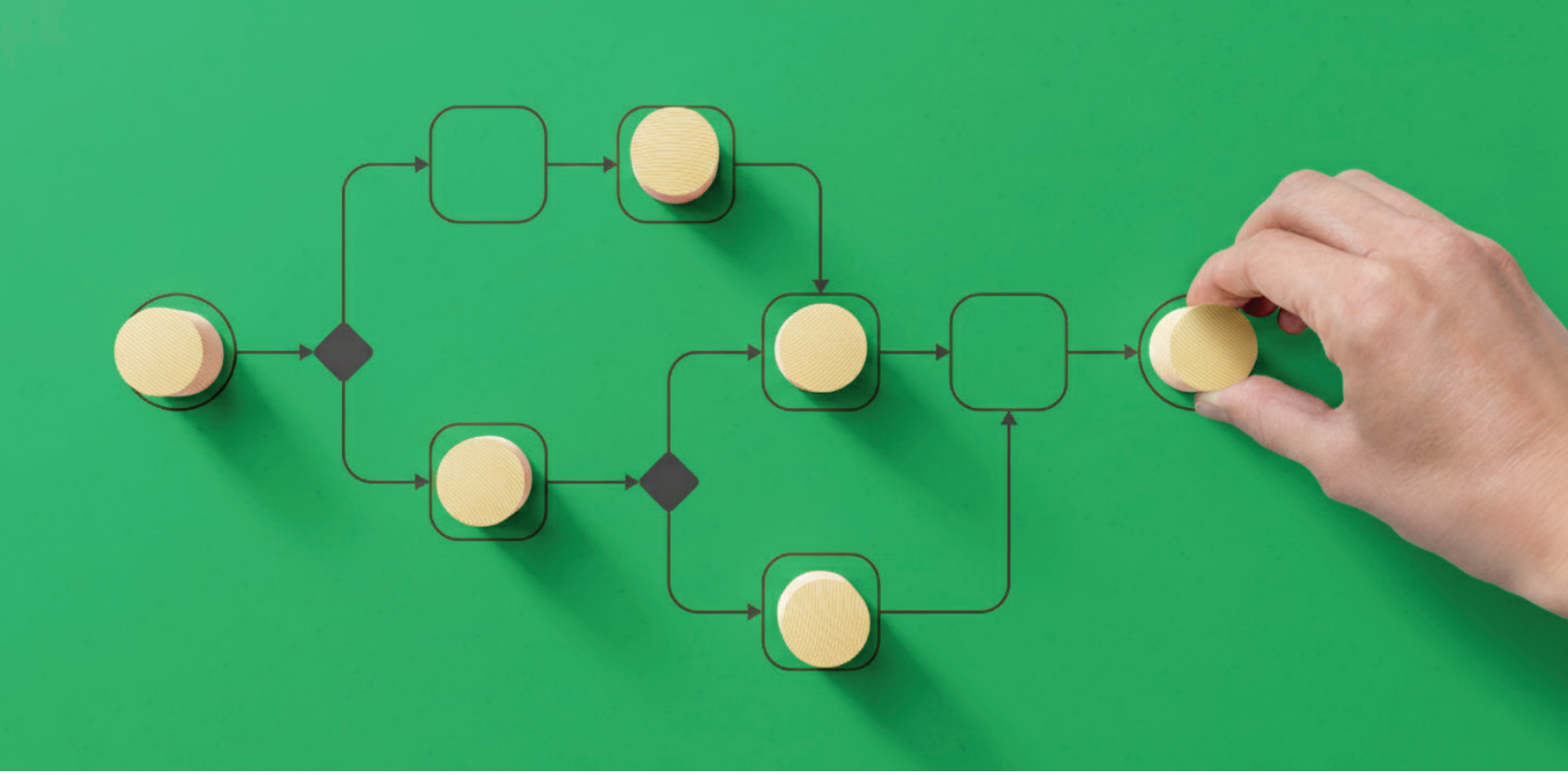


ARTIFICIAL INTELLIGENCE IN LARGE-SCALE ASSESSMENT PROGRAMS:

Applications and Considerations for State Education Agencies

February 2026

Will Lorié & Nathan Dadey

Center for Assessment

National Center for the Improvement
of Educational Assessment, Inc.
Dover, New Hampshire

ACKNOWLEDGMENTS:

We acknowledge the support and feedback from André Rupp, Sue Lottridge, Laura Hamilton, Ruhan Circi, Juan D'Brot, and other colleagues whose ideas and suggestions improved this paper. Special thanks to Catherine Gewertz, whose keen insights and thorough edits significantly improved the clarity of our work. Any errors and omissions are our own.

SUGGESTED CITATION:

Lorié, W., & Dadey, N. (2026). Artificial intelligence in large-scale assessment programs: Applications and considerations for state education agencies. Dover, NH: The National Center for the Improvement of Educational Assessment.

ARTIFICIAL INTELLIGENCE IN LARGE-SCALE ASSESSMENT PROGRAMS: *Applications and Considerations for State Education Agencies*

TABLE OF CONTENTS

ARTIFICIAL INTELLIGENCE IN LARGE-SCALE ASSESSMENT PROGRAMS:

Applications and Considerations for State Education Agencies

EXECUTIVE SUMMARY

This paper explores the potential of artificial intelligence (AI), especially generative AI (GenAI),[1] in educational assessment, with a focus on large-scale assessment (LSA) programs—specifically statewide summative assessments designed to fulfill federal accountability requirements.

We've structured the paper around the LSA life cycle, analyzing how AI currently influences, or could influence, each phase, from defining what we're measuring to reporting, and how it could support processes that cross stages. For each phase of the life cycle, we outline the main activities involved and then examine current and potential AI applications for that phase. We also offer considerations for state education agencies (SEAs) aiming to use AI in their programs at each stage. These phases are briefly described in the remainder of the Executive Summary.

