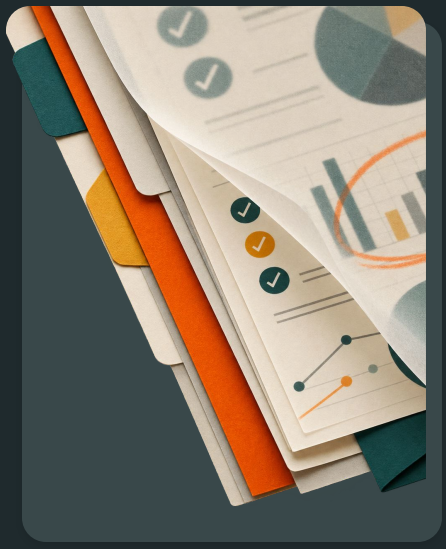
A QUALITY BAR YOU CAN INSPECT.



This is how transparency around responsible AI can—and should—work.

STANDARDS **EVIDENCE** **TRANSPARENCY**

DUOLINGO ENGLISH TEST / RESPONSIBLE AI

THE THESIS

A quality bar is only real if it is public.

The point is not to copy the DET. The point is to make consequential AI governable in public.

A reading path through the Duolingo English Test's responsible AI standards, transparency report, and technical paper trail.

01 STATE THE STANDARD

What do you commit to?

02 SHOW THE EVIDENCE

How do you know?

03 REPORT THE PERFORMANCE

What happened over time?

04 INVITE SCRUTINY

What will change next?

DET makes that loop visible.

Now follow the trail.

THE COMMITMENT

Publish the standard before you ask for trust.

The Duolingo English Test Responsible AI Standards

Jill Burstein • Duolingo Research Report DRR-26-02

AUGUST 21, 2026

> *The Standards are a living document.*
>
> Responsible AI Standards, p. 1

WHY IT MATTERS

A public standard turns values into reviewable commitments: goals, subgoals, human oversight, and an invitation for public comment.

THEN → What does validity mean when the test is digital-first?

FIRST PAGE / SOURCE COVER

The Duolingo English Test Responsible AI Standards

duolingo english test

Duolingo Research Report DRR-26-02
August 21, 2026 (18 pages)
https://englishtest.duolingo.com/research

Jill Burstein

Abstract

The Duolingo English Test (DET) Responsible AI (RAI) Standards are grounded in four ethical principles: Validity and Reliability, Fairness, Privacy and Security, and Accountability and Transparency. These principles are operationalized through concrete practices (Burstein, 2026) that guide AI use across the DET ecosystem, including test design, measurement, security, and validity argument frameworks, and test-taker experience factors (Burstein et al., 2026). These Standards were developed through cross-disciplinary collaboration, drawing on established principles in educational measurement, industry-agnostic ethical guidelines, and responsible AI frameworks. They also reflect Duolingo's internal AI strategy. First published in 2023, the Standards are a living document. In the spirit of promoting ongoing dialogue on ethical AI in education and assessment, this document remains open for public comment.

Keywords

AI for assessment, Duolingo English Test research, language assessment, responsible AI

Corresponding author:
Jill Burstein
Duolingo, Inc. 5900 Penn Ave, Pittsburgh, PA 15206, USA
Email: jill@duolingo.com

Source page shown at right • Research archive: englishtest.duolingo.com/research

VALIDITY

Validity starts with the construct—and the person taking the test.

Improving Test Validity and Accessibility with Digital-First Assessments

Naomi Care & Bryan Maddox • Assessment MicroAnalytics Ltd

2021

> *Digital-first high stakes assessments invite us to rethink the standards, user experience, and vocabulary of assessment validity.*
>
> Care & Maddox, p. 1

WHY IT MATTERS

Digital-first design can reduce construct-irrelevant variation—but only if the construct and user experience stay visible.

THEN → How do you monitor quality when the test runs continuously?

FIRST PAGE / SOURCE COVER

Improving Test Validity and Accessibility with Digital-First Assessments

duolingo english test

Naomi Care & Bryan Maddox
Assessment MicroAnalytics Ltd
Naomi@microanalytics.co.uk • Bryan@microanalytics.co.uk

Abstract

Digital-first high stakes assessments invite us to rethink the standards, user experience, and vocabulary of assessment validity. We explore the disruptive potential of digital-first high stakes assessments to set higher standards in test performance and validity. In this paper we describe the new lexicon that digital-first assessments have introduced into high stakes digital assessments such as *personalisation, user experience,* and *accessibility*. The features of digital high stakes assessments have the potential to reduce sources of construct-irrelevant variation and improve test validity. In conclusion we argue that the distinctive features of digital high stakes assessment challenge our understanding of good assessment design setting new standards for assessment performance and validity.

