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

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

Abstract

Artificial intelligence (AI) creates opportunities for assessments, such as efficiencies for item generation and scoring of spoken and written responses. At the same time, it poses risks (such as bias in AI-generated item content). Responsible AI (RAI) practices aim to mitigate risks associated with AI. This chapter addresses the critical role of RAI practices in achieving test quality (appropriateness of test score inferences), and test equity (fairness to all test takers)—key principles in this volume. To illustrate, the chapter presents a case study using the Duolingo English Test (DET)—an AI-powered, high-stakes English language assessment. The chapter discusses the DET RAI standards, their development and their relationship to domain-agnostic RAI principles. Further, it provides examples of specific RAI practices, showing how these practices meaningfully address the ethical principles of validity and reliability, fairness, privacy and security, and transparency and accountability standards to ensure test equity and quality.

Introduction

Test quality is achieved through evidence gathering that confirms an assessment's suitability for its intended purpose. *Test equity* is attained when test scores are fair–specifically, they do not favor or disadvantage a particular group. Classical argument-based test validity theory supports test quality and equity by constructing a *chain of inferences* to support test score interpretation. The theory guides the collection and evaluation of evidence to build a validity argument (Chapelle et al., 2008; Kane, 1992). The chain of inferences includes: *domain definition* (task types represent the target domain as defined), *evaluation* (test scores reflect language ability), *generalization* (test scores are reliable), *explanation* (test scores are attributable to the construct), *extrapolation* (test score are related to other language criteria), *utilization* (test scores are interpretable and meaningful for their purpose). While still highly relevant, classical validity theory predates assessments powered by artificial intelligence (AI). To evaluate the validity of interpretations and uses of such assessments, it is essential to evaluate AI capabilities, as they may affect evidence collection, measurement, and ultimately, test quality and equity.

AI-powered assessments are becoming increasingly common, offering many advantages, such as automated scoring of writing and speaking, and efficient creation of larger item banks through automated item generation. However, there are risks. For example, bias in AI-generated item content can impact test-taker outcomes (e.g., Belzak et al., 2025; Johnson et al, 2022); this can, potentially, lead to test inequity and diminished test quality. Therefore, AI-powered assessment calls for alignment with human-centered AI values which are enacted through responsible AI (RAI) guidelines and standards (von Davier & Burstein, 2024; Burstein, 2023; Auernhammer, 2020; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2017). Though some risks to

validity are similar across traditional and AI-based assessments, AI introduces unique risks. Current assessment standards[1] address AI to some extent (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014) (henceforth, AERA&APA&NCME, 2014). However, the expanded use of AI for assessment requires more comprehensive RAI standards and practices to mitigate risks to the test validity argument that affect test score interpretation (Association of Test Publishers (ATP), 2024).

This chapter illustrates how RAI assessment standards and practices help to achieve test equity and test quality of AI-powered assessments. To do so, the chapter presents a case study using the RAI standards for the Duolingo English Test (DET)—a digital-first, high-stakes English language assessment. The chapter (i) presents the standards; (ii) explains their development; (iii) validates the DET RAI standards with ethical principles for an industry-agnostic RAI governance framework—the National Institute for Standards and Technology's (NIST) Artificial Intelligence Risk Management Framework (NIST AI RMF) (NIST, 2023); (iv) illustrates their application on the DET, demonstrating how RAI standards and practices support test equity and quality; and, (v) discusses the RAI standards implications and current known limitations for assessment

Background and Related Work

AI[2] has been used for high-stakes assessments for quite a while for automated writing evaluation (AWE). Its use began on first-generation, computer-based assessments[3] (Lottridge et al., 2021; Shermis & Burstein, 2013, 2003; Foltz et al., 1999; Burstein et al, 1998), and speaking evaluation (ASE) (See Zechner & Evanini, 2019). Emerging more recently, *digital-first assessments* (DfA) were "born digital": DfA's were designed to be administered online and leverage AI for test design (e.g., automated item generation), measurement (e.g., automated essay scoring), and security (e.g., plagiarism detection) (See Belzak et al., 2025; Naismith et al., 2025). DfAs are made possible largely due to the availability of generative AI (OpenAI, 2023; Radford et al., 2019), which enables automated item generation at scale (Attali et al., 2022; Khan et al., 2021).

Classical assessment validity (Chapelle et al., 2008; Kane, 2013; 1992) and fairness (Kunnan, 2000) frameworks embody the ethical principles of validity and reliability, and fairness. While developed for paper-and-pencil and first-generation, computer-based assessments, these frameworks have laid the groundwork for modern assessment. However, they do not explicitly address the use of technology for assessment. The AERA, APA, & NCME (2014) *Standards* address automated scoring of written and spoken constructed responses, covering the scope of AI use at the time they were published.

Earlier frameworks and assessment standards also do not address aspects of AI use on tests that may impact test quality, such as the need for AI literacy training of human test developers, or broader societal issues, such as carbon emissions associated with large language model use (Faiz et al, 2024). It is not even clear how these new issues associated with AI use on assessments would be evaluated in the classical validity argument chain of inferences.

Given the growth of AI use on modern digital assessments, guidelines for RAI in learning and assessment have proliferated (such as ATP, 2024; Organization for Economic Cooperation and Development (OECD), 2023; International Test Commission (ITC) & Association of Test Publishers (ATP) (ITC-ATP), 2025; The International Privacy Subcommittee of the ATP Security Committee, 2021; U.S. Department of Education, Office of Educational Technology, 2023). *Guidelines* make high-level recommendations for mitigating AI risks. In contrast, standards

1 Standards that were available at the time this chapter was written.

2 Note that not all AI used for assessment is generative AI. In this chapter, we use the term AI to refer to AI and AI-adjacent approaches, including natural language processing (NLP) and statistical modeling approaches.

3 First-generation, computer-based assessments were typically first designed for paper-and-pencil formats and later moved to a computer-delivered format.

translate theoretical principles into practical, actionable guidance (e.g., AERA, APA, & NCME, 2014), offering concrete steps for implementing their underlying ideals. Failing to implement the ideals creates ethical debt that may compromise test quality and equity. This can lead to both short- and long-term harms—such as use of AI-generated content containing hallucinations (e.g., Ji et al., 2023).

The development of RAI assessment standards needs to draw from the extensive set of assessment guidelines and standards. Additionally, it should leverage the rich body of AI ethics literature, which outlines domain-agnostic, ethical principles such as fairness, transparency, explainability, privacy, security, trust, responsibility, justice, and autonomy (Memarian & Doleck, 2023; Floridi & Cowls, 2022; Fjeld et al., 2020; Jobin et al., 2019). Domain-agnostic ethical principles promote human responsibility. When combined with industry-specific standards (e.g., AERA, APA, & NCME, 2014) and guidelines (e.g., OECD, 2023; U.S. Department of Education, 2023; ITC-ATP, 2025), they help ensure alignment with the unique needs of educational assessment, supporting both test equity and quality.

Case Study: The Duolingo English Test RAI Standards

This section presents an overview of the Duolingo English Test (DET), and discusses the DET RAI standards, their development process, and their alignment with a domain-agnostic industry framework. Finally, the case study demonstrates how the systematic application of standards supports assessment quality.

The Duolingo English Test

The DET is a digital-first, high-stakes, computer-adaptive measure of English language proficiency, commonly used for admissions to English-medium higher education institutions (Naismith et al, 2025). It assesses a test-taker's ability to use English language skills that are required for speaking, writing, reading and listening, as well as for integrated skills associated with literacy, conversation, comprehension, and production. Integrated skills require multiple proficiencies, e.g., speaking and listening for conversation.

The DET leverages AI extensively, using automated item generation, automated writing and speaking evaluation, and automated plagiarism detection for test design, measurement, and security, respectively. The DET's test-taker experience also benefits from AI affordances. For instance, the DET's free practice test is made possible by automated item generation (i.e., to generate practice test items) and scoring (i.e., providing an instant score estimate).

The DET employs *human-in-the-loop* (HiTL) AI practices to support test quality and equity. The DET's HiTL approach is consistent with current education and assessment policy, aiming to contribute to the test's equity and quality. Recent education and assessment policies discuss HiTL AI as crucial human oversight at critical decision points (ATP, 2024). HiTL AI is also discussed as a real-time, human-system interaction process, whereby humans provide ongoing input to enhance AI performance (Wang, 2019). The DET leverages human judgment and oversight in test design, measurement, and security. For example, humans review automatically generated content during test design, label training data to build and evaluate writing and speaking models for measurement, and serve as proctors to referee AI-generated plagiarism flags for test security. The case study presented later provides illustrations of these current practices.