Construct Definition: Defining what we're measuring

AI can assist in developing assessment frameworks and specifications, enable new task types that incorporate AI agents, and support emerging constructs such as AI literacy. These innovations expand how knowledge and skills can be represented and measured, but also raise questions about validity, fairness, and standardization. Ensuring that the primary responsibility for defining and validating constructs remains in the hands of human experts is essential to maintaining the legitimacy and interpretive coherence of test scores.

Content Development: Creating tasks to measure knowledge and skills

GenAI systems can produce assessment items, synthesize passages and stimuli, and support human authors by providing feedback on item quality and alignment. These applications promise efficiency gains but also introduce challenges related to factual accuracy, bias, intellectual property, and content security. Effective implementation requires securely hosted models, rigorous human expert review, and reengineering of development workflows to preserve quality and defensibility.

[1] In this paper, we use "GenAI" when we are specifically referring to technologies that generate text, images, etc. We use "AI" when referring to AI more generally, which may include GenAI.

Field Testing and Equating: Building and maintaining tests

AI can model item characteristics, predict item difficulty, and identify potential sources of bias, reducing some of the burden of field testing and equating. Current research suggests that AI can augment—but not replace—empirical data from students, as prediction accuracy and generalizability remain limited. Validation studies will be needed before AI methods in field testing and equating can be relied upon operationally.

Administration: Delivering assessments to test-takers

AI can support more efficient and secure test delivery by assisting with scheduling, logistics, proctoring, monitoring, and accessibility tools. Emerging applications include adaptive interfaces and real-time analytics that help maintain smooth administration across varied settings. Because these systems rely on sensitive behavioral and technical data, they raise concerns about privacy, fairness, and transparency. As in other stages, human oversight remains essential to protect these.

Scoring: Rating responses to open-ended test items

Large language models now supplement traditional automated scoring systems in replicating human ratings for essays, short answers, and spoken responses. They also offer new possibilities for modeling rater disagreement and refining rubrics through interactive human-AI collaboration. For operational use, however, scoring systems must be deterministic, validated against human ratings, and deployed within secure, locally controlled environments.

Reporting: Connecting test results to intended interpretations and uses

AI enables more personalized and interpretable score reporting, ranging from elaborated performance-level descriptions to pattern-based or individualized narratives. These advances can make reports more meaningful but also complicate validation and comparability. To ensure that automatically generated interpretations remain defensible, reporting systems must document their evidentiary basis and adhere to principles of explainable AI.

Other Applications: Cross-cutting support for testing programs

Beyond the primary life cycle stages, AI is being applied to translation, alignment, standard setting, and program documentation. Early studies show that AI can streamline multilingual translation, assist with alignment judgments, and support standard-setting analyses. Used responsibly, these cross-cutting applications can improve efficiency and coherence across the assessment system while maintaining transparency and human oversight.

INTRODUCTION

Educational assessment is increasingly influenced by developments in artificial intelligence (AI), especially generative AI (GenAI).[2] This paper focuses on the intersection of AI and statewide summative assessments designed to meet federal accountability requirements. It complements summary work on applications and considerations of AI in assessment, such as that of Hao et al. (2024) and Bulut et al. (2024), and related guidance, such as that published by Duolingo (Burstein, 2025) and ETS (Johnson, 2025). Our paper builds on that work by foregrounding the life-cycle stages of large-scale assessment (LSA).

Although the technical and operational aspects of LSAs are implemented primarily at the vendor level, the use of AI is likely to be negotiated through requests for proposals and change requests that are best evaluated when state education agency (SEA) staff understand the possibilities and limitations of the technology. Accordingly, our primary audience is state education agency (SEA) staff interested in how AI can be leveraged to improve the quality of their statewide summative assessment programs, increase efficiency, or decrease costs, while also addressing risks. More broadly, this paper serves as a primer on AI, supporting SEA staff in considering whether and how to incorporate AI into their programs.

In this section we first introduce the key stages of the LSA life cycle and provide context on AI and GenAI in education. We then describe how AI is poised to impact the way LSA programs are designed and implemented. In the next section we examine each of the LSA life-cycle stages in detail: construct definition, content development, field testing and equating, administration, scoring, and reporting. For each stage we examine how AI currently impacts, or could impact, that stage and we highlight considerations for AI in that stage. We follow these stage-specific sections with a section on processes that are less routine (like standard setting) or that cut across the stages (such as documentation). We then conclude with an overview and overall implications for SEAs.

THE LARGE-SCALE ASSESSMENT LIFECYCLE

Large-scale assessment (LSA) refers to continuously operating testing programs that provide information for public reporting, school accountability, placement and exit decisions, or the improvement of educational programs or policies. LSAs, such as statewide summative testing programs, follow a structured life cycle with multiple interconnected stages. Although the implementation varies from program to program, all programs go through these stages. These stages are shown in Figure 1. In addition, although the life cycle is presented as linear, iterations between stages throughout development are typical.

Figure 1. The large-scale assessment life cycle

[2] In this paper, we use GenAI when we are specifically referring to technologies that generate text, images, etc. We use AI when referring to AI more generally, which may include GenAI.

These stages are:

1. **Construct Definition:** Determining what to measure and the approach to measuring it
2. **Content Development:** The authoring, curating, and review of test materials
3. **Field Testing & Equating:** Trying out items with representative samples and linking scales
4. **Administration:** Delivering assessments to test-takers
5. **Scoring:** Evaluating student responses, particularly for open-ended items
6. **Reporting:** Communicating results and their interpretations to intended audiences

Supporting these stages are other activities that are often implemented in differing ways across programs, and thus, their connections to each stage can vary. These activities are shown at the bottom of Figure 1. For example, standard-setting has typically been conducted after the first administration using panels of experts. However, some programs have shifted towards designs that involve standard setting activities included in the construct-definition and content-development stages (for example, embedded standard setting, Lewis & Cook, 2020).