Keywords

High-stakes assessment, digital-first assessment, user experience, validity, inclusive assessment.

About the authors

Naomi Care has an academic background in education, anthropology, and psychology. As an Analyst with Assessment MicroAnalytics, she leads work in the area of digital ethnography, and on neurodiversity, disability and chronic illness in assessment.

Bryan Maddox is Executive Director of Assessment MicroAnalytics, Professor in Educational Assessment at the University of East Anglia, and visiting professor at the Centre for Educational Measurement at the University of Oslo. Bryan has conducted assessment research in France, Luxembourg, Mongolia, Nepal, Senegal, Slovenia, the United States, and the UK.

Source page shown at right • Research archive: englishtest.duolingo.com/research

RELIABILITY + QA

Continuous administration needs continuous monitoring.

Maintaining and monitoring quality of a continuously administered digital assessment

Manqian Liao, Yigal Attali, J. R. Lockwood & Alina A. von Davier

2022 • Frontiers in Education

> *“Require continuous data monitoring and the ability to promptly identify, interpret, and correct anomalous results.*
>
> Liao et al., p. 1

WHY IT MATTERS

Quality assurance is a living control system, not a one-time validation study.

THEN → Who is the test fair for—and how will you know?

FIRST PAGE / SOURCE COVER

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 19 July 2022
DOI 10.3389/feduc.2022.857496

Check for updates

OPEN ACCESS

EDITED BY
Sedat Sen,
Harran University, Turkey

REVIEWED BY
Hollis Lai,
University of Alberta, Canada
Chad W. Buckendahl,
ACS Ventures LLC, United States

*CORRESPONDENCE
Manqian Liao
mancy@duolingo.com

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 18 January 2022
ACCEPTED 27 June 2022
PUBLISHED 19 July 2022

CITATION
Liao M, Attali Y, Lockwood JR and von Davier AA (2022) Maintaining and monitoring quality of a continuously administered digital assessment. *Front. Educ.* 7:857496. doi: 10.3389/feduc.2022.857496

Maintaining and monitoring quality of a continuously administered digital assessment

Manqian Liao*, Yigal Attali, J. R. Lockwood and Alina A. von Davier

Duolingo, Pittsburgh, PA, United States

Digital-first assessments are a new generation of high-stakes assessments that can be taken anytime and anywhere in the world. The flexibility, complexity, and high-stakes nature of these assessments pose quality assurance challenges and require continuous data monitoring and the ability to promptly identify, interpret, and correct anomalous results. In this manuscript, we illustrate the development of a quality assurance system for anomaly detection for a new high-stakes digital-first assessment, for which the population of test takers is still in flux. Various control charts and models are applied to detect and flag any abnormal changes in the assessment statistics, which are then reviewed by experts. The procedure of determining the causes of a score anomaly is demonstrated with a real-world example. Several categories of statistics, including scores, test taker profiles, repeaters, item analysis and item exposure, are monitored to provide context and evidence for evaluating the score anomaly as well as assure the quality of the assessment. The monitoring results and alerts are programmed to be automatically updated and delivered via an interactive dashboard every day.

KEYWORDS

digital-first assessment, quality assurance, language assessment, high-stakes assessment, anomaly detection

Introduction

Digital-first assessments are assessments designed to be delivered online, at scale, anytime and anywhere in the world, with a rapid turn-around of scoring and an online smooth sharing process with the receiving institutions. Digital-first assessments are administered continuously to individual test takers, in contrast to traditional large-scale assessments that are based on in-person administration to large groups of test takers in fixed locations. The advantages of digital-first assessments have manifested themselves during the pandemic, when traditional group assessments in brick-and-mortar test centers became impractical. However, this increased flexibility could come with challenges in evaluating and maintaining score validity. The test taking population composition of such a new and accessible test could be more sensitive to some external factors (e.g., admission deadlines). For example, on-demand accessibility can cause

Source page shown at right • Research archive: englishtest.duolingo.com/research

FAIRNESS

Fairness work must audit the ecosystem.

Fairness Auditing in a Digital-First Learning & Assessment Ecosystem

Geoff LaFlair, Jill Burstein & Alina von Davier • NCME 2023

APRIL 15, 2023

> *Auditing this content is critical for responsible AI.*
>
> LaFlair, Burstein & von Davier, p. 3

WHY IT MATTERS

Audit content, systems, and outcomes—with human review in the loop.

THEN → Can the score remain trustworthy under attack?