The Duolingo English Test Responsible AI Standards

The Four Standards

The DET's Responsible AI standards address test equity and quality through include four standards that represent ethical principles aligned with the test's goals. An overview of the standards is discussed below, and more details are provided in Section 3.3.

1. The ***Validity and Reliability*** standard is crucial to ensure that the test is suitable for its intended purpose. The Validity standard evaluates construct relevance and accuracy, and the Reliability standard focuses on consistency;
2. The ***Fairness*** standard promotes democratization and social justice through increased access, accommodations, and inclusion, representative test-taker demographics, and avoiding algorithms known to contain or generate bias;
3. The ***Privacy and Security*** standard ensures (a) compliance with relevant laws and regulations governing the collection and use of test taker data; (b) ensuring test-taker privacy and (c) providing secure test administration; and,
4. The ***Accountability and Transparency*** standard aims to gain trust from stakeholders through proper governance and documentation of AI used on the test.

Standards Development

Five key activities informed the choice of the four ethical principles used to create the DET RAI standards.

First, we conducted a literature review to identify commonly discussed ethical principles in the context of AI (e.g., Memarian & Doleck, 2023; Floridi & Cowls, 2022; Fjeld et al., 2020; Jobin et al., 2019). The review increased our understanding of which principles were applicable to the DET.

Second, to validate alignment between domain-agnostic, AI governance (e.g., NIST, 2023) and assessment-specific principles, we reviewed well-recognized assessment-specific standards (AERA, & APA, & NCME, 2014) and guidelines (including OECD 2023; U.S. Department of Education, Office of Educational Technology, 2023; and ITC-ATP, 2025).

Third, we consulted a cross-disciplinary group of experts from computational psychometrics, language assessment, law, machine learning, and security within Duolingo, and an external RAI expert from computer science[4].

Fourth, after identifying the four ethical principles, the external RAI expert helped to articulate the rationale and overall goal of each standard, and the more detailed subgoals (i.e., practical implementation of each standard).

Finally, the standards were published as a living document and remain open for public comment.

4 We thank Dr. Pascale Fung for her expert guidance in the development of the DET RAI Standards (Burstein, 2025)

Connections to NIST AI RMF

The DET RAI Standards were validated against the independent, national, industry-agnostic NIST AI RMF (2023) *trustworthiness characteristics*. Validation with an independent and industry-agnostic ethical framework demonstrates how our standards are aligned with prevailing best practices.

The NIST AI RMF's trustworthiness characteristics are similar to the DET RAI standards' ethical principles in that both identify characteristics of trustworthy AI. The NIST AI RMF emphasizes the following trustworthiness characteristics: Valid and Reliable; Safe, Secure and Resilient; Accountable and Transparent; Explainable and Interpretable; and, Privacy-Enhanced, Fair–with Harmful Bias Managed. Based on the standards development process, the DET RAI standards focus on standards which echo four of these characteristics in Table 1.

Table 1
NIST AI RMF* trustworthiness characteristics *& DET RAI Standards

NIST AI RMF trustworthiness characteristic	Description[5]	DET RAI Standards	Description
Valid and Reliable	Ensure objective evidence, fulfilling requirements for intended use. Perform consistently in expected conditions over a period of time.	Validity and Reliability	Ensure that the test is suitable for its intended purpose. Validity standards involve evaluating construct relevance and accuracy, while Reliability standards maintain consistent performance over time.
Fairness with harmful bias managed	Address concerns for equality and equity by addressing issues such as harmful bias and discrimination		Promote democratization and social justice through increased access, accommodations, and inclusion represent test-taker demographics, and avoid algorithms known to contain or generate bias
Privacy Enhanced; Secure and Resilient	Adheres to privacy values such as anonymity, confidentiality Maintain confidentiality, integrity, and availability through protection mechanisms, preventing unauthorized access	Privacy and Security	Ensure (a) compliance with relevant laws and regulations governing the collection and use of test taker data; (b) assurance of test taker privacy and (c) assurance of secure test administration.
Accountable & Transparent	Documents information about an AI system and its outputs for individuals interacting with the system	Accountability and Transparency	Provide thorough documentation and explanations

5 See NIST (2023) for complete descriptions.

How the DET RAI Standards Impact Test Quality and Equity

In this section, we illustrate the application of each of the four DET RAI Standards through practices aligned with their goals and subgoals. The examples show how these practices uphold the standards' ideals and support test quality and equity. To do so, we focus on two DET tasks—the *Interactive Reading* and *Writing Sample* tasks. We briefly describe these task types in Section 3.3.1 (also see Naismith et al., 2025 for more details), and discuss how each of the four standards is applied in test development, measurement, and security (Section 3.3.2).

Task Descriptions: Interactive Reading and Writing Sample

The *Interactive Reading* task is a measure of a test-taker's ability to read in academic contexts. The task contains five different item types, targeting different reading sub-constructs. The item types are: Vocabulary in Context; Text Completion; Reading Comprehension; Main Idea; and Possible Title. The item response formats include multiple choice reading comprehension questions, and a question in which test takers highlight a segment of the text to respond. (See Figures 2–6.)

Figure 2

Interactive Reading: Vocabulary in Context (Cloze)