In this paper, we do not directly address the various roles and responsibilities for designing, managing, and updating LSA programs or implementing AI applications. These programs are typically designed through a combination of state agency and vendor efforts and implemented primarily at the vendor level, as noted above. This is relevant to emerging AI applications, which will either be incorporated into LSA programs by design (for example, built into RFPs) or negotiated through change requests.

AI AND GENERATIVE AI

Artificial intelligence (AI) refers to "computer systems capable of performing tasks that historically required human intelligence, such as recognizing speech, making decisions, and identifying patterns" (Coursera Staff, 2024). This paper addresses AI generally but gives special attention to *generative AI* (GenAI), which involves computer systems creating *new* content, such as text, images, or music, based on patterns learned from existing data. The underlying technology of GenAI is part of a broader field of machine learning, with methods often considered under the umbrella of AI. These methods have been used in educational assessment for several years before the introduction of GenAI, particularly in automated scoring.

GenAI came to public attention with the release of ChatGPT by OpenAI on November 30, 2022 (OpenAI, 2022). GenAI builds upon a body of research on *language models*—systems that learn the patterns and relationships among words, allowing them to encode meaning and generate new text. Such language models had already been used in large-scale assessment applications (primarily for scoring open-ended responses) before the release of ChatGPT (Yan, Rupp, & Foltz, 2021). *Large language models* (LLMs) take this approach further: They are trained on massive datasets (at the scale of the Internet) and include billions of adjustable settings, called parameters, that allow them to produce highly fluent and context-aware outputs (Devlin, Chang, Lee, & Toutanova, 2019; Hao et al., 2024). The computational resources necessary to develop LLMs, often referred to as *pre-training* the model, are substantial, meaning that, as of the time of writing, only large companies or organizations can afford to create them (Hao et al., 2024).

Early versions of ChatGPT were LLMs that built on, processed, and generated text data. Other models, such as OpenAI's DALL-E (OpenAI, n.d.) and Midjourney, Inc.'s Midjourney (Midjourney, n.d.), are built on labeled image data; they process image and text data to generate images. Newer models are increasingly *multimodal*, meaning they can process and generate multiple forms of data, not just text.

LLMs vary in size (number of parameters). Larger models deliver higher-quality outputs; however, the pre-training method and the specific data used can significantly impact the results of an LLM. LLMs are "black boxes": There is very little technical information about how each one is built (including what specific data they use), and, in general, it is not feasible to explain their outputs. (In the context of LSA programs, this lack of transparency, which we'll address, raises questions about validity.)

Most current applications of LLMs in LSA programs adapt existing general-purpose models rather than train new ones from scratch. Adaptation approaches range from lightweight to highly technical. At the simplest level, practitioners use *prompt engineering* or *prompt chaining* to guide a hosted model such as ChatGPT, Gemini (Google DeepMind, n.d.), or Claude (Anthropic, n.d.) toward a desired behavior. Because these proprietary models are accessed through servers not under the test vendor's or the SEA's control, their use can raise data-governance and privacy concerns—issues discussed further in the "considerations" sections of this paper.

Given the sensitivity of assessment data, many emerging AI applications in LSA contexts rely instead on open-source, locally hosted models (e.g., LLaMA 2 [Meta AI, n.d.]; Mistral 7B [Mistral AI, n.d.]). Importantly, these models can be adapted through various techniques called *fine-tuning*, which updates only a small subset of parameters to specialize the model for particular tasks, such as generating test items or classifying student responses. Another important tool in adapting LLMs is *retrieval augmented generation*, or RAG, in which in which the model is connected to a curated database or document repository that it can "consult" during generation to retrieve relevant, up-to-date information rather than relying solely on what it learned during training. Although these approaches require more technical expertise than prompt engineering, they produce models that are both better aligned with specific use cases and deployable within secure, institution-controlled computing environments.

WHY AI IN LARGE-SCALE ASSESSMENT?

Before turning to applications of AI and GenAI in large-scale assessment, it is helpful to clarify why AI belongs in the LSA life cycle at all. The long histories of work on automated scoring and computer-assisted item development provides examples, demonstrating gains in scalability, consistency, and turnaround time (Yan, Rupp, & Foltz, 2021; Yaneva & von Davier, 2023). At a minimum, AI similarly offers greater efficiency and cost reductions across high-volume tasks. With GenAI, the case expands, as it represents a qualitative shift in capability. For example, because LSAs depend on sustained content creation (items, prompts, passages, rubrics, and feedback), GenAI aligns directly with a core need of LSA systems.

Beyond efficiency, the rationale for AI in LSA is strategic. As AI becomes embedded in schooling, work, and everyday life, fluency with AI tools will likely be reflected in state standards and key frameworks (e.g., ISTE Standards, see Sykora, 2024). This trajectory, in turn, will encourage innovative item formats that incorporate AI-mediated tasks and interactions, a theme we develop further in the section on construct definition.

In each LSA life-cycle section, we examine how AI is used or could be used at that stage of the life cycle. We then highlight considerations for incorporating AI into that stage.

There is considerable overlap in the features of AI across the life-cycle stages. Moreover, key considerations apply across several stages, including maintaining security, ensuring unbiasedness, and attending to the explainability of machine-generated output. We'll address these in the relevant sections of the LSA stages. Here, we elevate one frequently referenced consideration that cuts across all stages: The need to keep humans "in the loop" of processes involving AI.

