Cutting Edge Quality Assurance

Octavio Hidalgo

J.R. Lockwood

February 2023

Hello! 👋

Octavio Hidalgo
International Engagement

J.R. Lockwood
Assessment

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

The **Duolingo English Test** uses assessment technology to lower barriers and increase opportunities for English language learners everywhere.

duolingo english test development

As a **digital-first test** created with security considerations for online delivery, all test items have to meet two criteria:

1. Be grounded in academic research as predictive of English proficiency;
2. Meet our online delivery security protocols

Duolingo is a leading edtech company with 7+ years in online testing.

We're uniquely able to lead in test security with talent

in AI, machine learning, psychometrics, and humans.

overview

Accessible - Most affordable ELT, taken in 200+ countries

Comprehensive - Includes scores, video and writing samples

Fast - Certified results available in 2 days

Reliable - Research supports use in high-stakes decisions

Secure - Prevention and detection. Adaptive test, browser lockdown, remote proctoring, AI, and more

Scoring and quality assurance

Goals of Test Scoring Team

Ensuring three essential properties of test scores

- **Validity**: The scores measure what they're supposed to measure (i.e., Academic English proficiency).

- **Reliability**: Rapid retakes of the test produce similar scores

- **Comparability**: The meaning of scores remains consistent
  - across test-takers on any given day ("cross-sectional comparability")
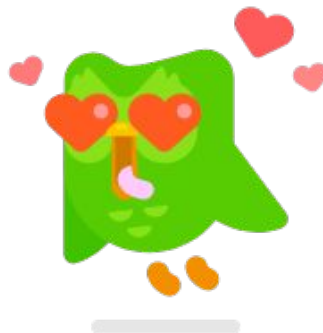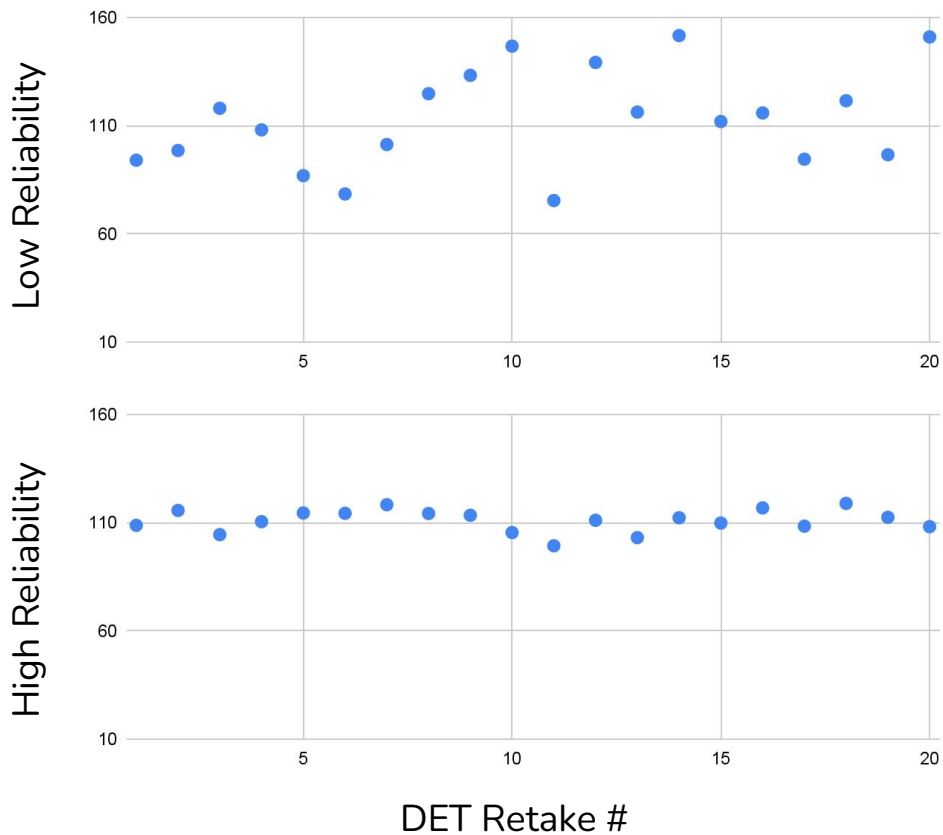 - across time ("longitudinal comparability")