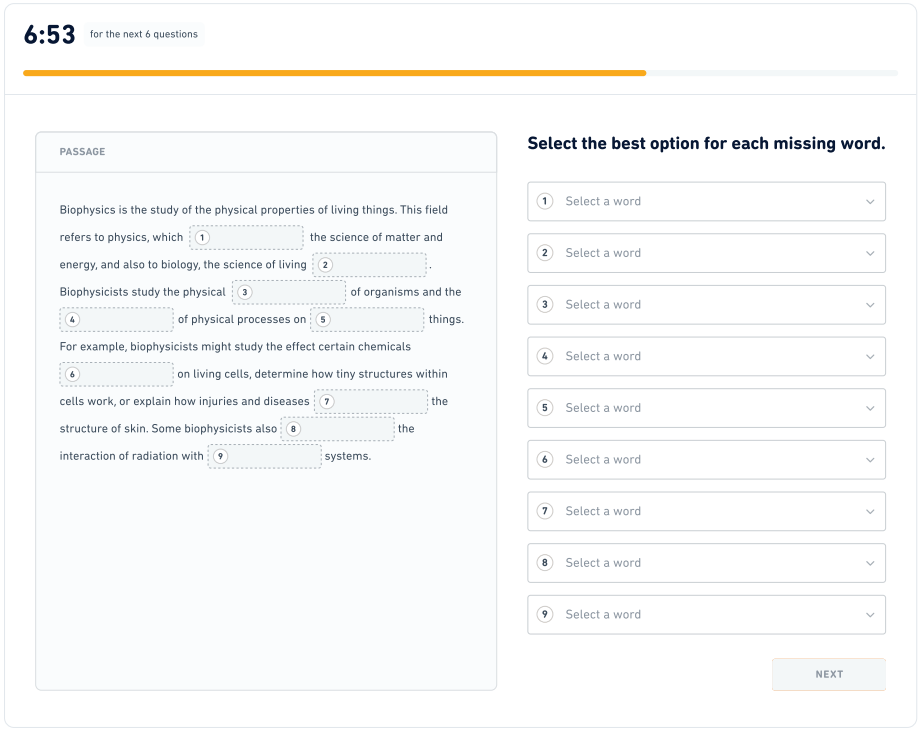


Figure 3

Interactive Reading: Text Completion

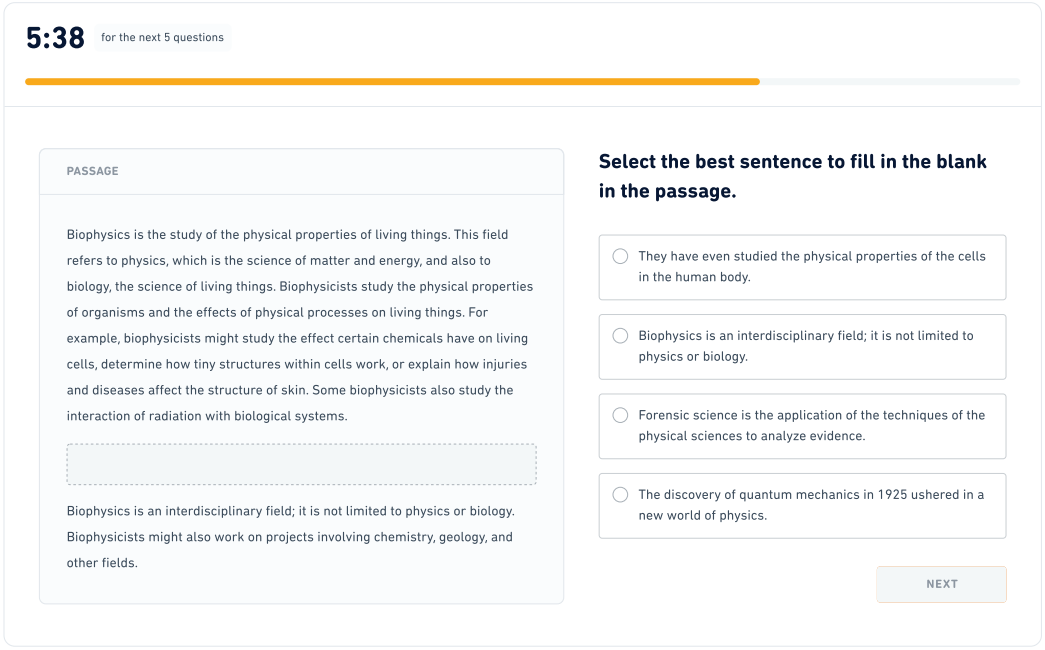


Figure 4

Interactive Reading: Comprehension Questions

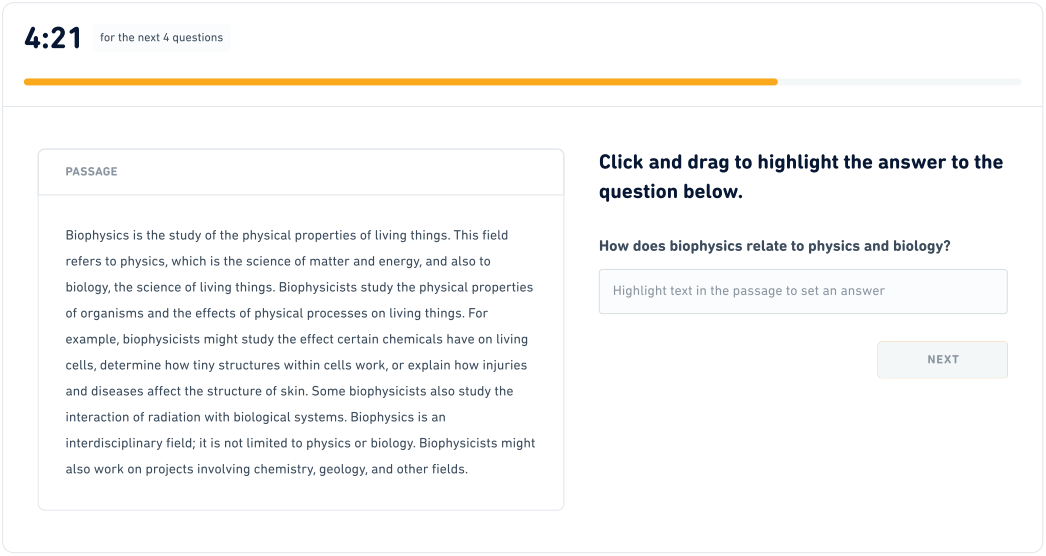


Figure 5

Interactive Reading: Main Idea

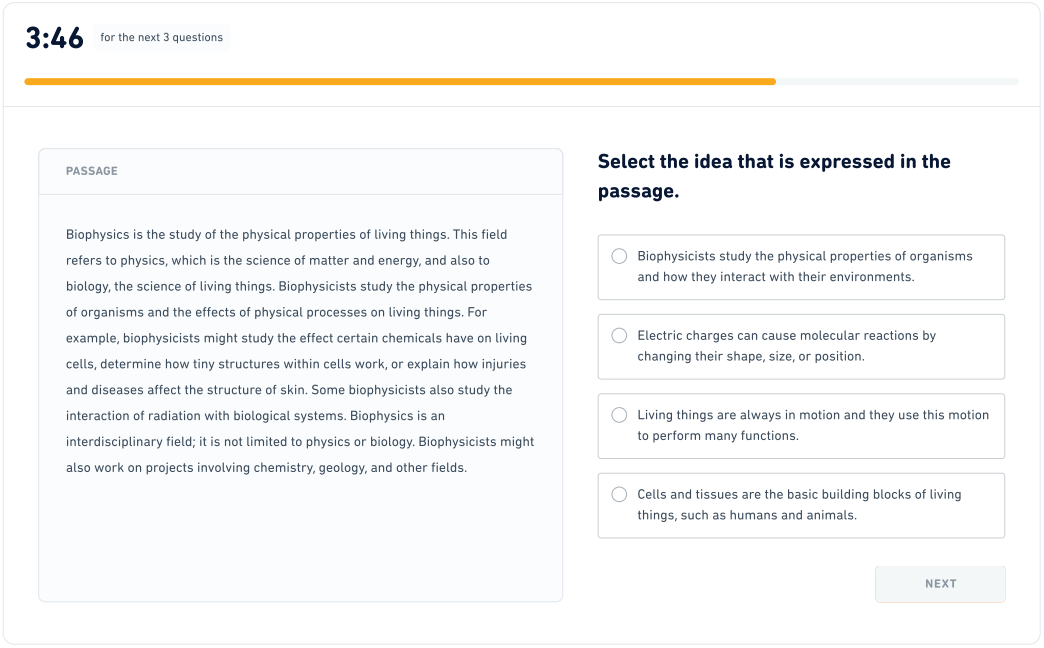
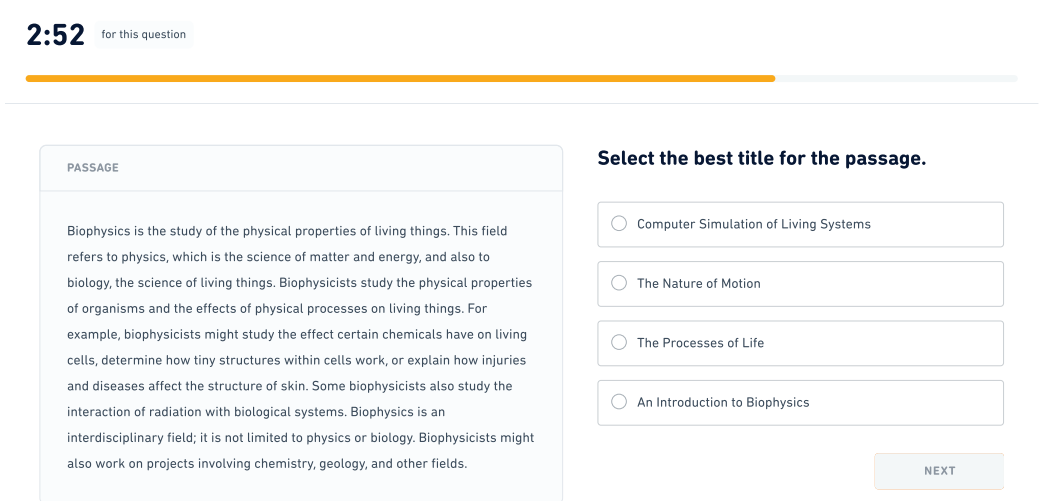


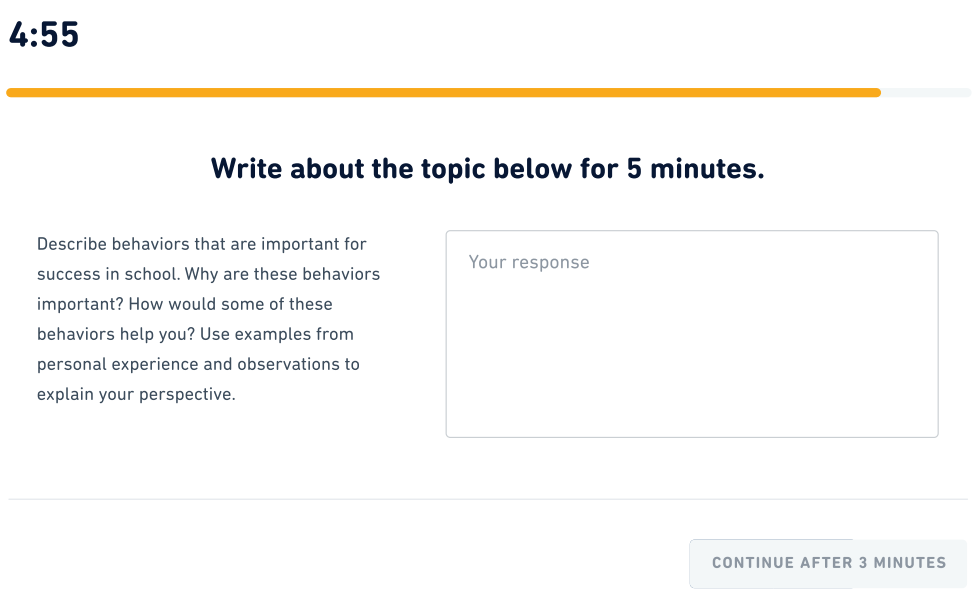
Figure 6

Interactive Reading: Possible Title



The *Writing Sample* is an independent, spontaneous writing task. Test takers receive a prompt, have 30 seconds to prepare, and then have five minutes to write their response. (See Figure 7.)