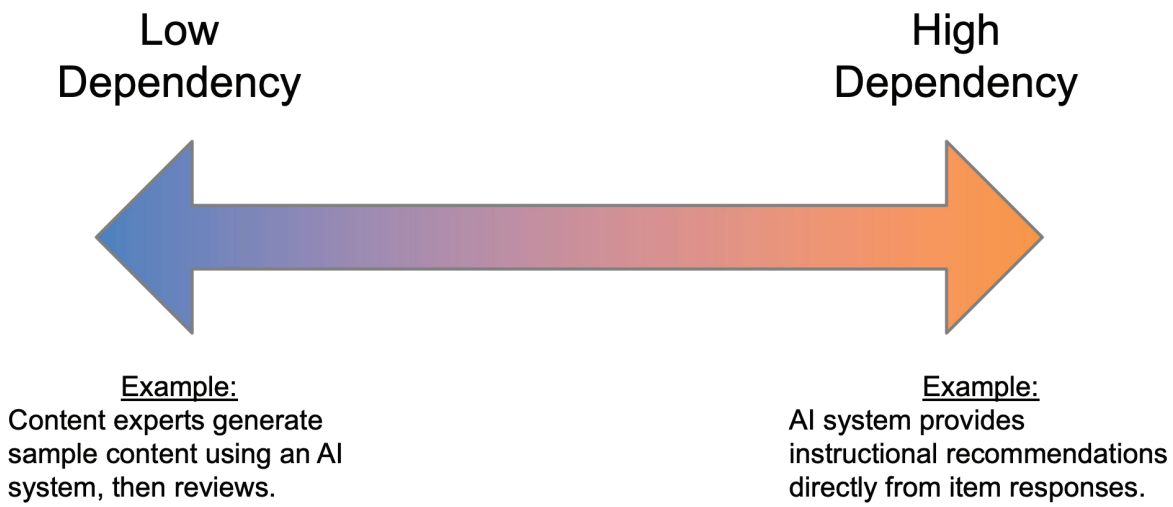
HUMANS IN THE LOOP

A consistent theme throughout the literature is the need for humans to be involved in virtually every application of GenAI, i.e., that humans should be in the loop. This idea of human-in-the-loop (HITL, Stanford Institute for Human-Centered AI, 2021), is more than just including humans in a process; it is about human ownership and oversight. The core idea of HITL is that in automated systems, a human should be responsible for overseeing the system and reviewing its functioning and output. A key reason why HITL has become important in GenAI is not only the widespread availability of GenAI but the fact that these systems often produce output that is irrelevant, incorrect, misleading, or unsafe. For LSA programs, failure to attend to HITL threatens the validity, fairness, and credibility of programs.

> *The more dependent humans are on the AI system, the greater the potential threat to validity.*

Having humans in the loop is critical to every LSA life-cycle stage, but in different ways, depending on characteristics of (1) the context and (2) the human user. Together, these two provide a heuristic for examining the level of dependency between the human and the AI system. The more dependent humans are on the AI system, the greater the potential threat to validity. Restated, the less human oversight there is of an AI system, the greater the potential threat to validity.

Figure 2. Human-In-The-Loop Relationship Continuum



An example of low human dependency is a situation in which content experts are responsible for developing and reviewing assessment items, and use a custom-built AI system to generate sample content. In this case, the experts' knowledge and their ability to accept, reject, or modify system suggestions create a low-risk HITL relationship. This is the ideal HITL relationship in LSA systems—one in which humans have the expertise to critically evaluate GenAI output and the flexibility, if the GenAI does not meet quality standards, to continue engaging productively in that specific LSA stage. (The content developers can always consult other sources for material.)

At the other end of the HITL relationship spectrum, the human relies on AI-generated outcomes and has limited knowledge or insight to assess the results. Consider a scenario, in the reporting stage of the LSA life cycle, where an AI system directly produces instructional recommendations to teachers based on student assessments. By "directly," we mean that the recommendation comes from the students' test responses without any model in between to interpret those responses. In this case, teachers' expertise in turning assessment data into instructional strategies wouldn't be enough to evaluate the AI recommendations critically. This is because the users—the teachers—would lack understanding of how the AI system processed student information to generate its suggestions. If these recommendations were part of a broader educational-technology program where all students are enrolled, it would increase the teacher's reliance on AI in this type of HITL relationship.

A key consideration for AI applications throughout all stages of the LSA life cycle is to ensure that HITL relationships have low dependency, as shown in the first example. When a specific AI application risks creating a high-dependency HITL relationship, that application should be changed to lower that dependency. Some strategies for rebalancing the HITL relationship in the second example include ensuring (1) that the AI assistant always bases recommendations on a well-founded assessment model, (2) that teachers can ask about the basis of the recommendations, and (3) that teachers can easily provide instructional interventions outside of the edtech program.

Finally, this HITL relationship should be viewed expansively and can be applied broadly. Consider a third example: automated scoring. Here there are multiple expert humans-in-the-loop and they engage in a well-structured process. This well-structured process involves experts in scoring

selecting representative student work at each score point (i.e., range-finding[3]) and scoring a sufficient number of student responses. Then measurement experts apply GenAI, likely in conjunction with other machine learning methods, to automatically score student work. Quality assurance and control checks are implemented throughout this process. This process involving HITL decreases the dependency on GenAI. Likewise, other processes could be implemented to decrease the dependency on GenA in other areas.

AI APPLICATIONS AND CONSIDERATIONS

CONSTRUCT DEFINITION

What are we measuring? How do we approach measuring it? These are the central questions of construct definition, which describes what we hope to capture when we test, such as how well a student understands and applies what is specified in a state's academic content standards for English language arts.

The line between construct definition and content development (the next phase) can get blurry. For our purposes here, we consider construct definition to include not just the literal definition of what we are assessing, but also high-level considerations on how to operationalize that definition, such as the specification of subdomains, item formats, expected alignment with standards, measurement conditions, target population, allowable accommodations and supports, and purpose and use statements. Thus, we include the development of test *frameworks*[4] and *content specifications*[5] in construct definition. Often, LSA programs develop a fairly limited set of content specifications, most notability a test blueprint. Given this, AI has the potential to greatly enhance and build out the construct definition stage.