FIRST PAGE / SOURCE COVER

NCME 2023

Fairness Auditing in a Digital-First Learning & Assessment Ecosystem

Presentation at the NCME 2023 Coordinated Session: Expanding the Conceptualization of Fairness for Digital Learning and Assessment

Geoff LaFlair, Jill Burstein & Alina von Davier
April 15, 2023

Source page shown at right • Research archive: englishtest.duolingo.com/research

SECURITY

Security protects the meaning of the score.

Digital-first assessments: A security framework

Geoffrey T. LaFlair, Thomas Langenfeld, Basim Baig, André Kenji Horie, Yigal Attali & Alina A. von Davier

2022 • Computer Assisted Learning

> *These tools and methods are essential to maintaining the security of assessments so that the score reliability is sustained.*

LaFlair et al., p. 1

WHY IT MATTERS

Security is not separate from validity: if score integrity fails, the score interpretation fails.

THEN → Where do measurement theory and AI ethics meet?

FIRST PAGE / SOURCE COVER

Check for updates

Received: 13 July 2021 | Revised: 25 January 2022 | Accepted: 20 February 2022

DOI: 10.1111/jcal.12665

ARTICLE

Journal of Computer Assisted Learning WILEY

Digital-first assessments: A security framework

Geoffrey T. LaFlair[1] | Thomas Langenfeld[2] | Basim Baig[1] | André Kenji Horie[1] | Yigal Attali[1] | Alina A. von Davier[1]

[1]Duolingo English Test, Duolingo, Pittsburgh, Pennsylvania, USA

[2]TEL Measurement Consulting, Iowa City, Iowa, USA

Correspondence
Alina A. von Davier, Duolingo English Test, Duolingo, Pittsburgh, PA, USA.
Email: avondavier@duolingo.com

Abstract

Background: Digital-first assessments leverage the affordances of technology in all elements of the assessment process: from design and development to score reporting and evaluation to create test taker-centric assessments.

Objectives: The goal of this paper is to describe the engineering, machine learning, and psychometric processes and technologies of a test security framework (part of a larger ecosystem; Burstein et al., 2021) that can be used to create systems that protect the integrity of test scores.

Methods: We use the Duolingo English Test to exemplify the processes and technologies that are presented. This includes methods for actively detecting and deterring malicious behaviour (e.g., a custom desktop app). It also includes methods for passively detecting and deterring malicious behaviour (e.g., a large item bank created through automatic generation methods). We describe the certification process that each test administration undergoes, which includes both automated and human review. Additionally, we describe our quality assurance dashboard which leverages psychometric data mining techniques to monitor test quality and inform decisions about item pool maintenance.

Results and Conclusions: As assessment developers transition to online delivery and to a design approach that places the test taker at the centre, it becomes increasingly important to take advantage of the tools and methodological advances in different fields (e.g., engineering, machine learning, psychometrics). These tools and methods are essential to maintaining the security of assessments so that the score reliability is sustained and the interpretations and uses of test scores remain valid.

KEYWORDS
computational psychometrics, digital-first assessment, machine learning, proctoring, test security

1 | DIGITAL-FIRST ASSESSMENTS: SECURITY CONSIDERATIONS

Large-scale testing programmes have traditionally been designed to enhance score reliability and the validity of score interpretations. For this reason, tests were administered under strict standardized conditions, and post-test evaluations required test takers to wait several weeks before receiving scores. Testing today is undergoing a transition where testing agencies are designing new models of administration and scoring (Dadey et al., 2018; von Davier, 2017). Test developers are

Source page shown at right • Research archive: englishtest.duolingo.com/research

VALIDATION + RAI

Measurement theory and AI ethics meet at the validity argument.

Where Assessment Validation and Responsible AI Meet

Jill Burstein & Geoffrey T. LaFlair • Duolingo, Inc.

2024 • arXiv:2411.02577

> *Validity, reliability, and fairness are core ethical principles embedded in classical argument-based assessment validation theory.*
>
> Burstein & LaFlair, p. 2

WHY IT MATTERS

The framework broadens beyond score accuracy to include human values, oversight, and social responsibility.

THEN → Can the principles travel into a field-facing case study?

FIRST PAGE / SOURCE COVER

Where Assessment Validation and Responsible AI Meet

Where Assessment Validation and Responsible AI Meet

Jill Burstein and Geoffrey T. LaFlair

Duolingo, Inc.

5900 Penn Avenue

Pittsburgh, PA 15206

{jill, geoff}@duolingo.com

Submitted to:
Special Issue of *Language Teaching Research Quarterly*

Source page shown at right • Research archive: englishtest.duolingo.com/research

FIELD-FACING SYNTHESIS

Make responsible AI concrete enough to teach.