Figure 7
Writing sample



Applying RAI Standards

Examples in this section illustrate how each of the four RAI standards' practices contribute to test quality and equity, referring to the alignment with relevant validity argument interferences.

We begin with the *Validity and Reliability,* and *Fairness* standards 1 and 2. In this context discuss a **six-step RAI process** for DET *task design*. The six-step process also addresses aspects of measurement (e.g., scoring), and security (e.g., item exposure) issues. Steps 1–6 are referred to throughout this section. Discussions for the *Privacy and Security, and Accountability and Transparency* standards 3 and 4 follow.

Validity and Reliability

The Validity and Reliability Standard focuses on test validity (i.e., suitability for its intended purpose) and reliability (i.e., yields consistent results), and has two main goals. The first goal aims to "specify processes required to build a validity argument", and the second goal aims to "evaluate AI used in test item creation, item calibration, and scoring." We illustrate with **four subgoals** from this standard, demonstrating how they contribute to test quality and equity.

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. (Subgoal 1.1.16)

Rationale. This subgoal aligns with the *domain definition* inference and involves defining the construct. Explicitly defining the construct is necessary for the design of any assessment. With regard to RAI, it is essential for construct fidelity when using AI to automatically generate high-quality text passages.

Implementation. Steps 1 and 2 describe the DET's task design process[7], using the Interactive Reading task to illustrate.

Step 1. *Articulates the target construct, using human subject-matter (assessment science) experts (SME).*

For the Interactive Reading task, the target construct is academic reading. The construct is defined as including a range of reading purposes and cognitive skills categorized into two main areas (Park et al., 2022) important reading skills in university study (Grabe, 2008): Reading for Orientation and *Reading for Information and Argument* (Council of Europe, 2020). Reading for Orientation entails searching for specific information within a text and quickly understanding its general idea with limited information (Giulia Cataldo & Oakhill, 2000; Guthrie, 1988; Guthrie & Kirsch, 1987). Reading for Information and Argument involves understanding main ideas, learning how ideas within a text connect to each other and to the reader's prior knowledge, integrating information from multiple texts or different parts of a long text, and using the carefully curated information to interpret the text or perform other tasks (Grabe & Stoller, 2020; Head & Eisenberg, 2009; Thompson et al., 2013).

Step 2. *Specifies a task and scoring system. This includes AI feature development to operationalize and evaluate the target construct articulated by human SMEs.*

Following the construct definition step, we outline task specifications for passage generation for the Interactive Reading task. These specifications and corresponding scoring systems—including feature development and evaluation—serve to operationalize elements of the target construct, as defined by subject matter experts (SMEs), and support automated passage generation.

The passages are automatically generated to support the Interactive Reading task and fall into two primary text categories: expository and narrative. These categories reflect the target language use domain. Expository texts, such as those found in textbooks (Thompson et al., 2013; Weir et al., 2009) and news articles (Head & Eisenberg, 2009), are particularly relevant for university students. Narrative texts, often used in academic contexts such as ethnographic reports and biographies (de Chazal, 2014), also represent important sources of academic reading.

The automatically generated passages are then used to assess specific aspects of the target construct. *Reading for orientation* is operationalized whereby passages require test takers to demonstrate their comprehension of specific ideas (through text highlighting) and vocabulary knowledge in context (through cloze items). *Reading for information and argument* is operationalized using tasks that require text completion, main idea selection, and passage title identification items.

6 The subgoal nomenclature should be read as follows. The *first numeral* refers to the Standard number, the *second numeral* to the Standard's **Goal number,** and the *third numeral* to the Standard's Subgoal number. For example, **Subgoal 1.1.1** refers to the Validity & Reliability Standard **1,** its Goal **1,** and its Subgoal **1**.

7 The remaining four steps are discussed later under subgoals 1.2.1 and 2.1.3.

Evaluate AI Scoring System Accuracy and Fairness, Leveraging Human Expertise (Subgoal 1.1.2)

Rationale. Aligned with Step 2, it is important to evaluate AI scorers during task development to ensure that expected scoring criteria can be satisfied using computationally-derived features. This is especially applicable to production items, such as open-ended writing and speaking items for which the AI-driven features need to align with human scoring rubrics (e.g., grammar error detection, text coherence) (described below). This is the AI analog to human scoring processes, and maps to the explanation inference.

Implementation. AI-driven feature development is most relevant for production tasks. Therefore, we illustrate the implementation of this subgoal using the *Writing Sample* task, where AWE is used to automatically score test-taker responses.

Human experts develop rubrics with criteria for rating writing that is consistent with the Common European Framework of Reference (CEFR) levels and descriptors (Council of Europe, 2020). The rubrics define criteria based on the quality of four writing sub-constructs: *content* (relevance and task achievement), *discourse coherence* (organization and cohesion), *lexis* (including sophistication and correct use of vocabulary in context), and *grammar* (complexity and accuracy). These rubrics are used to train human raters who annotate sizable samples of *Writing Sample* responses. To do this raters use the rubric criteria to assign scores of 1–6, which are aligned with the six CEFR levels (A1, A2, B1, B2, C1 and C2). Consistent with AERA, APA, and NCME (2014) standards and the *evaluation* inference, agreement rates between human raters are monitored to ensure that ratings are reliable and accurate (measurement), as these will be used to build the automated scoring models. Once the AI scoring models are built using a portion of the full human-rated data sets, the remaining portion is used to compute system-human agreement. This is one of the key quantitative evaluation metrics used for AI scoring model development.

Develop (a) explainable scoring methods, and (b) interpretable AI features used for scoring that have clear alignment with domain constructs (Subgoal 1.1.3)

Rationale. Consistent with the AERA&APA&NCME (2014) standards, this subgoal ensures that AI model scores are valid by requiring that scores are explainable, aligning with the explanation inference. The model features discussed below measure various aspects of the domain construct and support score explanation.

Implementation. The *Writing Sample* task is used to illustrate. The DET uses AI models to score open-ended speaking and writing responses. Drawing from the literature on NLP, linguistics, and AWE, human experts identify features to be included in the AI model. These example features represent the Writing Sample task's four sub-constructs.

- **Content:** Inverse document frequency (IDF) weighted word-similarity is used to measure the relevance of the response to the writing prompt (Burstein et al., 1998; Rei & Cummins, 2016).
- **Discourse coherence:** Sentence overlap, coreference counts, and latent semantic analysis (LSA)-based sentence similarity features (Foltz et al, 1998) similar to those implemented in the widely used Coh-Metrix (McNamara & Graesser, 2012) are used to evaluate lexical cohesion—a measure of coherence. In addition, a fine-tuned LLM trained to predict human ratings of coherence is used as a holistic coherence feature (Naismith et al., 2023).
- **Lexis:** Proportion of words by CEFR level (Xia et al., 2016) and differential word use (DWU) (Attali, 2011) are used as measures of lexical sophistication. DWU uses outputs from n-gram classification models to differentiate low and high proficiency test-takers.
- **Grammar:** Tree depth statistics (Schwarm & Ostendorf, 2005) are used as measures of grammatical complexity. Error rates of various grammatical error types, determined through grammatical error correction and classification (Bryant et al., 2017), are used as measures of grammatical accuracy.

To make scoring models explainable, the DET uses SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), whereby response scores are reduced to a sum of feature contributions, which can be aggregated into sub-construct contributions. Using SHAP supports powerful, explainable scoring models with complex, non-linear relationships (such as XGBoost models; See Chen et al., (2016).