AI also has the potential to reshape construct definition in large-scale assessment by influencing both what is assessed and decisions about how it is assessed (for example, when to assess, under what conditions, and using which item type(s)). These possibilities introduce new dimensions to familiar constructs and raise questions about standardization, fairness, and the role of AI fluency. In turn, they call for careful attention to whether and how the use of AI alters the intended meaning of test scores.

Applications

GenAI can influence the construct-definition stage of large-scale assessment in several ways. First, it can support framework development and specification design. Second, it can enable new types of test items, including interactive tasks that embed AI agents (such as chatbots). Third, GenAI introduces the possibility of directly assessing AI-related knowledge and skills. Finally, even when AI

[3] Range-finding is the process that identifies student responses (such as essays) that represent the different score categories (such as rubric score points).

[4] A test framework is a high-level document that describes what a test measures, its subdomains, the intended interpretation of its results, and its purpose and use.

[5] Content specifications translate the ideas of a test framework into guidelines for content developers. Content specifications specify item types, targeted standards, blueprints, scoring guidelines, and accessibility features.

knowledge is not itself part of the construct, AI fluency may become a prerequisite for interacting with assessment tasks, as test designs increasingly incorporate AI-based tools or supports.

Developing test frameworks and content specifications

The National Assessment of Educational Progress's recent framework for the 2028 science assessment (National Assessment Governing Board, 2023) is a prime example of a test framework, a high-level document describing a construct to be assessed.[6] Typically, framework documents draw on many sources, such as other frameworks, state academic content standards, standards from professional organizations, and widely adopted curricula.

GenAI systems can, in theory, draw upon all these sources to assist framework developers in producing framework documents and content specifications. LLMs' capacity to process language allows them to synthesize across sources and develop draft frameworks according to the instructions (prompts) provided by framework developers. Ideally, this would entail principled approaches to developing frameworks and content specifications, for example, by prompting AI systems to identify and prioritize common elements across the academic content standards of various states. GenAI might assist with optimizing blueprints or producing item specifications that capture the intent of state standards while incorporating best practices from other state testing programs. This untested application requires new software and low-dependency HITL relationships; expert humans would have to take seriously their role in reviewing such summaries and have recourse to alternative approaches.

AI-assisted test framework development will likely expand as assessment programs seek more efficient, coherent, and scalable ways to link test frameworks to item production.

Research in this area to date has been limited. Owen (n.d.) argues for developing a new specification framework for language testing that integrates AI. AI-assisted test framework development will likely expand as assessment programs seek more efficient, coherent, and scalable ways to link test frameworks to item production. Although we are not aware of any large-scale testing program currently incorporating AI in these ways, they represent a potential direct application of AI to the process of defining a construct.

Designing new kinds of test items

AI is poised to help large-scale test developers design better, more sophisticated items by ushering a change in how we assess constructs, including those incorporating generative interfaces (Rupp & Lorié, 2023). By incorporating GenAI agents (such as a chatbot) into tasks, test developers can assess how well a student can complete a performance task, given access to the AI resource. Such a task might assess higher-order thinking skills like understanding when there are insufficient resources to solve a problem and knowing what questions to ask to narrow the problem. For example, in a science assessment, students might need to determine which of several water treatment methods most effectively removes pollutants from runoff. By querying the chatbot for existing research on the methods under consideration, they would be able to rule out certain factors, and can design a more efficient experiment in a subsequent part of the task.

[6] Although NAEP framework documents do not include test blueprints and item specifications, we use "test framework" to encompass these more specific content-specification components of testing programs.

This extends to difficult-to-assess constructs. One example is individual assessment of constructs that involve interactions with others (such as collaborative problem-solving). AI is beginning to be used in these contexts. For example, Runge et al. (2024) report on using LLMs to generate dialogs for items assessing interactional competence. Goodwin et al. (2024) report on how GenAI can be used in a two-stage writing task to generate (ahead of testing) and to assign (dynamically, in response to a student's first stage response) follow-up prompts for an interactive writing task on the Duolingo English Test. The ongoing success of these efforts furnishes proof of concept that LSAs at the state level could take up.

Directly assessing AI-related knowledge and skills

The ability to use the internet for various tasks has become an expectation in nearly all college work and careers; fluency with AI is now moving in that same direction. Prominent education organizations have begun to articulate what this fluency entails. For example, Digital Promise (Lee et al., 2024) defines AI literacy as the capacity to understand, evaluate, and use emerging technologies responsibly, while the European Commission and OECD (2025) propose an international AI literacy framework that emphasizes understanding how AI works, creating and managing AI systems, and recognizing their social and ethical implications.

This growing emphasis on AI literacy is likely to influence what is tested and how. Hao et al. (2024) call on the measurement community to "recognize and adapt to the changing landscape of labor and technology" (p. 19). They outline three broad categories of skills in this evolving landscape: (1) using AI-powered systems effectively, (2) evaluating the outputs of those systems, and (3) recognizing bias and potential harms in their applications. As with digital literacy, the emergence of AI literacy will create demand for assessments that directly measure such competencies.

In practice, direct assessment of AI-related knowledge and skills could take several forms. Scenario-based tasks might present students with an AI-generated summary or image and ask them to identify possible errors, bias, or misuse of data. Data-reasoning items could ask students to explain how the choice of training data affects the fairness of an algorithm. Design tasks might require students to craft effective prompts for an AI system or to outline steps for validating AI-produced results against authoritative sources. Ethical-decision items could present short case studies and ask students to evaluate the risks or recommend responsible uses.

At present, no large-scale U.S. state assessment directly measures AI knowledge and skills. However, the increasing need to prepare students for an AI-enabled world will likely accelerate the integration of AI literacy into academic standards and curricula—a direction recently endorsed at the federal level (Trump, 2025). Once such expectations are codified in standards, they will become part of the constructs that large-scale assessments aim to measure.

Requiring AI fluency to respond to assessment tasks

Even when AI-related knowledge and skills are not directly included among the targets of assessment, the growing presence of AI in classrooms may nonetheless lead to the expectation that students possess a minimal level of fluency in AI to respond to certain novel items.