Validity

r = 0.78

Reliability

DET Score

Low Reliability
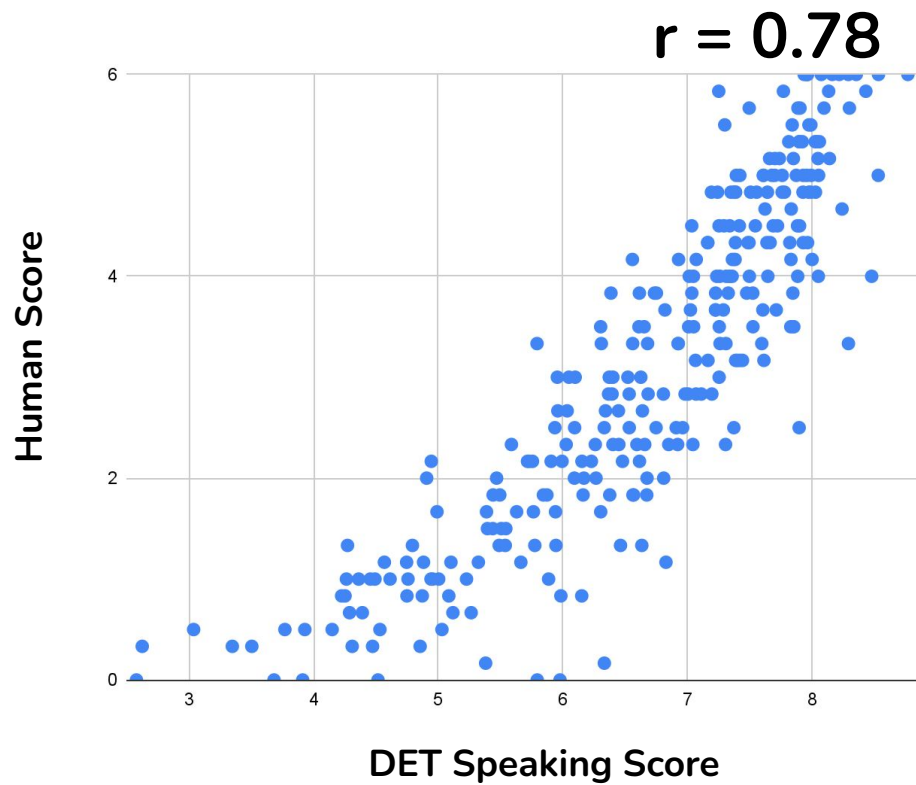
High Reliability

DET Retake #

Maximizing Reliability Under Constraints

- Reliability can be increased by making the test longer, but a key value proposition of the DET is ~1 hour testing time

- Two major strands of work over the last two years increased reliability without making the test longer:
  - Developed methods to extract more information about proficiency from the response data we collect (e.g., novel psychometric models)
  - Improved the adaptive assignment (CAT) algorithm to better match item difficulty to person proficiency

Comparability

- The meaning of scores remains consistent
  - across test-takers on any given day ("cross-sectional comparability")
  - across time ("longitudinal comparability")

Cross-Sectional Comparability

Test-Taker #1

Item R	Item K	Item A	…	Item J

110

Completely Different Questions

Comparable

Test-Taker #2

Item J	Item Z	Item L	…	Item B

120

Disentangling Items from Persons

- Score comparability requires separating item attributes (e.g. difficulty) from person attributes (e.g., proficiency)

- Two complementary operations
  - **Item Calibration**: learning about an item's attributes, controlling for the proficiency of the people who responded to it
  - **Test Scoring**: learning about a person's proficiency from their responses, controlling for the attributes of the specific items that were presented to them

Item Calibration and Test Scoring

- **Item Calibration**: Blend machine learning and psychometrics
  - Machine learning: Use cutting-edge language models to identify characteristics of items related to their psychometric attributes
  - Psychometrics: Combine predictions from these models with response data to estimate item attributes relevant to scoring

- **Test Scoring**: Use psychometric methods to produce the most accurate estimate of a person's proficiency from their item responses

Longitudinal Comparability

- Scores should also be comparable across test takers who take the test at different times
- We routinely improve the test:
    - Expand item banks
    - Improve item selection algorithms
    - Apply cutting-edge methods for scoring spoken and written responses
    - Increase construct coverage through new item types