Identify AI methods for item creation, leveraging human expertise to efficiently create valid and reliable test items (Subgoal 1.2.1).

Rationale. Large item banks mitigate the long-standing assessment of security issues associated with item exposure and pre-knowledge (Chen et al., 2003; LaFlair et al., 2022; Way, 1998). This subgoal manages automated item generation to efficiently create large item banks using large language models (LLM), such as GPT-4. It aligns with the *domain definition* inference in that it attends to construct relevance.

Implementation. For illustration, we switch back to the Interactive Reading task type, since it has a more complex item generation process than the Writing Sample item type. Steps 3–4 below discuss the implementation.

Step 3. ***Creates prompts that elicit content and questions from a large language model (LLM) (such as, GPT-4). Prompts are expected to meet the specifications for DET tasks. To generate content and texts at scale for the Interactive Reading task, machine learning and assessment scientists collaborate to develop prompts that align with the task specifications (See Subgoal 1.1.1).***

Step 4. ***Uses LLMs (such as GPT-4) for large-scale generation of questions and content.***

The text types for Interactive Reading (expository and narrative) can be generated at scale via the prompting process (using prompts created in Step 3). One approach to this is to employ in-context learning (Dong et al., 2022), where exemplar texts are provided as part of the prompt submitted to the LLM. In this approach, prototypes of narrative and expository readings are shown to the LLM to generate reading passage outputs. The outputs could be conditioned on any relevant target characteristic, such as a list of STEM subjects or topics.

For *main idea* and *possible title* item types (mentioned earlier), potential **answers** are generated and evaluated based on their similarity to the passage. For *comprehension questions*, the model generates questions and answers, filtering out any content with undesirable characteristics (such as, extreme lengths or poor alignment with the passage). For the *text completion* item type, candidate target sentences are identified based on the probabilistic likelihood of their occurrence in the text. For vocabulary in context items, a different process is used: words for deletion are selected based on likelihood, rank order, syntactic and semantic information, and distance from other elided words. To generate **distractors** for the main idea, possible title, and text completion items, alternative texts, questions, and keys are generated. The keys for the alternate questions are used as distractors for the passage and questions that will be on test. Candidate distractors are selected for human review based on metrics such as vector embedding similarity and LLM log probability. Candidate distractors for vocabulary in context items are then selected from the model's likelihood ranking (targeting lower likelihood values) for the candidate words (Attali et al., 2022). With regard to quality, human review of generated content and items is discussed later as part of Step 5.

Fairness

The Fairness Standard addresses test equity explicitly. It aims to ensure that test takers have equal opportunity to succeed and that AI is free of algorithmic bias. It consists of two main goals. The first goal aims to "specify how the use of AI facilitates test-taker access, accessibility, and inclusion," and the second goal aims to "specify test-taker demographic representation, and algorithms known to contain or generate bias." We focus on two subgoals that address fairness and bias (FAB) review of items, and the mitigation of algorithmic bias.

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 (Subgoal 2.1.5)

Rationale. By focusing on access, accessibility, and inclusion, this standard aims to create a more equitable testing environment for individuals from diverse backgrounds and with varying needs. One of the ways this is achieved is developing and applying item reviews to increase inclusion and eliminate potential biases in automatically-generated test content. This aligns with the explanation inference in that the review manages task (passage) characteristics.

Implementation. Humans review the content and tasks to identify sensitive content and low quality items. The human review process improves the item design process and the prompt development based on human review and feedback. These processes are achieved through Steps 5–6.

Step 5. *Requires human review of content and tasks.*

To mitigate any negative impact of automatically-generated content that is poorly constructed or distracting, a human review process is conducted to remove such content. This includes both an Item Quality Review (IQR) and a Fairness and Bias (FAB) review. Such reviews are a long-standing tradition in test development (AERA, & APA, & NCME, 2014).

IQR evaluates the content and questions to ensure that they represent the relevant text types (narrative and expository) and sub-constructs targeted by the questions. Factual accuracy is also checked. For both types of reviews, reviewers are trained using in-house materials developed in-house by assessment developers. IQRs are tailored to each task type. Where relevant, they are informed by state-of-the-art item writing guidelines (Haladyna et al., 2002). F*AB review* ensures that items do not contain content that may introduce culturally sensitive or inaccessible topics that might upset or distract the test-taker and introduce construct-irrelevant variance. What is different about human review of AI-generated from human-generated content is required awareness about potential issues specifically associated with LLM outputs which are unlikely to occur with human-created items, such as LLM hallucinations.

The F*AB guidelines* are an expanded version of Zieky (2015); they are tailored by the DET test developers through regular discussions to avoid inclusion of sensitive content. Additionally, DET test developers survey test takers to understand what types of content they would like to read when taking the test, allowing for test-taker input during test design. While these test-taker surveys do not create fully democratic involvement in the test design process (Jin, 2023; Shohamy, 2001), they incorporate for test taker input about the test content.

Step 6. *Aims to improve the item design process and the prompt development, using human review and feedback.*

Here, we close the feedback loop between content generation and feedback that is collected from reviewers, weekly research discussions, and test-taker content surveys responses. Information gathered from these sources is used to improve item generation procedures. The information collected through the reviews can be used to improve the prompts for the LLMs, our automated filters of generated content, and our FAB and IQR reviewing guidelines.

This six-step process:
1) aligns the AI-assisted item development process with traditional approaches (e.g., construct articulation, development of specifications, content review) and
2) introduces human evaluation of AI outputs (e.g., content generation and automated scoring) which are likely to have characteristics unique to LLMs.

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 (Subgoal 2.2.1)

Rationale. It is crucial to document and evaluate demographic representation in datasets used for AI-powered assessments. This helps to create inclusive data sets that represent the test-taker population, and mitigate biases in test-taker outcomes, downstream. This is aligned with the evaluation inference as it builds in the awareness of demographic groups.

Implementation. An important part of developing DET's automated writing scorer for the Writing Sample task type is curating human-rated datasets used to train and evaluate scoring models. When building these datasets, Writing Sample responses are selected to include a roughly equal number of test takers identifying as male or female from the seven most common first-language (L1) backgrounds in the test-taker population. L1[8] backgrounds—Arabic, Mandarin Chinese, Telugu, English, Spanish, Gujarati, and Bengali—represent a broad range of language families. This ensures that the model is trained and evaluated on a diverse range of L1 backgrounds, promoting measurement quality.

Evaluate and document bias associated with automatically-generated item content (e.g., Fairness and Bias Review Guidelines), and proficiency measurement (Subgoal 2.2.3)

Rationale. It is essential to evaluate and document known algorithmic bias in AI used in assessment processes, such as test security, design, and measurement. This includes managing potential bias associated with automatically-generated item content and proficiency measurement. This is aligned with the evaluation inference.

Implementation. The DET implements this subgoal for proficiency measurement (scoring) by using Differential Rater Functioning (DRF) analysis (Jin & Eckes, 2021; Myford & Wolfe, 2004) on all scorers for open-ended writing and speaking tasks, including the Writing Sample task. Specifically, these scoring models are evaluated on the representative dataset (mentioned in subgoal 2.2.1) to quantify any bias they may have with respect to sensitive background characteristics (e.g., gender or L1) after controlling response quality. We perform this kind of analysis at both the feature and score level to identify potential differential performance test-takers groups. A similar analysis called differential item functioning (DIF) is used to detect bias at the item level (Holland & Wainer, 2012) caused by automatic item generation. The DET conducts differential item functioning (DIF) on its item bank (Belzak et al., 2023), flags items with potential bias, and sends them back for FAB review.

