Author

Jill Burstein, Principal Assessment Scientist, Duolingo English Test

Contributors

Kevin Yancey, Staff AI Research Engineer, Duolingo English Test;
Klinton Bicknell, Senior AI Research Manager, Duolingo;
Carl Gottlieb, Privacy Lead & Data Protection Officer, Duolingo;
Mark Zheng, Senior Corporate Counsel, Duolingo;
Alina von Davier, Chief of Assessment, Duolingo English Test.

Reviewers

Pascale Fung, Chair Professor, Department of Electronic and Computer Engineering, Hong Kong University of Science & Technology
Ramsey Cardwell, Assessment Scientist, Duolingo English Test,
Sophie Wodzak, Senior Communications Manager, Duolingo English Test.
Invitation for public comment

We invite public comment on the Duolingo English Test Responsible AI Standards. These standards were developed to advance thinking in the field of assessment with regard to the ethical and social issues around the use of AI for testing. Further, the standards were informed by the ITC-ATP guidelines for technology-based assessment and ATP guidelines for AI in the testing industry. Embracing thoughtful industry guidelines helped us to formulate these responsible AI standards, which contribute to the Duolingo English Test’s validity, reliability, fairness, and security. The Duolingo English Test assessment research team — composed of experts in computational psychometrics, language assessment, and machine learning — developed the standards in collaboration with experts from our legal and security teams. We believe that public engagement with stakeholders across communities impacted by AI in testing will promote our goal of using AI for good.

Contact us at englishtest-research@duolingo.com with the subject: “Responsible AI Standards”
Contents

Introduction ................................................................. 4

1. Validity & Reliability .................................................. 5
   1.1. Specify processes required to build a validity argument.... 5
   1.2. Evaluate AI used in test item creation, item calibration, and scoring. . 6

2. Fairness ................................................................. 7
   2.1. Specify how the use of AI facilitates test-taker access, accessibility and inclusion. 7
   2.2. Specify test-taker demographic representation, and algorithms known to contain or generate bias. . 7

3. Privacy & Security ................................................... 8
   3.1. Specify methods to ensure privacy and security associated with data origin, data collection and processing, and data management. . 8
   3.2. Specify how to maintain test-taker privacy, item security, and test-taker security during test administration. . 9
   3.3. Specify fair and reliable test security proctoring protocols, item pool development, and psychometric procedures for test security. 10

4. Accountability & Transparency .................................... 10
   4.1. Assess how AI processes impact stakeholders. ............... 11
   4.2. Document AI used for building the validity argument, test item creation, test item calibration, and scoring. ...................... 11
   4.3. Document processes for human-in-the-loop interactions with AI. ....... 12
   4.4. Document human expert qualifications required for human-in-the-loop activities that support AI for the DET. . 13
   4.5. Disseminate research about use of AI to various stakeholder communities. . 13
   4.6. Publish information about how AI is used on the DET, and usage of test-taker data. 13

References ................................................................. 14
Introduction

Artificial intelligence (AI) is now instantiated in digital learning and assessment platforms. Many sectors, including tech, government, legal, and military sectors, now have used formalized principles to develop responsible AI standards. While there is a substantial literature around responsible AI more generally (e.g., Fjeld et al., 2020; Gianni et al., 2022; and, NIST, 2023), traditional validity frameworks (such as, Xi, 2010a; Chapelle et al., 2008; Kunnan, 2000; and, Kane, 1992) pre-date AI advances, and do not provide formal standards for the use of AI in assessment. The AERA/APA/NCME Standards (2014) pre-date modern AI advances, and include limited discussion about the use of AI and technology in educational measurement. Some research discusses AI application in terms of validity (such as Huggins-Manley et al., 2022, Williamson et al., 2012, and Xi, 2010b). In earlier work, Aiken and Epstein (2000) discuss ethical considerations for AI in education. More recently, Dignum (2021) proposed a high-level vision for responsible AI for education, and Dieterle et al (2022) and OECD (2023) discuss guidelines and issues associated with AI in testing. The Duolingo English Test (DET)’s Responsible AI Standards were informed by the ATP (2021) and ITC-ATP (2022) guidelines, which provide comprehensive and relevant guidelines about AI and technology use for assessment. New guidelines for responsible AI are continually being developed (Department for Science, Technology & Innovation, 2023).

