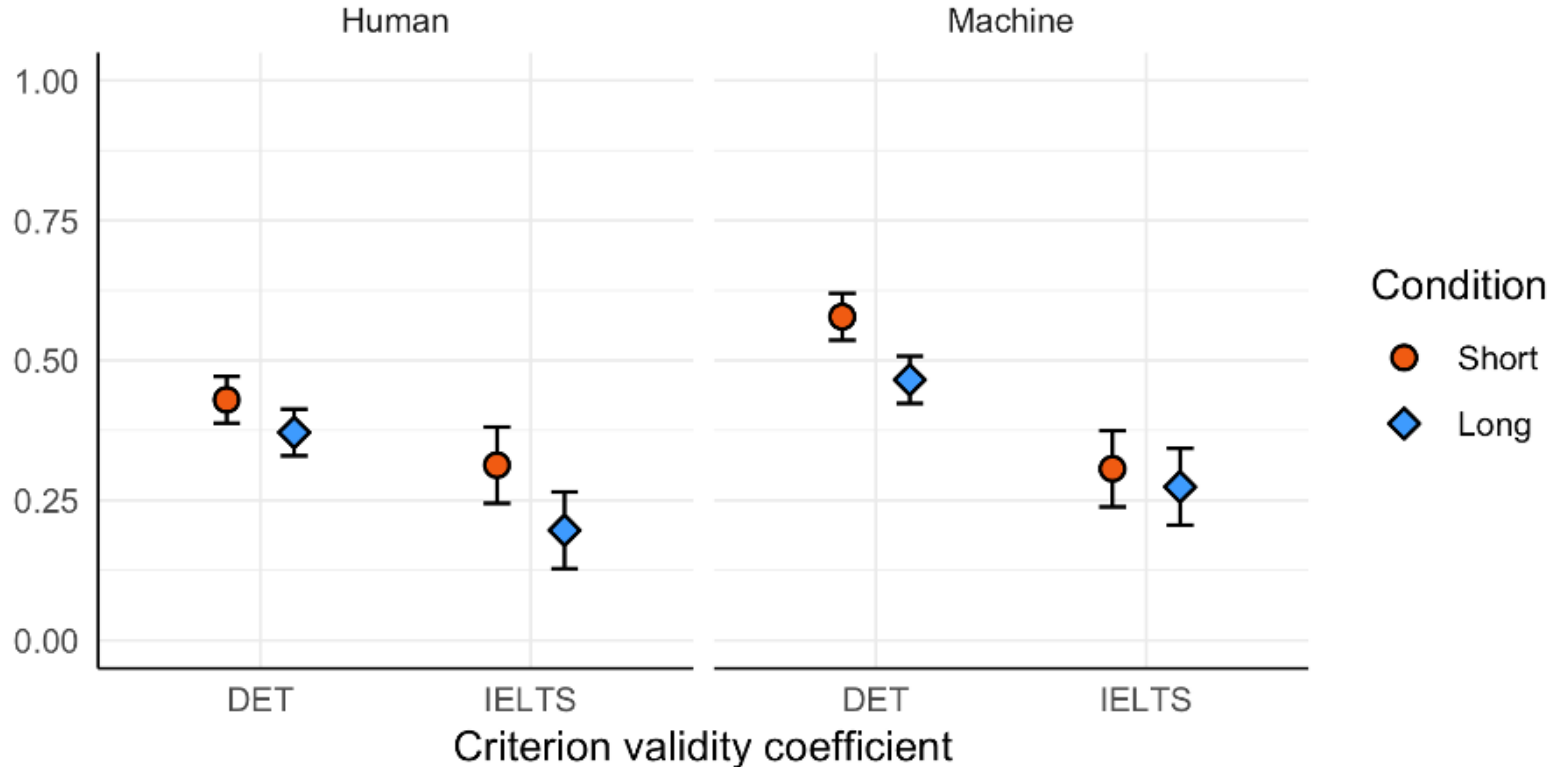
Also collected: keystroke analysis and rater interviews

Results

Reliability (Test-retest reliability)