Making a Better Airplane While Flying

- We need a system to track testing data in (effectively) real time to make sure that everything goes according to plan
  - And be able to remediate quickly if it does not
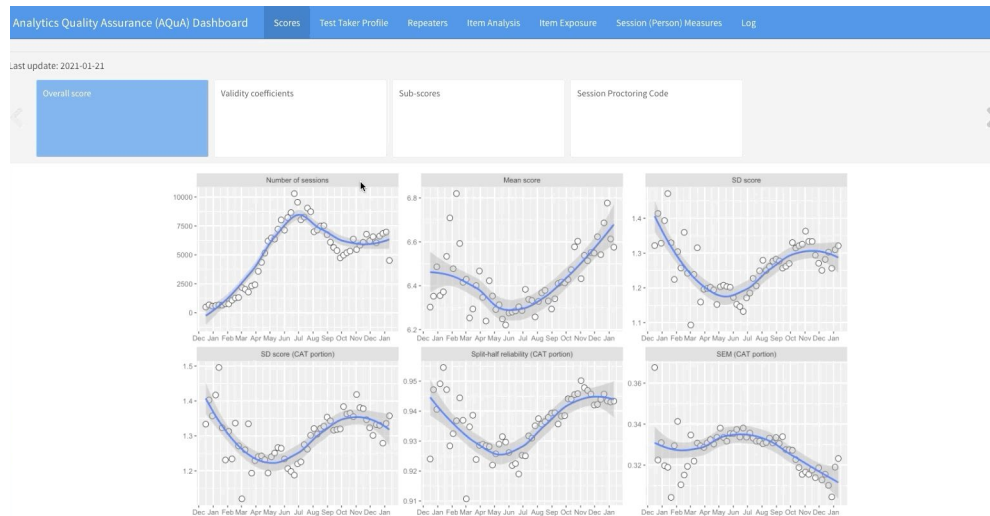- We built a tool to help us monitor statistical summaries of DET testing data:

AQuAA

- **A**nalytics for **Qu**ality **A**ssurance in **A**ssessment (Liao et al., 2021)

What is AQuAA?

A tool for evaluating validity, reliability, and longitudinal comparability of scores

Both a monitoring system and an alert system

Liao, M., Attali, Y., & von Davier, A. A. (2021). AQuAA: Analytics for Quality Assurance in Assessment. In *Proceedings of the Educational Data Mining Conference (Virtual)* (pp. 787-792).

Why do we need AQuAA?

- **To maintain quality of the DET a high-stakes assessment**
  - Ensures that expected quality standards are met at all the stages of the test (Allalouf, 2007)
  - Integrates and monitors the components in the test validity framework
- **To alert us about any score anomalies**
  - Implements statistical methods to identify anomalies
  - Promptly notifies us about any score anomalies
  - Provides contextual information to help us understand the causes of anomalies

| Test security | Domain description | Scoring | Generalization | Explanation | Extrapolation | Utilization |

Why Do We Need AQuAA?

Fact: The composition of the DET test-taker population fluctuates constantly

Consequence: It is difficult to judge whether a fluctuating score trend is a result of changing composition or a score comparability issue

Approach: A daily alert system that flags daily outliers in score metrics
- Use statistical models to predict today's statistic (e.g. average score) from test-taker background variables and recent (30-day) trends
- Raise an alert if today's statistic deviates too much from the prediction

Once an alert is raised:

- Can the alert be explained by a change in test taker population?
- Does the score alert signal an issue in score comparability?

Update AQuAA Data

↓

Review alarms/anomalies in AQuAA

↓

Are the alarms/anomalies caused by validity issues?

No →

Yes ↓

Investigate causes of the issue

↓

Fix the validity issue

↓

Document the validity issue

Improve AQuAA
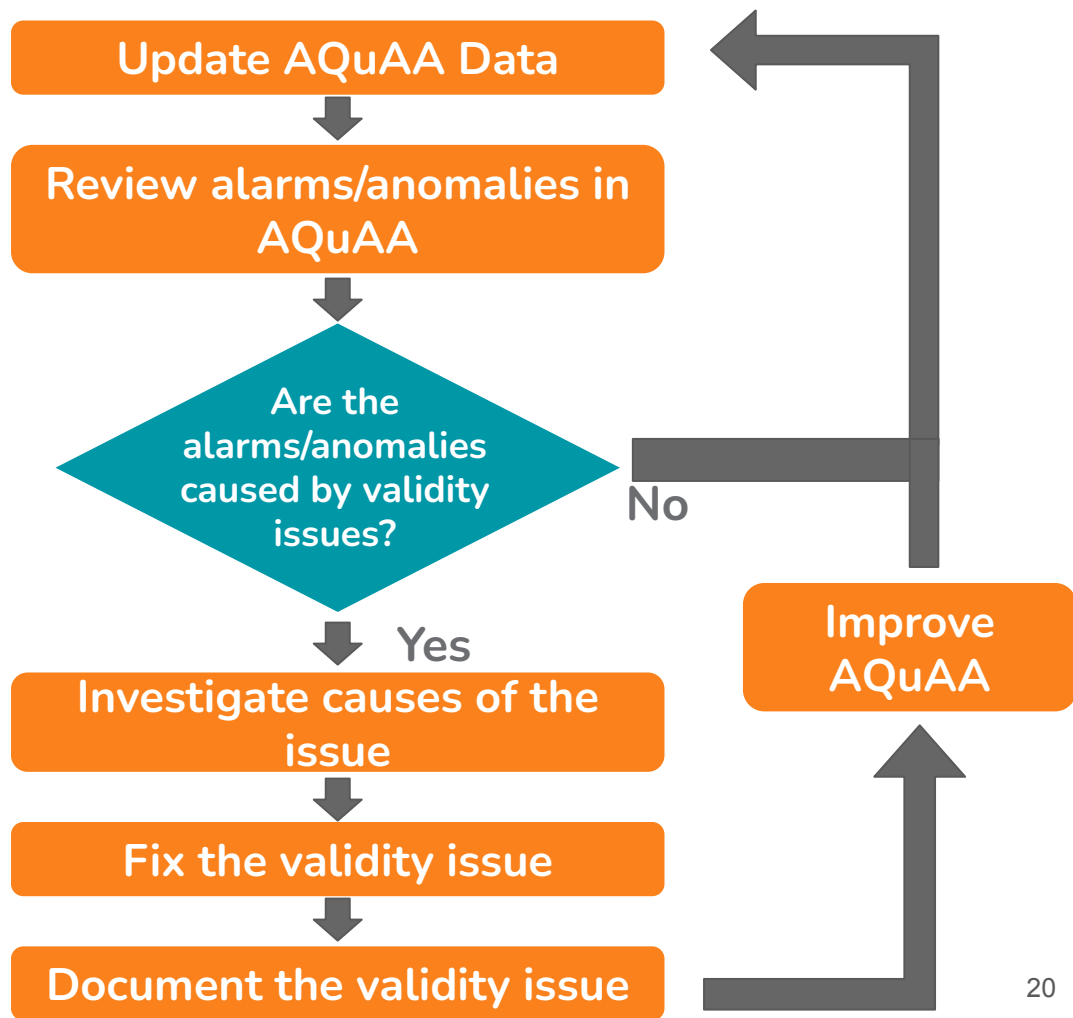
How is AQuAA updated?