8 L1 refers to the test-takers' self-reported first language.

Privacy and Security

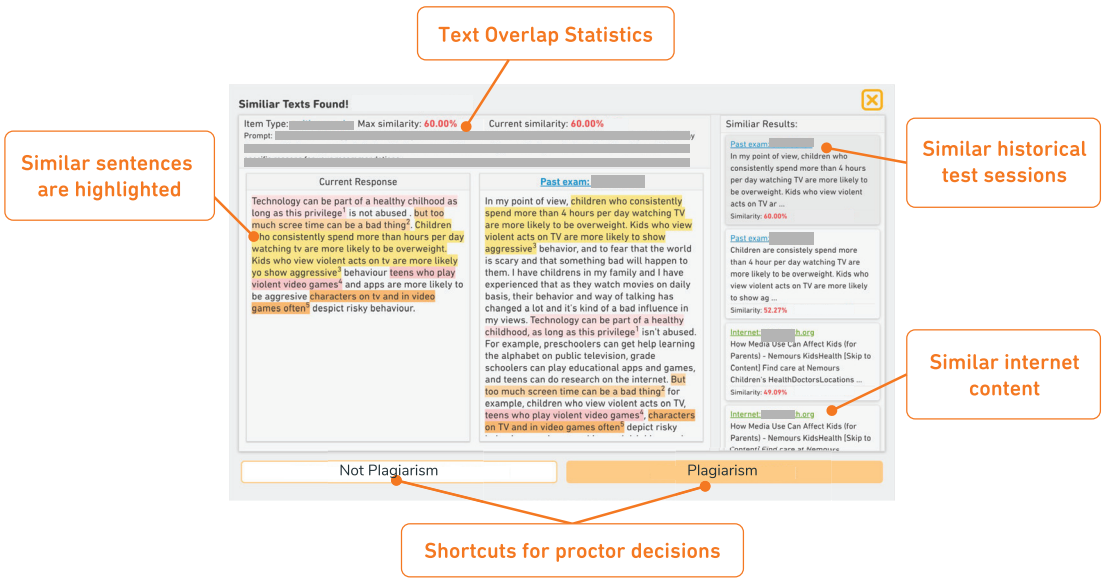
The Privacy and Security Standard seeks to ensure that the test administration process is secure, fair, and reliable, while protecting test-taker privacy and preventing cheating. This standard consists of three *goals*. The first goal aims to "specify methods to ensure privacy and security associated with data origin, data collection and processing, and data management". The *second* goal aims to "specify how to maintain test-taker privacy, item security, and test-taker security during test administration". The *third* goal aims to "specify fair and reliable test security proctoring protocols, item pool development, and psychometric procedures for test security." In this section, we highlight a subgoal from this third goal.

Define, document, and implement human-in-the-loop AI proctoring protocols that fairly and reliably identify novel and known cheating behaviors (Subgoal 3.3.1)

Rationale. This subgoal focuses on ways to use AI to identify (evaluate) cheating behaviors, and develop protocols. It supports DET proctors'[9] use of AI-enabled tools to make informed, equitable decisions about cheating behaviors observed on high-stakes assessments. For instance, evidence associated with test takers hiring other people to help them test, or using texts written by others), and, most recently, using AI tools, such as LLMs (e.g., ChatGPT) to generate responses (Khalil & Er, 2023). This is aligned with the *evaluation* inference.

Implementation. We illustrate the implementation of this subgoal on *traditional plagiarism*. Traditional plagiarism is a known problem on high-stakes language assessments (Wang et al., 2019). For example, test takers memorize long, generic essay responses that they superficially adapt to respond to a writing prompt on an assessment. Such cases can often be detected using AI models that quantify feature overlap between texts (Foltýnek et al., 2020). However, it is important to distinguish between *deceitful* plagiarism and *benign* text overlap (Chandrasoma et al., 2004; Pecorari & Petrić, 2014).

During proctoring, the DET uses AI to compare test-taker writing responses to a database of relevant Internet content and writing responses from historical DET sessions. Matches are flagged and shown to proctors (See Figure 1). The DET's plagiarism tool displays the sources where matches were found, and highlights the overlapping text. This demonstrates human-in-the-loop AI (as defined in Section 3.1), as AI tools help human proctors with decision-making. (Note that subgoals within the Accountability and Transparency standard address *AI literacy* requirements to ensure that proctors understand how AI tools work.)



9 Note that the DET employs asynchronous proctoring (See Duolingo English Test, 2021)

Accountability and Transparency

The Accountability and Transparency Standard seeks to build trust with stakeholders. The standard is satisfied through six goals related to the DET's documentation and dissemination about AI use. (See Burstein, 2025 for details about the six goals). We illustrate standards' application with the first (4.1), second (4.2) and fifth (4.5) goals. The first goal is to "assess how AI processes impact stakeholders". The second goal is related to documenting how "AI is used for building the validity argument, test item creation, test item calibration, and scoring". The fifth goal focuses on "disseminating research about use of AI to various stakeholder communities". Arguably, goals in this standard result in documentation of evidence that supports all inferences in a validity argument, since the documentation offers stakeholders explanation about different aspects AI use for test design, measurement and security.

Document external factors that result in a need to modify AI (Subgoal 4.1.3)

Rationale. External factors can affect the impact of AI use on an assessment. For example, changes in the test-taker population may increase bias as captured by differential item functioning (DIF) or differential rater functioning (DRF).

Implementation. The DET's Analytics for Quality Assurance for Assessment system (AQuAA; Liao et al., 2022) addresses this subgoal. The AQuAA system provides weekly reports on metrics that reflect the quality and comparability of the test scores over time, particularly with respect to shifts in test-taker demographics.

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

Rationale. This documentation ensures that internal stakeholders are fully informed about AI use as they perform their tasks and/or make changes to any part of the test. This ensures that stakeholder actions do not compromise the test's validity, reliability, or fairness.

Implementation. To help satisfy this goal and all its subgoals, the DET documents and controls the use of AI through its Exam Change Proposal (ECP) process. For example, when a new scoring model is developed for task types, such as the Writing Sample, the evidence for the scoring model's validity, reliability, and fairness is collected and documented in an ECP document. Similarly, when a new item is developed for operational use, the evidence that supports the launch of the item is documented in an ECP. Before proposed changes are implemented, the document is reviewed and approved by multiple experts, including DET senior assessment and AI researchers.

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

Rationale. Outcomes from high-stakes assessments can profoundly impact test takers' educational goals. Test developers should clearly communicate with stakeholder communities about how AI is used across the assessment ecosystem for test design, measurement, and security.

Implementation. To reach different stakeholder audiences, DET researchers regularly disseminate research such as through blog posts, white papers, and peer-reviewed articles. For example, the launch of the Interactive Reading task was accompanied by a white paper describing the task and what it measures (Park et al., 2022), and a subsequent technical, peer-reviewed article describing the procedures for automated generation of the task (Attali et al., 2022).

Limitations and Future Work

Organizational RAI guidelines and standards are not a one-time exercise (PwC, 2024). Organizations that build and deploy AI-powered assessments should commit to integrating RAI into the full assessment ecosystem as a test is developed and deployed. Standards development should evolve in tandem with new versions and applications of AI introduced into the test, while also addressing emerging risks that could affect test equity and quality. Known limitations for the DET RAI standards are discussed here.

AI advances. GPT-4o was released (OpenAI, 2024) only slightly ahead of the time this chapter was being written. GPT-4o is a much more powerful LLM than had previously existed. Its multimodal generation capabilities creates opportunities, such as use for innovative item types. Even more advanced models (e.g., GPT 4.5 and GPT 5.0) have been introduced as we finalized the chapter, and model improvements are likely to continue prior to and following the publication of this chapter. At the same time, the advances present additional risks (such as deep fakes which have implications for test security). To maintain test equity and quality as the technology evolves, assessment developers need to consider how this new technology can be responsibly used for assessment.

Fairness. Fairness issues span across all standards. For example, the use of AI to detect traditional plagiarism[10] was described in the Privacy and Security standard. Since it is acknowledged that AI exhibits biases, it is possible that these detectors introduce biases into the plagiarism evaluation. Given the pervasive nature of fairness issues, one approach to consider is making fairness a cross-standard narrative, decreasing the likelihood that fairness issues fall between the cracks.

Additional RAI Standards. The DET RAI includes four standards: Validity and Reliability, Fairness, Privacy and Security, and Accountability and Transparency. These topics were chosen through a process of literature review, consultation with experts, and internal deliberation. However, naturally, some topics were left out. For example, two important issues the DET RAI Standards do not cover are environmental and labor impacts. The DET's environmental impacts include the carbon emissions and water consumption of the generative AI applications involved in the assessment process, as these impacts are notoriously high (Strubell, Ganesh & McCallum, 2023). The DET is working on estimating the environmental impact of the AI models used on the test. Such impact also includes substitution effects associated with test takers taking the DET instead of an alternative English language proficiency test. For example, if each DET test session were instead replaced by a physical, in-person test session at the closest test center, we estimate that it would require approximately tens of millions of additional kilometers in travel each year. Labor impacts could affect Duolingo's employees as a result of the adoption of generative AI. Currently, there have been no negative impacts as the company's employees have been retrained to integrate generative AI assistance.