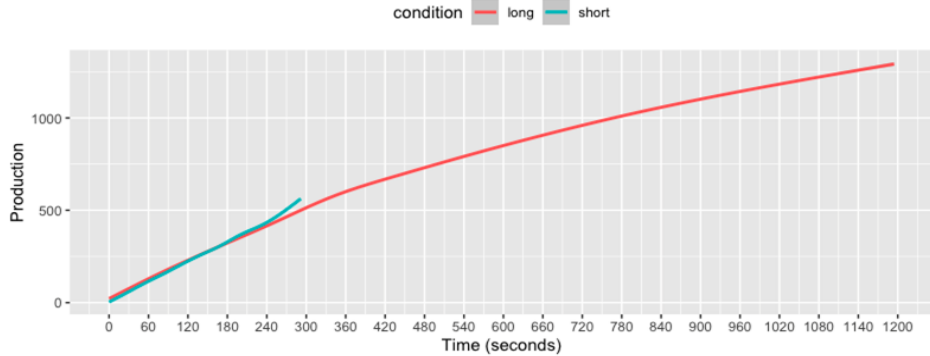


Validity (Correlation with external

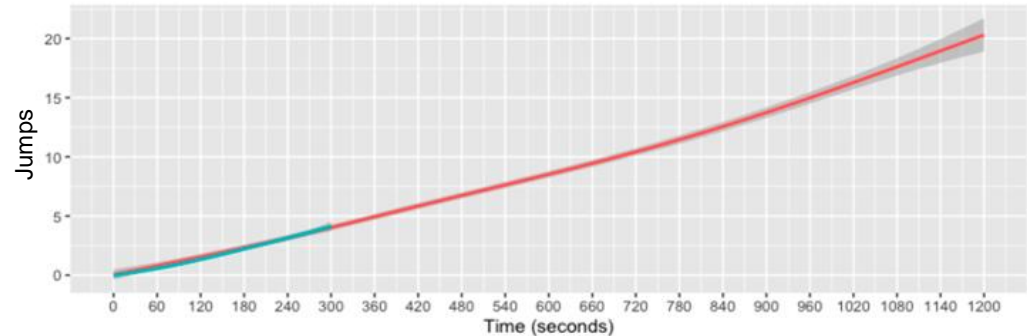


Writing processes (Keystroke analysis)

Writing speed



Editing



Effect of raters not knowing the duration

I think knowing that was five minutes would have probably helped me give a higher award there. My rationale for thinking I would change it was I didn't think there was enough evidence. So in only five minutes, if that's what they came up with, I think that's pretty good. -Rater 1

If I had known some of these responses were produced in the short timeframe, that might have helped with the scoring in their favor. -Rater 2

Implications



In the context of the DET...

- full range of proficiency can be demonstrated in shorter writing tasks
- human- and machine-scoring are seemingly attending to the same aspects of writing
- **but...** it is important to calibrate rubrics, task expectations, and score interpretation because shorter durations lead to lower scores

Predictive validity research

We also look at the predictive validity

- **What?:** A predictive validity study demonstrates the meaning of a test score as a predictive measure of academic success (e.g., DET score in higher education institutions).
- **Why?:** For the DET, predictive validity evidence will demonstrate the test score value to stakeholders.
 - strengthening the test credibility
 - growing acceptance

Our overall results are positive

Based on the data we received from 6 institutions in the US:

- We're seeing comparable results across DET, IELTS, and TOEFL
 - Test score–GPA relationship
 - Good academic standing
- Test score–GPA relationship can differ by
 - Level of study (undergraduate vs graduate)
 - Domain of study (STEM vs non-STEM)
- Small sample sizes of some subgroups (e.g., graduate non-STEM)
 - More data needed for reliable results

7 security questions to ask all test providers

QUESTION 1

How do you externally validate your security?

Have you had independent experts try to cheat on your test?



QUESTION 2

What's your ratio of proctors to test takers?

Do the proctors do it live or are they able to rely on recordings to pause/rewind?



QUESTION 3

**Is your proctoring
done in-house or
outsourced?**



QUESTION 4

If there's a whistleblower complaint on a prior test session, what records do you have to verify the complaint?



How do you use technology to enhance human proctoring?

How do you demonstrate the efficacy and accuracy of this technology?



What is your decertification policy?

Do you review prior test results periodically to reaffirm their validity?

How do you communicate with test takers and accepting institutions when a test is no longer found to be valid?



QUESTION 7

How likely is a test taker able to know the questions they will get ahead of time?



Next steps

NEXT STEPS

Connect with us

We're located in the Exhibition Hall
at booth 10!

Schedule a follow-up chat post-
conference by contacting
alyson.murray@duolingo.com



NEXT STEPS

Stay up to date

 Connect with Duolingo English Test

See the latest research
englishtest.duolingo.com/research



Thanks!
Questions?