The following process is implemented daily:

Import data from database → Calculate metrics → Visualize the metrics with plots and dashboard

Intermediate data files

What data are analyzed in AQuAA?

- Imported from the test database
  - Person-level and item-level scores
  - Background variables of the test takers
- Data are aggregated by week or day
  - Each point in AQuAA represents test sessions of a week or a day
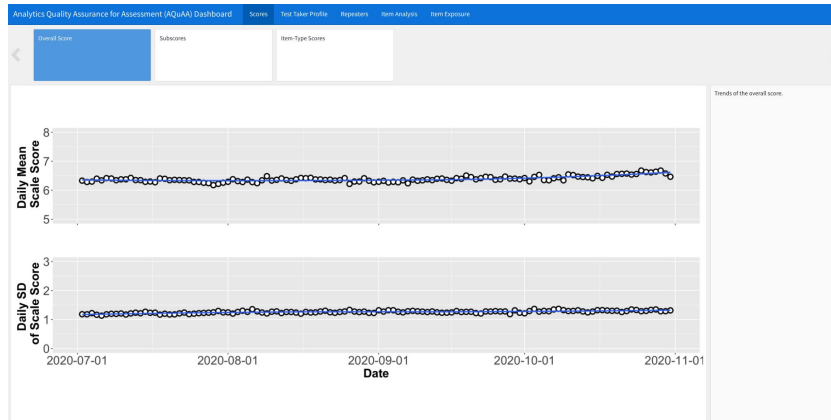
What metrics are included in AQuAA?

Metric Categories

Scores

Test-Taker profile

Repeaters

Scores

Purpose

- Track important summary statistics of the test scores and subscores over time
- Identify & account for seasonality

Example Statistics

- Mean and SD
- Inter-correlations
- Reliability and SEM
- Correlations with other test scores

Example questions

- Do the mean scores in 2022 remain the same as those in 2021?
- Do the mean scores after the launch of test change remain the same as those before the launch?

Test-Taker Profile

Purpose

- Track the change in composition of the test taker population over time

Example questions

- Did any countries experience an abrupt change in test volume over the past weeks?
- Can such changes be explained by admissions deadlines?

Example Test-Taker Background Variables

- Country
- Native language
- Gender
- Age
- Intent

Repeaters

Purpose

- Track prevalence and performance of repeaters (i.e., test takers who took the test more than once)

Example Statistics

- Frequency of repeaters
- Distribution of time between repeated tests
- Distribution of the score differences between repeated tests
- Test-retest reliability

Example questions

- Are scores from different sessions from the same person highly correlated?
- Are daily means for repeaters abnormally different from their past scores?

Takeaways

- AQuAA is designed to constantly monitor performance of the DET system and alert us to any anomalies
  - Helps us to meet professional standards for score reliability, validity, and comparability
- When a new feature (e.g., a new item type) is launched on the DET, we closely monitor AQuAA to ensure that the launch is smooth and the test scores have the appropriate properties

For More Details

More details about the DET, AQuAA, and supporting research at
englishtest.duolingo.com/research

thanks!
questions?