The DET is a digital-first assessment. The test uses human-in-the-loop AI for test design, such as for automated generation of test item content; for measurement, such as for item parameter estimation and automated scoring of written and spoken responses; and for test security, such as for test proctoring. Further, human-in-the-loop AI fosters a positive test-taker experience. Examples include: automated scoring of practice tests so test takers can quickly get a sense of their test performance, and use of computer adaptive testing for a shorter test (Burstein et al, 2022). The DET values human expertise as an essential part of decision-making, particularly since the DET has high-stakes implications for both test-takers. Human expertise supports accuracy and fairness of AI system outputs that may impact test security, test development, and measurement of test-taker proficiency. The DET’s theoretical language assessment ecosystem specifies an integrated set of assessment frameworks in which processes and decisions support valid, fair, and reliable test scores for the intended purpose of admissions in English-medium universities (Burstein et al, 2022).

The DET’s use of AI is in support of Duolingo’ mission — to develop high-quality

---

1 See the NIST (2023) fact sheet.
education and provide universal access. The mission reflects the principles of AI for Good – i.e., the DET’s use of AI to support a valid, reliable, fair, and secure test that also promotes a positive test taker experience. To ensure the test’s overall validity, reliability, fairness, and security with regard to use of AI embedded in DET ecosystem processes (Burstein et al, 2022), DET research teams use the Responsible AI Standards to a) conduct on-going auditing of test processes for design, measurement and security; b) conduct validity and reliability studies; and c) maintain documentation for the theoretical, qualitative, and quantitative research required to satisfy the DET Responsible AI Standards.

In the DET Responsible AI Standards, the term “AI” refers to AI systems and AI-adjacent methods\(^2\) and disciplines, such as computational psychometric methods (von Davier, et al., 2022; von Davier, 2017). The DET’s Responsible AI Standards guide research and documentation for validity and reliability, fairness, privacy and security, and accountability and transparency.

### 1. Validity & Reliability

**Rationale:** Validity and Reliability standards are crucial to ensure that the test is suitable for its intended purpose. Validity standards involve evaluating construct relevance and accuracy (Kane, 1992; Kane, 2013), while Reliability standards focus on consistency.

**Goals summary:** To specify processes required to build a validity argument, and to evaluate AI used in test item creation, item calibration, and scoring.

**Goal 1.1. Specify processes required to build a validity argument.**

Processes include theoretical and empirical evaluations that directly inform or address AI used to build a validity argument for test score use.

1. Develop a description for the test target domain – i.e., English language proficiency – to ensure that test items are aligned with the domain being measured.

2. Evaluate AI scoring system accuracy and fairness, leveraging human expertise. Examples include agreement with human raters, accuracy of system features used for scoring constructed responses, and

\(^2\) The concept of AI-adjacent methods refers to computational approaches that may not involve AI.
evaluations of scoring bias.

3. Develop (a) explainable scoring methods, and (b) interpretable AI features used for scoring that have clear alignment with domain constructs.

4. Conduct empirical investigations of item reliability, ensuring reliability for AI-generated items.

5. Evaluate extrapolation through empirical investigation to illustrate relationships between automatically-generated items, test-taker scores, and relevant external measures that suggest proficiency in English skills. Examples of external measures include relationships to other tests, relationships between test-taker’s linguistic input and the target domain.

Goal 1.2. Evaluate AI used in test item creation, item calibration, and scoring.

1. Identify AI methods for item creation, leveraging human expertise to efficiently create valid and reliable test items. An example of human expertise is human review of items from automated item generation.

2. Conduct human evaluations of the quality of items created using AI, such as reviewing outputs from automated item generation.

3. Identify AI methods that can be efficiently used for valid and reliable test item calibration.

4. Conduct evaluations that confirm the accuracy of AI for predicting item parameters (such as item difficulty), leveraging human expertise for quality assurance.

5. Identify AI methods that efficiently produce valid and reliable scores for test-taker responses.

6. Conduct evaluations that confirm the accuracy of AI for scoring test-taker responses, leveraging human expertise for quality assurance.
2. Fairness

Rationale: Fairness standards are required to promote democratization and social justice through increased access, accommodations, and inclusion (ITC-ATP, 2022; Burstein et al., 2022; Cardwell et al., 2022; and Care & Maddox, 2021), represent test-taker demographics, and avoid algorithms known to contain or generate bias (Belzak, 2022; Johnson et al., 2022).

Goals Summary: To specify how the use of AI facilitates test-taker access, accessibility and inclusion; and to specify test-taker demographic representation, and algorithms known to contain or generate bias.

Goal 2.1. Specify how the use of AI facilitates test-taker access, accessibility and inclusion.

1. Identify AI methods to increase test-taker access globally, as part of the DET test-taker experience mission. For example, the DET is available remotely, online, and 24/7. Other access considerations include, but are not limited to, test costs, access to devices, and testing time.

2. Adopt design principles required in compliance with accessibility standards for test takers who require accommodations, due to factors such as low vision, or physical limitations. Ensure that AI or AI-adjacent capabilities do not impact design, such that accessibility compliance might be violated.