Responsible Artificial Intelligence for Test Equity and Quality: The Duolingo English Test as a Case Study

Jill Burstein, Geoffrey T. LaFlair, Kevin Yancey, Alina A. von Davier & Ravit Dotan

2025 • Handbook for Assessment in the Service of Learning, Vol. I

> *Responsible AI practices aim to mitigate risks associated with AI.*
>
> Burstein et al., p. 3

WHY IT MATTERS

A field-facing case study makes responsible AI concrete: opportunities, risks, principles, and practices.

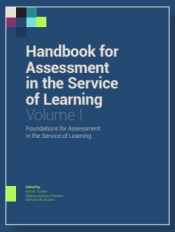
THEN → What does it look like in the applied chapter?

FIRST PAGE / SOURCE COVER

Responsible Artificial Intelligence for Test Equity and Quality: The Duolingo English Test as a Case Study

Jill Burstein, Geoffrey T. LaFlair, Kevin Yancey, Alina A. von Davier, and Ravit Dotan

UMassAmherst
University Libraries



Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

Source page shown at right • Research archive: englishtest.duolingo.com/research

APPLIED PRACTICE

Treat RAI practices as evidence—not slogans.

Responsible Artificial Intelligence for Test Equity and Quality: The Duolingo English Test as a Case Study

Jill Burstein, Geoffrey T. LaFlair, Kevin Yancey, Alina A. von Davier & Ravit Dotan

2025 • Case Study, Vol. I, Chapter 16

> *Test quality is achieved through evidence gathering that confirms an assessment's suitability for its intended purpose.*
>
> Burstein et al., p. 2

WHY IT MATTERS

The key move is to connect RAI practices to both test quality and test equity.

THEN → How do you operationalize the quality loop day to day?

FIRST PAGE / SOURCE COVER

CASE STUDY

Responsible Artificial Intelligence for Test Equity and Quality: The Duolingo English Test as a Case Study

Jill Burstein, Geoffrey T. LaFlair, Kevin Yancey, Alina A. von Davier, and Ravit Dotan

Source page shown at right • Research archive: englishtest.duolingo.com/research

QUALITY ASSURANCE

Operationalize the loop.

AQuAA: Analytics for Quality Assurance in Assessment

Manqian Liao, Yigal Attali & Alina von Davier • Duolingo

Poster • Analytics for digital-first assessment

> *Import data → check and clean → track metrics → identify irregularities → communicate results.*

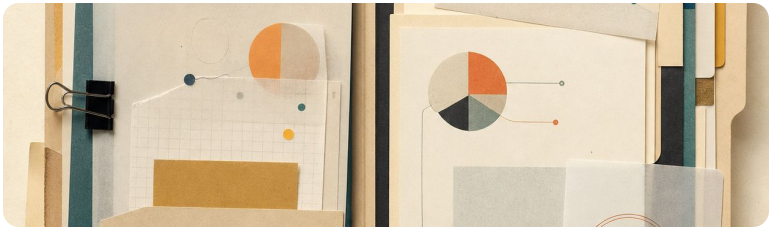
AQuAA poster, process diagram

WHY IT MATTERS

AQuAA makes quality assurance visible as a repeatable loop: monitor, flag, investigate, fix, document, communicate.

THEN → What does performance against the standard look like in public?

FIRST PAGE / SOURCE COVER



Key Terminology

- Analytics for Quality Assurance in Assessment (AQuAA): A quality assurance (QA) system that is designed for digital-first assessment.
- Duolingo English Test: A digital-first, high-stakes assessment from which AQuAA originated.
- Computational Psychometrics: A blend of theory-driven psychometrics and data-driven algorithms (e.g., von Davier et al., 2021); the theoretical foundation of AQuAA.

Motivation

AQuAA is motivated by the Duolingo English Test:

- High-stakes → Have great impact on people's lives → Need rigorous quality assurance (QA) procedure (Allalouf, 2007)
- Digital-first → can be administered anytime and anywhere → more frequent QA with flexible timeline

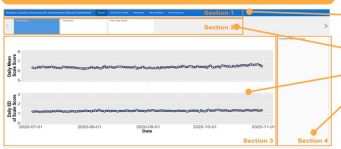
What statistics are tracked in AQuAA?

- Scores (e.g., mean and SD of the test scores)
- Test taker profile (e.g., test volume by gender, country, native language, operating system, etc.)
- Repeaters (e.g., test-retest reliability)
- Item analysis (e.g., item difficulty)
- Item exposure (e.g., item exposure rate)

AQuAA: Analytics for Quality Assurance in Assessment

Manqian Liao, Yigal Attali, Alina von Davier

Duolingo



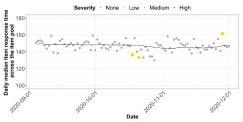
How is AQuAA developed and updated?