Concordance Process

About Concordances

- Concordances facilitate the interchangeable use of multiple tests

- Concordances historically are updated very infrequently
    - Traditional tests are slow to change
    - TOEFL <> IELTS concordance is from 2010

- TOEFL & IELTS have existed 40+ years
    - Greater familiarity among score users
    - Established score minimums

- DET is newer, and innovates more often, than other tests

Last DET Concordance

1. The last concordance between DET and IELTS/TOEFL was reported in 2019.

2. Reasons for updating concordance:
   a. DET is adopted in many more universities than in 2019.
   b. The DET population has changed quite a bit over the last three years.
   c. People are more familiar with the assessment than they were in 2019.

3. The previous concordance:
   a. Used only self-report data.
   b. Investigated sensitivity to two methodological decision points.

Test-Taker Intentions

	2019	**2020**	**2021**	**2022 (ytd)**
Graduate	10%	38%	36%	38%
Undergraduate	25%	39%	42%	39%

Concordance Challenges

Getting Official Score Data

- We set up a process for soliciting official TOEFL/IELTS score reports.

- From May 16, former DET test takers who met eligibility criteria were contacted via email and a banner on the DET website.

- Eligible TOEFL/IELTS submissions were compensated $35 or free retest

- Final **official score report** analytic sample sizes:

  - TOEFL n = 328

  - IELTS n = 1,643

Reporting Bias (Paired Data)
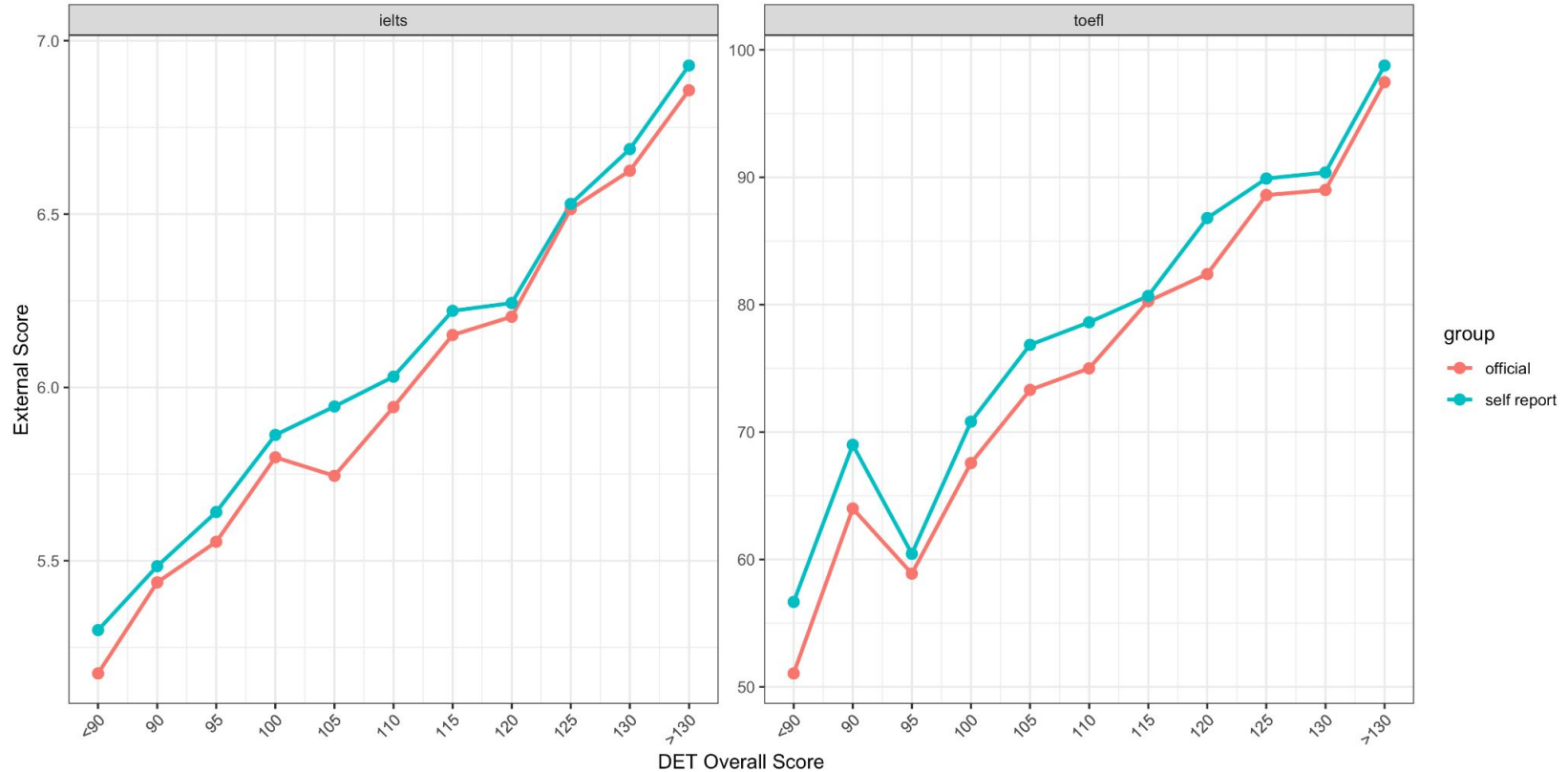
Concordance Results