3. Develop and apply fairness and bias item review principles for inclusion that eliminate construct-irrelevant barriers, and ensure that cultural and linguistic factors do not impede accessibility and inclusion for the DET test-taker population.

Goal 2.2. Specify test-taker demographic representation, and algorithms known to contain or generate bias.

1. Evaluate and document demographic representation in data sets used to build AI. Documentation should describe how representative (inclusive) the data are with regard to DET test takers. For example, the selection of data for AI system development, such as human rater scoring of written responses, should consider the underrepresentation
of test-taker language groups which could lead to bias in test-taker outcomes.

2. Evaluate and document known algorithmic bias in AI used in DET ecosystem processes (i.e., test security, design, and measurement).

3. Evaluate and document bias associated with automatically-generated item content (e.g., fairness and bias review guidelines), and proficiency measurement.

3. Privacy & Security

Rationale: The Privacy and Security standards are needed to ensure that we (a) comply with relevant laws and regulations governing the collection and use of test taker data; (b) ensure test taker privacy and (c) to ensure secure test administration. (See Liao et al., 2022a; Liao et al., 2022b; LaFlair et al., 2022; Wodzak, 2021; and, Duolingo English Test: Security, Proctoring, and Accommodations, 2021).

Goals Summary: To specify methods to ensure privacy and security associated with data origin, data collection and processing, and data management; to specify how to maintain test-taker privacy, item security, and test-taker security during test administration; and to specify fair and reliable test security proctoring protocols, item pool development and psychometric procedures for test security.

Goal 3.1. Specify methods to ensure privacy and security associated with data origin, data collection and processing, and data management.

1. Ensure that data provenance, governance, and management comply with the Duolingo privacy policy, external privacy policies (where appropriate), and applicable laws such as the European Union’s General Data Protection Regulation (GDPR). (See also the EC Proposed AI Regulation, US National Conference of State Legislatures, 2021).

2. Define and document data requirements with regard to DET intended uses, stakeholders, and the geographic areas where the DET is administered that do not violate privacy terms or security (such as including personally-identifying information without consent).

3. Define, document, and implement methods to ensure that data provenance complies with Duolingo privacy policy, and security policies
with regard to the origin of the data (e.g., open-access corpora, test taker), how it was obtained (e.g., test-taker consent), and changes applied.

4. Define, document, and implement methods to ensure that data governance complies with Duolingo privacy policy and, where appropriate, external privacy or security policies during data collection and processing, including data cleaning, annotation, enrichment, and aggregation, sharing, and use. Document how stakeholder data is used, including but not limited to biometric and personally-identifying data (e.g., IDs for security), process (such as keystroke profiles), and product response data and test scores.

5. Define, document, and implement data management procedures to ensure compliance with Duolingo and, where appropriate, external privacy or security policies.

Goal 3.2. Specify how to maintain test-taker privacy, item security, and test-taker security during test administration.

1. Define, document, and implement methods for test-taker verification in the context of test onboarding to ensure that test-taker identity can be authenticated.

2. Define, document, and implement methods to ensure that verified test-taker identity (i.e., personally-identifiable information) is secure.

3. Define and document test-taking rules, such as prohibiting headphones, except in cases of accommodations, and mitigate actual or perceived cheating behaviors to support test-taker integrity (see Test Rules in FAQs; Test Security Rules).

4. Define, document, and implement test administration processes that mitigate cheating through use of external resources – i.e., test administration through a desktop application.
Goal 3.3. Specify fair and reliable test security proctoring protocols, item pool development, and psychometric procedures for test security.


2. Define and document the algorithm used for proctoring support, proctor training for use of AI, and bias management – i.e., how proctors identify and report perceived AI bias.

3. Define, document, and implement methods (such as, automated item generation) to support scaling of a large, and continuously refreshed test item pool. Methods include human expert monitoring protocols for tracking item exposure and test overlap. Larger item pools mitigate the risk that a single test taker, or multiple test takers are likely to see the same test item, or set of ordered test items during repeated sessions (i.e., a test taker registers for and takes the test multiple times).

4. Define, document, and implement security protocols to prevent item breach from external attackers.

5. Define, document, and implement psychometric procedures, such as test-retest reliability that can track anomalies in test-taker performance that can reveal test-taker cheating behavior.

4. Accountability & Transparency

Rationale: To gain trust from stakeholders, it is essential that the DET have Accountability & Transparency standards for proper governance of AI used on the test. Through documentation and explanations, we are holding ourselves accountable.

Goals Summary: To assess how AI processes impact stakeholders; to document AI used for building the validity argument, test item creation, test item calibration, and scoring; to document processes for human-in-the-loop interactions with AI; document human expert qualifications required for human-in-the-loop activities that support AI for the DET; to disseminate research about use of AI to various stakeholder communities; and to publish information about how AI is used on the DET, and usage of test-taker data.
Goal 4.1. Assess how AI processes impact stakeholders.