The process below is implemented with R:

Import data from database → Check and clean data → Track metrics and statistics → Identify patterns and irregularities → Communicate results

How to identify irregularities in the statistics?

- Take into account the changing baseline due to the changing test taking population
- Flag days when the statistics shift but the shift cannot be fully explained by the change in test taking population

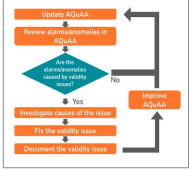


duolingo english test

AQuAA Demo

- Section 1: Navigation bar containing five tabs corresponding to the five categories of statistics monitored in AQuAA.
- Section 2: Pages that correspond to the subcategories of statistics.
- Section 3: The major section of the dashboard where figures are displayed.
- Section 4: The commentary section that display the text description and numerical results.

How is AQuAA used for quality assurance?



Selected References

- Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. Educational Measurement: Issues and Practice, 26(1), 36–46.
- von Davier, A. A., Mislevy, R., & Hao, J. (2021). Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment. Springer. doi: 10.1007/978-3-030-74394-9

Source page shown at right • Research archive: englishtest.duolingo.com/research

TRANSPARENCY REPORT

Then publish how you performed.

The Duolingo English Test Responsible AI Transparency Report

Jill Burstein • Duolingo Research Report DRR-26-08

AUGUST 13, 2026 • REPORTING PERIOD: SUMMER 2024–SUMMER 2026

> *The report serves as a public companion to the DET RAI Standards.*
>
> DET RAI Transparency Report, p. 3

WHY IT MATTERS

Transparency is the evidence return: performance against standards, on a defined reporting period, for real stakeholders.

THEN → What should school systems demand?

FIRST PAGE / SOURCE COVER

The Duolingo English Test Responsible AI Transparency Report

duolingo english test

Duolingo Research Report DRR-26-08
August 13, 2026 (25 pages)
https://englishtest.duolingo.com/research

Jill Burstein

Abstract

The Duolingo English Test (DET) is a high-stakes, digital-first, AI-powered assessment. Universities, employers, and government agencies use DET scores to make consequential decisions about test takers (Naismith et al., 2026). In line with the practice of responsible AI (RAI) transparency reporting in the tech industry, this inaugural DET RAI Transparency Report documents how the DET upheld its commitments to RAI practice during a specific reporting period: Summer 2024 to Summer 2026. It is written with three primary audiences in mind: 1) *test takers*; 2) *institutional stakeholders*; and, 3) the *educational measurement and language testing research communities*. The report further serves as a public companion to the DET RAI Standards (Burstein, 2026)*. It shows how those commitments were upheld through RAI practice. The report is organized around the four DET RAI Standards: Validity and Reliability, Fairness, Privacy and Security, and Accountability and Transparency. Consistent with the DET RAI Standards, the report shows how DET RAI commitments are aligned with established professional guidance, including American Educational Research Association, & American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 2014); the International Test Commission and Association of Test Publishers' [ITC & ATP] Guidelines for Technology-Based Assessment (ITC & ATP, 2025)†; the International Language Testing Association [ILTA] Code of Ethics (ILTA, 2026), and the National Institute of Standards and Technology [NIST] AI-governance framework (NIST, 2023). Finally, the report discusses work-in-progress associated with the DET's commitments moving forward.

Keywords

Duolingo English Test, digital-first assessment, language assessment, responsible AI, transparency report

*While we had three stakeholder groups in mind in writing this inaugural report, the report might also serve others, such as governmental agencies.
†Note that this was updated from the original 2022 version.

Corresponding author:
Jill Burstein
Duolingo, Inc. 5900 Penn Ave, Pittsburgh, PA 15206, USA
Email: jill@duolingo.com

Source page shown at right • Research archive: englishtest.duolingo.com/research

THE TAKEAWAY

School systems should demand this quality bar.

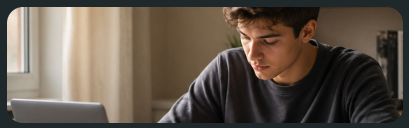
If an AI-powered assessment informs a consequential decision, its standards, evidence, monitoring, and limits should be visible to the people who rely on it.

01 **What are your AI standards?**

02 **What evidence shows performance against them?**

03 **What do you publish—and how often?**

04 **Where do humans intervene, audit, and appeal?**



Read the paper trail → englishtest.duolingo.com/research

A curated reading path through the attached source set.