Determining Concordance Methods

- Considering the construct coverage of the three tests
  - DET, TOEFL, & IELTS measure similar skills, used for similar purposes
  - Built from different frameworks and different test blueprints

- Analyzed sensitivity of concordance results to five methodological decisions

- At different stages of the concordance study, we presented to, and got feedback from, our advisory boards:
  - Technical Advisory Board
  - Admissions Advisory Board
  - Graduate Admissions Advisory Committee

Concordance Analysis Steps

1. Clean data and disambiguate DET/external scores

2. Adjust self-report data for reporting bias and combine with official data

3. Weight and smooth data distributions

4. Estimate concordance using equating methods

Correlations at Each Step

	Official	**Self**	**Combined**
TOEFL	.71	.69	.70
IELTS	.65	.64	.64

Final Concordance Comparison

DET	2019	2022
100	74–79	65–69
105	80–85	70–75
110	86–91	76–81
115	92–96	82–86
120	97–102	87–92
125	103–107	93–97
130	108–112	98–103
135	113–115	104–108
140	116–118	109–112

IELTS	2019	2022
6	95–100	105–115
6.5	105–110	120–125
7	115–120	130–135
7.5	125–130	140–145
8	135–140	150–155

New 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

utilizing the test

Search by Name or Email

SEARCH

ADD FILTERS

13 Results   BUY COUPONS   NOTIFICATIONS   EXPORT

Surname	Given Name	Test Taken	Test Sent	Overall	Literacy	Comprehension	Conversation	Production	Certificate	Interview	
ma		Jul 15, 2019	Mar 20, 2020	100	90	115	105	70	VIEW	VIEW	⌄
Rey	ie	Jul 15, 2019	Mar 20, 2020	120	115	130	110	90	VIEW	VIEW	⌄
Git	an elista	Jul 15, 2019	Mar 20, 2020	125	115	140	120	85	VIEW	VIEW	⌄
Ana	va	Jul 29, 2019	Aug 9, 2019	115	110	130	110	85	VIEW	VIEW	⌄
Din		Jul 17, 2019	Aug 9, 2019	60	50	75	50	20	VIEW	VIEW	⌄
Rey	ie	Jul 15, 2019	Aug 9, 2019	120	115	130	110	90	VIEW	VIEW	⌄
Ha	n	Jul 16, 2019	Aug 9, 2019	105	100	120	100	70	VIEW	VIEW	⌄
Cru	a	Apr 10, 2018	Aug 9, 2019	70 ⓘ	-	-	-	-	VIEW	VIEW	⌄
Zha		Aug 30, 2018	Aug 9, 2019	82 ⓘ	-	-	-	-	VIEW	VIEW	⌄
	Christian	Jul 15	Aug 8								

Q Search by Name or Email   **SEARCH**

▽ ADD FILTERS

14 Results   🏷 BUY COUPONS   ✉ NOTIFICATIONS   ⬇ EXPORT

Surname ⌄	Given Name	Test Taken ⌄	Test Sent ⌄	Overall ⌄	Literacy ⌄	Comprehension ⌄	Conversation ⌄	Production ⌄	Certificate	Interview
ta s n	Dec 10, 2022	Dec 12, 2022	110	105	125	100	70	VIEW	VIEW ⌄
na	E n	Jul 11, 2020	Nov 3, 2022	135	130	145	135	115	VIEW	VIEW ⌄
	A ne	May 24, 2022	May 26, 2022	140	145	150	115	100	VIEW	VIEW ⌄
	M x	May 22, 2022	May 24, 2022	115	120	125	100	85	VIEW	VIEW ⌄
sia	K tc	Jul 29, 2019	Aug 9, 2019	115	110	130	110	85	VIEW	VIEW ⌃

Email
◼◼◼◼◼◼◼◼◼◼

DOB
Jun 29, 2000

Coupon
HONEST999

Date (30/11/2020), AO (Fleisch), Notes: (BA: Marketing Admit)