The DET's approach to these and other important AI ethics issues is incremental, aiming to increase the scope of the DET RAI standards over time.

10 The DET recently introduced methods to detect plagiarism behaviors associated with the use of LLMs, and a manuscript describing the methods is in preparation.

5. Implications & Conclusion

Intended implications of this chapter were to increase AI responsibility in assessment with attention to how it may impact *test quality* and *equity*. To do this, we provided a case study that showcases RAI standards customized for an English language proficiency assessment; explains the standards' development process; validates the standards against an AI industry standard–i.e., the NIST AI RMF trustworthiness characteristics; illustrates the standards' implementation; and, facilitates an opportunity for critical professional and public engagement.

The DET RAI standards illustrate one example of how RAI standards and practices can be developed and applied for assessment. Through concrete examples of standards application, this chapter demonstrates how RAI standards contribute to test quality and equity, and ensure that test score interpretations are trustworthy and appropriate. The broader assessment community is invited to consider the DET RAI standards if they choose to develop standards for other assessments.

Acknowledgements

We are grateful to the anonymous reviewers for their insightful comments. Many thanks to our Duolingo English Test colleagues: Mancy Liao, for providing content for plagiarism detection; and, Ed Fu, for content related to environmental impact.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational & psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Association of Test Publishers. (2024). *Creating responsible and ethical AI policies for assessment organizations* (July 12, 2024).

Attali, Y. (2011). *Differential word use for content assessment. Journal of Educational Measurement, 48*(1), 1–22. https://doi.org/10.1111/j.1745-3984.2010.00126.x

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.903077

Auernhammer, J. (2020). Human-Centered AI: The role of human-centered design research in the development of AI. In S. Boess, M. Cheung, & R. Cain (Eds.), *Synergy—DRS international conference 2020* (pp. 1315–1333). https://doi.org/10.21606/drs.2020.282

Belzak, W. C., Baig, B., Naismith, R., Hastings, R., Horie, A. K., LaFlair, G., Liao, M., Niu, C., Shih, Y. S. (2025). *Duolingo English Test: Security and score integrity. DRR-25–01.* Duolingo English Test. https://duolingo-papers.s3.us-east-1.amazonaws.com/reports/DET_Security_Report.pdf

Belzak, W. C., Naismith, B., & Burstein, J. (2023, June). Ensuring fairness of human-and AI-generated test items. In *International Conference on Artificial Intelligence in Education* (pp. 701–707). Springer Nature Switzerland.

Burstein, J. (2025). *The Duolingo English Test Responsible AI Standards* (Duolingo Research Report DRR-25–05, Version 3). Duolingo. https://go.duolingo.com/ResponsibleAI

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *36th annual meeting of the Association for Computational Linguistics and 17th international conference on Computational Linguistics: Vol. 1* (pp. 206–210). Association for Computational Linguistics. http://dx.doi.org/10.3115/980845.980879

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test* [Research report]. Duolingo English Test. https://duolingo-papers.s3.amazonaws.com/other/det-assessment-ecosystem-mpr.pdf

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics: Vol. 1. Long papers* (pp. 793–805). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P17-1074

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge. https://doi.org/10.4324/9780203854693

Chandrasoma, R., Thompson, C., & Pennycook, A. (2004). Beyond plagiarism: Transgressive and nontransgressive intertextuality [Publisher: Taylor & Francis]. *Journal of Language, Identity, and Education, 3*(3), 171–193.

Chen, S., Ankenmann, R. D., & Spray, J. A. (2003). Exploring the relationship between item exposure rate and item overlap rate in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129–145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x.

Chen, T., & Guestrin, C. (2016, August). *Xgboost: A scalable tree boosting system.* In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794).

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment–companion volume*. Council of Europe Publishing. www.coe.int/lang-cefr

de Chazal, E. (2014). *English for academic purposes*. Oxford University Press.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., & Sui, Z. (2022). *A survey for in-context learning.* arXiv preprint arXiv:2301.00234.

Duolingo English Test. (2021). *Duolingo English Test: Security, proctoring, and accommodations.*[White Paper]. https://duolingo-papers.s3.amazonaws.com/other/det-security-proctoring-whitepaper-2021-11.pdf

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., & Jiang, L. (2023). *LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models.* The Twelfth International Conference on Learning Representations, https://openreview.net/forum?id=aIok3ZD9to

Fiesler, Casey and Garrett, Natalie. (16 Sept 2020). *Ethical Tech Starts with Addressing Ethical Debt, Wired Ideas:* https://www.wired.com/story/opinion-ethical-tech-starts-with-addressing-ethical-debt/

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI* [Research report]. Berkman Klein Center for Internet & Society at Harvard University. https://dx.doi.org/10.2139/ssrn.3518482

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In S. Carta (Ed.), *Machine learning and the city: Applications in architecture and urban design* (pp. 535–545). John Wiley & Sons Ltd. https://doi.org/10.1002/9781119815075.ch45

Foltýnek, T., Bjelobaba, S., Glendinning, I., Khan, Z. R., Santos, R., Pavletic, P., & Kravjar, J. (2023). ENAI recommendations on the ethical use of Artificial Intelligence in education. *International Journal for Educational Integrity, 19*(1).

Foltýnek, T., Dlabolová, D., Anohina-Naumeca, A., Razı, S., Kravjar, J., Kamzola, L., Guerrero-Dib, J., Çelik, Ö., & Weber-Wulff, D. (2020). *Testing of support tools for plagiarism detection. International Journal of Educational Technology in Higher Education, 17*(1), 46. https://doi.org/10.1186/s41239-020-00192-4

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2). http://imej.wfu.edu/articles/1999/2/04/index.asp

Giulia Cataldo, M., & Oakhill, J. (2000). Why are poor comprehenders inefficient searchers? An investigation into the effects of text representation and spatial memory on the ability to locate information in text. *Journal of Educational Psychology, 92*(4), 791–799. https://doi.org/10.1037/0022-0663.92.4.791

Grabe, W. (2008). *Reading in a second language: Moving from theory to practice.* Cambridge University Press. https://doi.org/10.1017/CBO9781139150484

Grabe, W., & Stoller, F. L. (2020). *Teaching and researching reading* (3rd ed.). Routledge.

Guthrie, J. T. (1988). Locating information in documents: Examination of a cognitive model. *Reading Research Quarterly, 23*(2), 178. https://doi.org/10.2307/747801

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology, 79*(3), 220–227. https://doi.org/10.1037/0022-0663.79.3.220

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Head, A. J., & Eisenberg, M. B. (2009). *Lessons learned: How college students seek information in the digital age* [Progress report]. University of Washington, The Information School. http://www.ssrn.com/abstract=2281478

Holland, P. W., & Wainer, H. (2012). Differential item functioning. Routledge. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (Version 2). IEEE. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

The International Privacy Subcommittee of the ATP Security Committee. (2021, July 6). *Artificial intelligence and the testing industry: A primer.* Association of Test Publishers.

International Test Commission & Association of Test Publishers (2025). *Guidelines for technology-based assessment.* Association of Test Publishers; International Test Commission.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

Jin, Y. (2023). Test-taker insights for language assessment policies and practices. *Language Testing, 40*(1), 193–203. https://doi.org/10.1177/02655322221117136

Jin, K.-Y., & Eckes, T. (2021). Detecting differential rater functioning in severity and centrality: The dual DRF facets model. *Educational and Psychological Measurement 82*(4), 757–781. https://doi.org/10.1177/00131644211043207

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, *59*(3), 338–361.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://psycnet.apa.org/doi/10.1037/0033-2909.112.3.527

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73. https://doi.org/10.1111/jedm.12000

Khalil, M., & Er, E. (2023, June). Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection. In *International Conference on Human-Computer Interaction* (pp. 475–487). Springer Nature Switzerland.

Khan, S., Hamer, J., & Almeida, T. (2021). Generate: A NLG system for educational content creation. In I.-H. Hsiao, S. Sahebi, F. Bouchet, J. -J.Vie (Eds.), *Proceedings of the 14th international conference on educational data mining* (pp. 736–740). Educational Data Mining.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Studies in language testing: Vol. 9. Fairness and validation in language assessment: Selected papers from the 19th Language Testing colloquium, Orlando, Florida* (pp. 1–14). Cambridge University Press.