Returning to the previous example regarding the task with the built-in GenAI agent, students need to be skilled at interacting with chatbots to query them successfully and efficiently. Another example is allowing students to use a specially designed AI tool when writing essays. Here, students would need a level of fluency using AI as a writing aid to engage in such a task successfully.

Considerations

In the construct-definition stage, AI has implications not only for assessment constructs but also for the processes used to define and operationalize them. When GenAI systems assist in synthesizing frameworks and specifications, designing new kinds of test items, or directly assessing AI literacy, the core concern for SEAs remains the same: ensuring that constructs are valid, bounded, and defensible.

Framework and specification design

When GenAI systems are used to support the development of test frameworks and content specifications, developers must ensure that the resulting constructs remain traceable to human decisions and established sources such as academic content standards, curricula, and professional guidelines. Because LLMs synthesize patterns across diverse sources, they may introduce or omit emphases that subtly alter the intended construct. To preserve construct fidelity, prompts and outputs should be version-controlled, with documentation linking every AI-generated element back to its human-reviewed antecedent. In this context, the *NIST AI Risk Management Framework* (NIST, 2023) offers a useful guide: its emphasis on "trustworthy AI" that is valid, transparent, and accountable mirrors the evidentiary expectations in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Validity of AI-enabled tasks

Embedding AI components within test tasks, such as conversational agents that respond to examinee queries, introduces validity risks. These systems must provide accurate, construct-relevant information, remain within the scope of the intended domain, and preserve equivalent conditions for all test takers. Generative systems can produce fluent but incorrect information, or hallucinations (Ji et al., 2023). Inaccurate or inconsistent responses can compromise fairness if some students receive misleading or advantaged assistance. Likewise, the probabilistic nature of many GenAI models can challenge the standardization requirements of large-scale assessments, where equivalent opportunities and conditions are essential. In practice, this means that within-test AI systems must be constrained, set to exhibit deterministic behavior, rigorously tested for factual accuracy, and subject to human oversight during pilot phases.

Two other related but distinct risks arise when adapting general-purpose AI models to assessment contexts. *Domain drift* occurs when a model draws on knowledge or reasoning beyond the intended construct. *Gaming*, by contrast, occurs when test takers exploit the model's conversational or inferential capabilities to obtain unintended information or assistance. Both threaten score validity in different ways: the former compromises construct representation, while the latter affects score meaning and comparability. Consistent with the *NIST framework* (2023) categories of safety, security, and validity, mitigation should include (1) limiting model access to domain-specific knowledge bases, (2) disabling conversational memory or open-ended prompting, (3) logging and auditing all examinee-AI interactions, and (4) conducting "red-team" validation—that is, testing how AI-savvy users might intentionally probe for weaknesses or hints. These practices parallel established test-security protocols but extend them to AI behavior.

CONTENT DEVELOPMENT

In educational assessment, content development involves creating all the materials and supporting technology that students, test administrators, and other users interact with during the assessment. First and foremost, this involves the development of test items, which can range from traditional selected-response formats to open-response items to performance tasks. Stimuli, such as text passages, images, scripts, or prompts for interactive tasks are part of content development as well. Materials that teachers or raters use to make inferences about student skills, such as scoring guides, are also included.

Along with human scoring and field testing, content development is one of the costliest parts of the assessment development process. Recent estimates of per-item costs for the National Assessment of Educational Progress (NAEP) range from $1,000 to over $20,000 depending on item type, with a per-item average of $3,700 for item creation and review. (National Academies of Sciences, Engineering, and Medicine [NASEM], 2022). Consequently, there is great interest in whether GenAI can directly generate large volumes of high-quality items or enhance the efficiency of the development process. Recent research syntheses consistently cite cost, scale, and timeline as key motivations for AI-assisted item generation (e.g., Tan, Armoush, Mazzullo, Bulut, & Gierl, 2025; Sommer & Arendasy, 2025; Song, Du & Zheng, 2025).

GenAI has the potential to support initial content development, either by generating content directly (to be later reviewed by experts) or as a tool to assist item writers. It also has the potential to facilitate the review and revision of content. In all these applications, human oversight remains critical. Multiple recent research synthesis (Artsi et al., 2024; Tan et al., 2025, Sommer & Arendasy, 2025; Song et al., 2025) all conclude that human-in-the-loop oversight is needed in content development. Introducing GenAI in the content development stage raises several key considerations, including those related to intellectual property (IP), technical infrastructure, accuracy, bias, security, and efficiency. We recommend, at least initially, that SEAs seeking to use GenAI in this stage of the LSA life cycle deploy GenAI primarily as a tool used by human experts, closely attend to efficiency and IP risk, and utilize securely hosted GenAI systems.

Applications

Initial content development

By now, many educators have logged onto GenAI-based chatbots like ChatGPT, Gemini or Claude and asked them to create a test item or even an entire assessment. With some additional prompting and revision, many have generated something that could be edited and presented to students. Naturally, the question then is, "Can we do this for large-scale tests, and do it at scale?"

The answer, in short, is a *qualified* yes. Although GenAI's capacity to generate assessment content is increasingly well documented, the quality of that content remains uneven and highly dependent on human review and revision. Looking across a wide body of studies, a number of recent research syntheses (Artsi et al., 2024; Tan et al., 2025, Sommer & Arendasy, 2025; Song et al., 2025) have found that GenAI has been successfully used to conduct automatic item generation (AIG) or automatic question generation (AQG), generally with humans-in-the-loop at various stages of the

generation process. In this literature, the items are predominantly selected response type items and are generally developed for higher education and professional certification. There is relatively little literature exploring AI item generation in K-12 contexts (e.g., Song et al., 2025), and similarly, there is relatively little literature on the use of AI to create other content, like images, scenarios or passages, outside and separate from the creation of items or questions. Also, in the studies in which content was evaluated for quality, which is not every study, the majority of GenAI content generally passed initial content review by subject-matter experts and displayed generally acceptable psychometric properties.