61/255

		Jul 17, 2019	Aug 9, 2019	60	50	75	50	20	VIEW	VIEW ⌄
	rj e	Jul 15, 2019	Aug 9, 2019	120	115	130	110	90	VIEW	VIEW ⌄
ovikj		Jul 16, 2019	Aug 9, 2019	105	100	120	100	70	VIEW	VIEW ⌄
	i	Apr 10, 2018	Aug 9, 2019	70 ⓘ	-	-	-	-	VIEW	VIEW ⌄
	n	Aug 30, 2018	Aug 9, 2019	82 ⓘ	-	-	-	-	VIEW	VIEW ⌄
	n ista	Jul 15, 2019	Aug 9, 2019	125	115	140	120	85	VIEW	VIEW ⌄
nka	V	Jul 15, 2019	Aug 9, 2019	100	90	115	105	70	VIEW	VIEW ⌄
		Jul 15, 2019	Aug 9, 2019	125	115	135	125	100	VIEW	VIEW ⌄
ar		Jul 15, 2019	Aug 9, 2019	125	110	140	140	90	VIEW	VIEW ⌄

115 ▬▬▬▬▬ 2019-07-29
Score is on a scale of 160

110 Literacy   **130** Comprehension   **110** Conversation   **85** Production

VIDEO INTERVIEW

Explain a quality about yourself that you are proud of. How does this quality make you a better person? What can your friends learn from you because of this quality?

⟲ ▶ ⏭
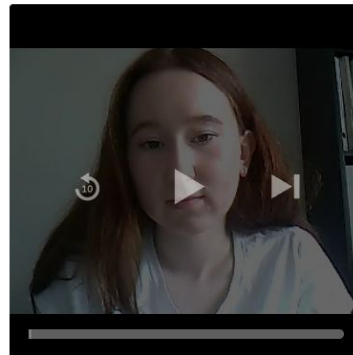
‹ 1 of 2 ›

WRITING SAMPLE

What do you think is the most important thing you need in order to have a happy life? Explain the reasons for your choice.

RESPONSE

The most important thing I need in order to have a happy life is loving people, though they are not things. There is a wrong opinion that money makes people happy and that if you can buy everything you need, thismaterial things will make you happy. It is not true because person as a social creature is in need of true love and caring which only mankind is able to give.

‹ 2 of 2 ›

Score Setting

Score comparison

Understand how Duolingo English Test scores relate to the TOEFL, IELTS, and CEFR. Select each button to compare.

TOEFL **IELTS** **CEFR**

	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

Conditional percentiles of subscores

Overall	Literacy									Conversation									Comprehension									Production								
	10th	20th	30th	40th	50th	60th	70th	80th	90th	10th	20th	30th	40th	50th	60th	70th	80th	90th	10th	20th	30th	40th	50th	60th	70th	80th	90th	10th	20th	30th	40th	50th	60th	70th	80th	90th
65	55	55	60	60	65	65	65	70	70	45	45	50	55	55	60	60	65	70	70	70	75	75	75	75	80	80	80	25	30	35	35	40	40	45	50	55
70	60	65	65	65	70	70	70	75	75	50	50	55	60	60	65	65	70	75	75	75	75	80	80	80	80	85	85	30	35	40	40	45	50	50	55	60
75	65	70	70	70	75	75	75	80	80	55	55	60	65	65	70	70	75	80	80	80	80	85	85	85	85	90	90	35	40	45	45	50	55	55	60	65
80	70	75	75	75	80	80	80	85	85	60	60	65	70	70	75	75	80	85	85	85	85	90	90	90	90	95	95	40	45	50	55	55	60	65	65	70
85	75	80	80	85	85	85	90	90	95	65	65	70	75	75	80	80	85	90	85	90	90	90	95	95	95	100	100	45	50	55	60	60	65	70	75	80
90	80	85	85	90	90	90	95	95	100	70	75	75	80	80	85	85	90	95	90	95	95	95	100	100	100	100	105	55	60	60	65	70	70	75	80	85
95	90	90	90	95	95	95	100	100	105	75	80	80	85	85	90	90	95	100	95	100	100	100	100	105	105	105	110	60	65	70	70	75	75	80	85	90
100	95	95	100	100	100	105	105	105	110	80	80	85	90	90	95	95	100	100	100	105	105	105	105	110	110	110	115	65	70	75	75	80	85	85	90	95
105	100	100	105	105	105	110	110	110	115	85	85	90	95	95	95	100	105	105	105	105	110	110	110	115	115	115	120	70	75	80	80	85	90	90	95	100
110	105	105	110	110	110	115	115	115	120	90	90	95	100	100	100	105	105	110	110	110	115	115	115	120	120	120	125	75	80	85	85	90	95	95	100	105
115	110	110	115	115	115	120	120	120	125	95	95	100	105	105	105	110	110	115	115	120	120	120	120	125	125	125	130	80	85	90	90	95	95	100	105	110
120	115	115	120	120	120	125	125	125	130	100	100	105	105	110	110	115	115	120	120	125	125	125	130	130	130	135	135	80	85	90	95	95	100	105	105	115
125	120	120	125	125	125	130	130	130	135	105	105	110	110	115	115	120	120	125	125	130	130	130	135	135	135	140	140	85	90	95	100	100	105	105	110	115
130	125	125	130	130	130	135	135	135	140	105	110	115	115	120	120	125	125	130	130	135	135	135	140	140	140	145	145	90	95	100	100	105	110	110	115	120
135	130	130	135	135	135	140	140	140	145	110	115	120	120	125	125	130	130	135	135	140	140	145	145	145	145	150	150	95	100	105	105	110	110	115	120	125
140	135	140	140	140	140	145	145	145	150	115	120	125	125	130	130	135	135	140	140	145	145	150	150	150	155	155	155	100	105	110	110	115	115	120	125	130
145	140	145	145	145	145	150	150	150	155	125	130	130	135	135	135	140	140	145	145	150	150	150	155	155	155	155	160	110	110	115	120	120	125	125	130	135
150	150	150	150	150	155	155	155	155	160	130	135	135	140	140	145	145	145	150	150	155	155	155	155	155	160	160	160	115	120	125	125	130	135	135	140	145
155	155	155	155	160	160	160	160	160	160	140	140	145	145	145	150	150	150	155	155	155	160	160	160	160	160	160	160	130	130	135	135	140	140	140	145	150
160	160	160	160	160	160	160	160	160	160	145	145	150	150	150	155	160	160	160	160	160	160	160	160	160	160	160	160	135	140	140	145	150	150	150	155	155