Liao, M., Attali, Y., Lockwood, J. R., & von Davier, A. A. (2022). Maintaining and monitoring quality of a continuously administered digital assessment. *Frontiers in Education, 7.* https://doi.org/10.3389/feduc.2022.857496

Lottridge, S., Godek, B., Jafari, A., & Patel, M. (2021). *Comparing the robustness of deep learning and classical automated scoring approaches to gaming strategies* [Technical report]. Cambium Assessment Inc.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems: Vol. 30.* Curran Associates, Inc.

McNamara, D. S., & A. C. Graesser. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing." *Applied natural language processing: Identification, investigation and resolution* (pp. 188–205). IGI Global.

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, *5*. https://doi.org/10.1016/j.caeai.2023.100152

Myford, C., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.

Naismith, B., Cardwell, R., LaFlair, G., Nydick, S., & Kostromitina, M. (2025). *Duolingo English Test: Technical manual* (Duolingo Research Report). Duolingo. https://go.duolingo.com/dettechnicalmanual

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. *In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA 2023, pp. 394–403). Toronto, Canada: Association for Computational Linguistics.

National Institute of Standards and Technology (2023). *Artificial intelligence risk management framework (AI RMF 1.0).* U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

OpenAI (2024). https://openai.com/index/hello-gpt-4o/

OpenAI (2023). *GPT-4 Technical Report*. https://arxiv.org/pdf/2303.08774.pdf

OECD (2023). *Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI* (No. 349). OECD Publishing. https://doi.org/10.1787/2448f04b-en

Park, Y., LaFlair, G. T., Attali, Y., Runge, A., & Goodwin, S. (2022). *Interactive reading—The Duolingo English Test* [White paper]. Duolingo English Test. https://doi.org/10.46999/RCXB1889

Pecorari, D., & Petrić, B. (2014). Plagiarism in second-language writing. *Language Teaching, 47*(3), 269–302. https://doi.org/10.1017/S0261444814000056

PriceWaterhouseCoopers (PwC) (2024). *PwC's 2024 US Responsible AI Survey.* https://www.pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-survey.html

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 1–18.

Rei, M., & Cummins, R. (2016). Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 283–288). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W16-0533

Schwarm, S., & Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL'05), pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics. https://doi.org/10.17705/1thci.00131

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation.* Routledge.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language testing, 18*(4), 373–391. https://journals.sagepub.com/doi/abs/10.1177/026553220101800404

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). *Energy and policy considerations for deep learning in NLP.* In Proceedings of the Annual Meeting of the Association for Computational Linguistics, (pp. 3645–3650).

Thompson, C., Morton, J., & Storch, N. (2013). Where from, who, why and how? A study of the use of sources by first year L2 university students. *Journal of English for Academic Purposes, 12*(2), 99–109. https://doi.org/10.1016/j.jeap.2012.11.004

U.S. Department of Education, Office of Educational Technology (2023). *Artificial intelligence and future of teaching and learning: Insights and recommendations* [Report]. https://www2.ed.gov/documents/ai-report/ai-report.pdf

Von Davier, A., & Burstein, J. (2024). *AI in the Assessment Ecosystem: Implications for Fairness, Bias, and Equity.* In Artificial Intelligence in Education: The Intersection of Technology and Pedagogy, Springer Nature Switzerland.

Wang, G. (2019, October 20). H*umans in the Loop: The Design of Interactive AI Systems:* https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems

Wang, X., Evanini, K., Mulholland, M., Qian, Y., & Bruno, J. V. (2019). *Application of an Automatic Plagiarism Detection System in a Large-scale Assessment of English Speaking Proficiency.* Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 435–443.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17–27. https://doi.org/10.1111/j.1745-3992.1998.tb00632.x

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). *The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university* (vol. 9; pp. 97–156). British Council; IELTS Australia.

Xia, M., Kochmar, E., & Briscoe, T. (2016). Text readability assessment for second language learners. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 12–22). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W16-0502

Zechner, K., & Evanini, K. (Eds.). (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Zieky, M. J. (2015). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81–99). Routledge.

About the authors

Jill Burstein is Principal Assessment Scientist at Duolingo, leading validity and efficacy research for the Duolingo English Test – Duolingo's English language proficiency test. Her career has been motivated by social impact, working on AI-driven, education technology to enhance equity and access for learners and test takers. Her research lies at the intersection of artificial intelligence and natural language processing, educational measurement, equity in education, learning analytics, and linguistics. Dr. Burstein pioneered the first automated writing evaluation system used in large-scale, high-stakes assessment, as well as early commercial online writing instruction tools. She holds numerous patents for this work, and has published extensively in the field of AI in education, including topics in automated writing evaluation, digital assessment, responsible AI, and writing analytics. Her recent work focuses on responsible AI for digital assessment, and wrote the Duolingo English Test Responsible AI Standards, the first standards for an assessment program. Additionally, she is a co-founder of SIG EDU, an ACL Special Interest Group on Building Educational Applications. Dr. Burstein holds a Ph.D. in Linguistics from the Graduate Center, City University of New York.

Geoffrey T. LaFlair is a Principal Assessment Scientist at Duolingo where he co-leads Assessment Research and Development for the Duolingo English Test. He holds an MA in TESOL from Central Michigan University and a Ph.D. in Applied Linguistics from Northern Arizona University. 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 ESL at the University of Kentucky. His research interests are situated at the intersection of language assessment, psychometrics, and natural language processing, focusing on the application of research from these fields in researching and developing operational language assessments.

Kevin Yancey is a Senior Staff AI Researcher at Duolingo, leading the engineering and AI functions for Research & Development on the Duolingo English Test.As an expert software engineer and AI researcher who has also taught and studied abroad in two foreign countries, he is passionate about the applications of technology to second language learning and assessment.His work in AI specializes in the field of Natural Language Processing (NLP), where he has made innovative contributions to automatic readability estimation, automatic writing evaluation, and estimating item response theory (IRT) item parameters for L2 assessments using explanatory models with NLP features.

Alina A. von Davier is a researcher, innovator, and an executive leader with over 20 years of experience in EdTech and in the assessment industries. She is the Chief of Assessment at Duolingo, leading the Duolingo English Test research and development area. She is the Founder and CEO of EdAstra Tech. She is an American Educational Research Association (AERA) Fellow and serves as an Honorary Research Fellow at University of Oxford, and a Senior Research Fellow Carnegie Mellon University. Her research spans computational psychometrics, machine learning, and education. Dr. von Davier's work has been widely recognized in the academic community. She received the Brad Hanson award twice from National Council on Measurement in Education (NCME) for her pioneering work on computational psychometrics, and her work on adaptive testing. She received ATP's Career Award for her contributions to assessment. She was a finalist for the Innovator award from the EdTech Digest. The AERA awarded her the Division D Signification Contribution Educational Measurement and Research Methodology Award for her publications "Computerized Multistage Testing: Theory and Applications" (2014) and an edited volume on test equating, "Statistical Models for Test Equating, Scaling, and Linking" (2011).

Ravit Dotan, Ph.D., is a renowned tech ethicist specializing in artificial intelligence (AI) and data technologies. She aids tech companies, investors, and procurement teams in developing and implementing responsible AI strategies, conducts research on these topics and creates resources. Dr. Dotan was recognized as one of the 100 Brilliant Women in AI Ethics for 2023 and has received accolades such as the 2022 "Distinguished Paper" Award from the FAccT conference. Her views are frequently featured in prominent publications like the *New York Times, The Financial Times,* AP News, and TechCrunch. Dr. Dotan holds a Ph.D. in Philosophy from UC Berkeley and has extensive experience in AI ethics research, teaching, and advocacy for diversity and inclusion in academia. You can find Dr. Dotan's resources on her AI Ethics Treasure Chest and LinkedIn page.

About the Study Group

The Study Group exists to advance the best of artificial intelligence, assessment, and data practice, technology, and policy; uncover future design needs and opportunities for educational systems; and generate recommendations to better meet the needs of students, families, and educators.

Date of Publication
March 2026

Case Study Citation: Burstein, J., LaFlair, G. T., Yancey, K., von Davier, A. A., & Dotan, R. (2026). *Responsible artificial intelligence for test equity and quality: The Duolingo English Test as a case study.* The Study Group.

Chapter Citation: Burstein, J., LaFlair, G. T., Yancey, K., von Davier, A. A., & Dotan, R. (2025). Responsible artificial intelligence for test equity and quality: The Duolingo English Test as a case study. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume I: Foundations for assessment in the service of learning.* University of Massachusetts Amherst Libraries.