However, these studies also show that not *every* AI-generated item, passage or other kind of content is of high quality. Work that has examined quality, as summarized by Sommer & Arendasy (2025) and partially by Song et al. (2025), shows that both human review and field testing are needed, meaning that AI alone cannot be trusted to produce high-quality content. Human review helps ensure that the content quality is high in terms of the substantive, content domain, whereas field testing helps ensure the psychometric quality of the content. We explore these two aspects of quality—content-based quality and psychometric quality—below.

> *Although GenAI's capacity to generate assessment content is increasingly well documented, the quality of that content remains uneven and highly dependent on human review and revision.*

The number of GenAI-generated items that pass initial content review or content validation varies widely in the literature, depending on the development process, domain assessed, and review criteria. In the more moderate range of acceptance, Attali et al. (2022) retained about 58% of reading passages and associated item sets (454 out of 789 passages) after expert review for the DuoLingo English Test. Similarly, Shultz et al. (2025) found that 38 of 60 (63%) generated pharmacotherapy items met content validity requirements. On the higher end, Runge et al. (2024) found that 725 out of 900 listening comprehension tasks (81%) for the Duolingo English Test successfully passed expert review. Studies with approaches similar to Runge et al. in complexity have likewise found high acceptance rates (e.g., Bhandari et al., 2024; Kaya et al., 2025; Law et al., 2025). Again, the variability in acceptance rates is likely due to the intersections of the approach to prompting, the robustness of the development approach, and the complexity of the content domain. Chan et al. (2024), for example, found that advanced prompting (e.g. chain-of-thought) produces higher acceptance rates than more simple prompts.

In addition to acceptance rates based on content-expert review, there is evidence that the content developed by GenAI can be *qualitatively different* from that produced by human item writers. For example, when Feng et al. (2024) applied five different approaches using LLMs (ChatGPT, GPT-4, and Mistral) to generate distractors for multiple-choice questions in middle school mathematics, they found that human-generated distractors were judged significantly more relevant to the question stem and were more likely to be selected by actual students. Likewise, Chauhan et al. (2025) found that the distractors for GenAI tend to be flawed at a higher rate than human-authored items. More generally, a review of studies by Artsi et al. (2024) on LLM-based medical question generation noted that the only study comparing items produced by LLMs to those authored by humans demonstrated

that the latter were more relevant (as determined by expert ratings). Finally, Sommer and Arendasy (2025) further caution that even when GenAI items appear to be appropriate, the GenAI items may induce subtle shifts in construct representation, cognitive demand, or elicited response processes.

The results for psychometric quality are similar, but more limited. Song et al. (2025) notes that there are relatively few studies that examine item performance using student response data, and that even fewer employ formal psychometric models. Studies that do evaluate the psychometric quality of items have found that, like content quality, psychometric quality is dependent on approach to prompting, the robustness of the development approach, and the complexity of the content domain. Several studies have found reasonable distributions of item difficulties (e.g., Cheung et al., 2023; Bhandari et al., 2024; Kıyak et al., 2025), but have also found that AI-generated items often tend to be on the easier side of the item difficulty distribution (e.g., Attali et al. 2022; Baudin, 2025; Kaya et al., 2025; Kıyak et al., 2025; Law et al., 2025; Shultz et al., 2025) and often have significant numbers of items with discrimination values falling below common conventions (e.g., item total correlations < .30; e.g., Attali et al., 2022; Baudin, 2025; Chauhan et al., 2025; Law et al., 2025; Shultz et al., 2025), although, less frequently, discrimination was similar or higher than human-authored items (e.g., Bhandari et al., 2024; Kaya et al., 2025). Substantively, findings on item difficulty suggest that GenAI may struggle to produce complex items (e.g., a higher level of rigor or cognitive complexity; e.g., Baudin, 2025; Law et al., 2025). The findings on discrimination may be related to the easiness of the items (e.g., Runge et al., 2024) or that GenAI has been found to struggle in developing plausible distractors (e.g., Attali et al., 2022).

Finally, looking across the literature, there are, as of yet, no best practices or specific consensus on the generation of items or their evaluation, beyond the need for human-in-the-loop development and evaluation. There are high-quality practices to be sure, but it remains to be seen whether any specific approach emerges as best practice.

... there are, as of yet, no best practices or specific consensus on the generation of items or their evaluation ...

Review and revision of items

The consensus in the literature is that content created using GenAI must be thoroughly reviewed by content developers before publication, especially in high-stakes settings (e.g., Tan et al., 2025; Sommer & Arendasy, 2025). GenAI can sometimes produce novel content that appears plausible but is incorrect (i.e., hallucinate). Diagnosing these errors can be problematic, as it may require deep content expertise. This shows that while GenAI can assist and may expedite content development, "humans [should] remain in charge, using AI as a tool to augment decision-making rather than replace it." (Stanford Institute for Human-Centered AI, 2021). This approach is not very different from many everyday uses of AI systems, where human users prompt the system but must evaluate the output.

These reviews by content developers often lead to revisions or rejections, ensuring that the content approved is worth field testing. In high-stakes contexts, none of the items used in field testing should be incorrect, and certainly not for operational use. Naturally, there is interest in whether AI can assist with reviewing test items or questions.

To organize emerging approaches, Gorgun and Bulut (2024) propose three methods for evaluating automatically generated items: human evaluation, metric-based evaluation, and post-hoc analysis.

The first of these is the most widely used. It is standard practice for assessment programs, regardless of the use of GenAI, and the safest (minimizing the risk of a flawed item being published), but also the least efficient. Post-hoc analysis occupies the opposite end of the risk spectrum. To assess item quality, items are given to students and the responses are submitted to psychometric analysis. (In the section on field testing, we address how AI can be used to assess item quality ahead of administration by predicting the results of post-hoc analysis, thus providing a kind of item quality review.) In the middle is metric-based evaluation, where automatically generated items are compared to a reference set.