conditional percentiles of subscores

Overall	Literacy 10th	20th	30th	40th	50th	60th	70th	80th	90th	Conversation 10th	20th	30th	40th	50th	60th	70th	80th	90th	Comprehension 10th	20th	30th	40th	50th	60th	70th	80th	90th	Production 10th	20th	30th	40th	50th	60th	70th	80th	90th
105	100	100	105	105	105	110	110	110	115	85	85	90	95	95	95	100	105	105	105	105	110	110	110	115	115	115	120	70	75	80	80	85	90	90	95	100
110	105	105	110	110	110	115	115	115	120	90	90	95	100	100	100	105	105	110	110	110	115	115	115	120	120	120	125	75	80	85	85	90	95	95	100	105
115	110	110	115	115	115	120	120	120	125	95	95	100	105	105	105	110	110	115	115	120	120	120	120	125	125	125	130	80	85	90	90	95	95	100	105	110
120	115	115	120	120	120	125	125	125	130	100	100	105	105	110	110	115	115	120	120	125	125	125	130	130	130	135	135	80	85	90	95	95	100	105	105	115
125	120	120	125	125	125	130	130	130	135	105	105	110	110	115	115	120	120	125	125	130	130	130	135	135	135	140	140	85	90	95	100	100	105	105	110	115
130	125	125	130	130	130	135	135	135	140	105	110	115	115	120	120	125	125	130	130	135	135	135	140	140	140	145	145	90	95	100	100	105	110	110	115	120
135	130	130	135	135	135	140	140	140	145	110	115	120	120	125	125	130	130	135	135	140	140	145	145	145	145	150	150	95	100	105	105	110	110	115	120	125
140	135	140	140	140	140	145	145	145	150	115	120	125	125	130	130	135	135	140	140	145	145	145	150	150	150	155	155	100	105	110	110	115	115	120	125	130
145	140	145	145	145	145	150	150	150	155	125	130	130	135	135	135	140	140	145	145	150	150	150	155	155	155	155	160	110	110	115	120	120	125	125	130	135
150	150	150	150	150	155	155	155	155	160	130	135	135	140	140	145	145	145	150	150	155	155	155	155	155	160	160	160	115	120	125	125	130	135	135	140	145
155	155	155	155	160	160	160	160	160	160	140	140	145	145	145	150	150	150	155	155	155	160	160	160	160	160	160	160	130	130	135	135	140	140	140	145	150
160	160	160	160	160	160	160	160	160	160	145	145	150	150	150	155	160	160	160	160	160	160	160	160	160	160	160	160	135	140	140	145	150	150	150	155	155

Thoughts on Using Concordance Tables

- The DET does not recommend specific score minimums
    - We are happy to discuss factors/methods to consider
- Ideally, concordance tables are one of several factors in setting minimums
- Also want to consider predictive validity evidence
    - We are currently recruiting for a predictive validity meta-analysis

Ongoing Collection of Concordance Data

We are continuing to collect TOEFL & IELTS score reports

- Increase the sample size
- Continue to monitor relationships between tests
- Further validate self-reported data