Stakeholders include test takers and organizations who use the DET, such as universities.

1. Document how ML algorithms (a) are used on the test and DET support resources,³ (b) are used for test design, measurement and security, and (c) impact stakeholders (i.e., high-stakes decisions based on DET outcomes).

2. Document unintended risks resulting from AI, such as biased scores or construct-irrelevant variance related to design, such as a test-taker’s unfamiliarity with “drag-and-drop” options. Unintended risks may result in negative consequences for stakeholders, such as unfair or inappropriate admissions decisions.

3. Document external factors that result in a need to modify AI. Examples might include: a) institutional policy changes, such as, new language proficiency requirements require new item types; b) modifications in institutional use cases, such as, different rating practices; and, c) demographic changes, such as, increases in particular language groups that might impact differential item functioning (DIF).

4. Document how AI is used for DET stakeholder support. Examples of support include test readiness resources for test takers, and score interpretation guidance for organizations.

Goal 4.2. Document AI used for building the validity argument, test item creation, test item calibration, and scoring.

1. Document theoretical claims, and empirical studies to support the DET validity argument – i.e., evidence that the test is suitable for its intended purpose.

2. For item creation, document rationales for, and descriptions of item generation methods, including data and algorithms, human expert

³ DET support mechanisms include: test readiness resources for test takers, score interpretation guidance for organizations, and how associated AI contributes to impact (e.g., practice test automated scoring is out of sync with certified test scoring.)
processes (such as fairness and bias review), and evaluation methods that provide a clear explanation of system performance metrics, and fairness evaluations, such as mitigating bias with regard to item content generation.

3. For **test item calibration**, document rationales for, and descriptions of AI for predicting item parameters, such as item difficulty.

4. For **test scoring**, document (a) rationales, descriptions, and alignment between item subconstructs and computationally-derived features used for scoring. This is relevant for constructed-response tasks involving spoken and written responses; and, (b) rationales for, and descriptions of measurement methods used to generate DET scores.

---

**Goal 4.3. Document processes for human-in-the-loop interactions with AI.**

These processes include, system development, evaluation, and data preparation and oversight.

1. Document **sustainable processes requiring 100% human expertise** (e.g., data annotation), or human expert supervision, such as fairness & bias review, monitor language model hallucination for automatically-generated item content. Sustainability is measured in terms of time, costs, and required resources (e.g., need for third parties, such as hiring contractors to support human FAB review, and annotation supporting AI development).

2. Document **qualifications of human experts**, such as software engineers, AI researchers, and assessment researchers, who are responsible for supervision during system development and evaluation, and piloting and operational deployment phases.

3. Document **supports to help individuals understand and carry out their responsibilities** in relation to interacting with AI. Supports may include system UX, alert and reporting functions, and rubrics.

4. Document **content to help individuals’ understand how AI is applied on the DET**. Document (1) AI systems’ intended uses, such as text generation, scoring, test security, (2) empirical evaluations and interpretations of AI system behavior, and (3) acknowledgement of potential automation bias – specifically, favoring system outputs.
Goal 4.4. Document human expert qualifications required for human-in-the-loop activities that support AI for the DET.

1. Document qualifications criteria for human experts that are specific to activities that support human-in-the-loop AI. Such documentation may be applied for hiring practices, such as job descriptions.

Goal 4.5. Disseminate research about use of AI to various stakeholder communities.

1. Document research to illustrate how the DET validity argument was constructed with attention to AI.

2. Disseminate research about theoretical, and quantitative, qualitative and mixed-methods research through peer-reviewed, external publications and presentations See DET Research page.

3. Prioritize peer-reviewed, open-access venues and publications, and provide public access to peer-reviewed presentations.

4. Document, disseminate, and update white papers about internal DET research through the DET website, such as DET Research page.

5. Disseminate external media publications, such as blogs, that provide clear and plain language explanations of complex DET processes, such as test security. These publications are intended to render complex concepts transparent and accessible for the broader stakeholder community.

Goal 4.6. Publish information about how AI is used on the DET, and usage of test-taker data.

1. Display on the DET website documentation about how AI is used on the DET so that it is understandable to stakeholders. For example, include content about automated test item creation and scoring, and test security on the DET website. Information should be publicly-available, such as in FAQs on the DET website, or a section devoted to explanations about how
AI are used on the test.

2. Display documentation on the DET website about how stakeholder data is used, including but not limited to biometric and personally-identifying data, such as IDs for security; process, such as keystroke profiles; and, product response data and test scores.

References