The literature confirms that LLMs may be helpful in item review through metric-based evaluation, but AIs cannot be relied upon to identify all flawed items. For example, Gorgun and Bulut (2024) report that an open-source model (Llama 3-8B) trained on instructional materials misclassified 18% of bad reading comprehension items as good. Similarly, Bedi et al. (2024) found that an ensemble of LLM models misclassified 17% of medical licensure items deemed flawed by clinical reviewers. Thus, as with item generation applications of AI, item review applications of AI also require human oversight; therefore, if the context calls for a very small tolerance for error (as is the case for LSAs), the need for human review of items cannot currently be eliminated, even for complex workflows that include AI assistance in review.

S. Lottridge (personal communication, June 16, 2025) suggested that initial item development and review/revision of items can be considered together as *AI supports for item developers*. Lottridge envisions human authors choosing how to utilize AI for initial item development and then receiving feedback on those items. The feedback would be provided by a system that leverages LLMs trained on an LSA program's data. Such a system, she explains, can inform human authors about the similarity of draft items to those already in the bank, how well aligned the item is to the intended content standard, the grade level of the vocabulary, the predicted difficulty of the item, potential biases, etc. The test developer can then make revisions as needed based on that feedback.

Considerations

The differences between a classroom teacher using a freely available GenAI tool to help develop tests for their students and a test vendor doing the same for an LSA program emerge quickly when we consider that statewide summative assessment content must be secure, be produced at scale, pass review by subject-matter experts, and ultimately function well from a measurement perspective. (The quality of classroom assessments also matters, of course: even if they do not face the same security and psychometric demands, low-quality tests can influence course grades, affect instructional decisions, and ultimately carry negative consequences for students.)

Implementing GenAI-assisted content authoring for LSAs requires addressing several issues, including accuracy, bias, intellectual property, technical infrastructure, security, and efficiency.

Accuracy and bias

GenAI introduces new pathways for bias propagation ...

As noted in the section on construct definition, GenAI systems are prone to factual inaccuracies and reasoning errors, often referred to as *hallucinations*, where the model produces fluent but incorrect or fabricated information (Ji et al., 2023). In the context of large-scale assessment, such inaccuracies can compromise the validity of test items by introducing content that misrepresents standards, misaligns with item specifications, or includes distractors that

are implausible or ambiguous. Ensuring accuracy in GenAI-assisted item authoring requires rigorous human review, prompt and output validation, and systematic comparison against content specifications and style guides.

Bias presents an equally critical challenge. LLMs reflect patterns present in their training data, which may encode social, cultural, or linguistic biases, which can have negative effects in education applications (Weissburg et al., 2024; Zhao, Singh, & Li, 2024). When used in test development, these biases can manifest as stereotypes and uneven language complexity across demographic groups that disadvantage specific populations. While bias in human-authored items is also a longstanding concern, GenAI introduces new pathways for bias propagation, especially when prompts or data fed to systems (through retrieval augmented generation) are not representative of diverse student populations. However, recent work on Hawai'i's Kaiapuni Assessment of Educational Outcomes (KĀ'EO) program has shown that AI can also be intentionally used to address and improve cultural and linguistic integrity of items (Kūkea-Shultz & Brockmann, 2025).

Ultimately, both accuracy and bias must be addressed in AI based test development workflows, ensuring that AI-assisted content adheres to the same validity and fairness standards as human-authored items.

Intellectual property

There are at least two concerns involving intellectual property. First, that many LLMs have been trained on, and therefore may reproduce, copyrighted material. Second, that what is produced by LLMs cannot be copyrighted without sufficient human revision and authorship.

> *… what is produced by LLMs cannot be copyrighted without sufficient human revision and authorship.*

In terms of the first, many proprietary GenAI chatbots, such as ChatGPT, Claude, and Gemini, have been trained on datasets that include copyrighted material, raising questions about the ownership of their outputs (Chuks-Okeke, Linero, & Leong, 2024). Even LLMs that are open-source may contain copyrighted material as well. Moreover, LLMs can sometimes reproduce segments of their training data verbatim without attribution, which can lead to potential instances of plagiarism (Mittal, 2024). The risks these issues pose are not yet clear, but the fact that these issues might not even be detectable could by itself represent a potential liability.

In terms of the second issue, the copyright of generated material, the U.S. Copyright Office recently clarified that AI-generated content lacking sufficient human authorship is not eligible for copyright protection (U.S. Copyright Office, 2025). This means that AI-generated content must be "sufficiently" modified by humans to retain copyright, which is important as copyright is the primary means by which action can be taken when assessment content is inappropriately exposed (e.g., using a copyright claim to get test content removed from a social media site). However, it may be that copyright is less of a concern if assessment content can be developed at a greater volume at reduced cost.

To attend to intellectual property, SEAs may want to establish policies and procurement requirements that ensure AI-assisted content development complies with intellectual property law

and ethical standards. SEAs can ask for clear disclosure about the sources and licensing status of any models used, and about the safeguards vendors have in place to prevent the reproduction of copyrighted material. Vendors should also confirm that AI-generated materials have sufficient human authorship for copyright protection.

Technical infrastructure

Despite their considerable power, implementing GenAI models requires extensive adaptation and technical infrastructure. This infrastructure can be thought of on a continuum of increasing sophistication and customization.

At the most accessible end, content developers can use proprietary systems like ChatGPT, Gemini or Claude directly with custom prompts. Moving toward greater technical complexity, organizations may still use proprietary systems, but develop custom tools like custom interfaces that make API (Application Programming Interface) calls to the proprietary GenAI system. Similarly, an organization could develop a corpus of high-quality examples that could be drawn on by the GenAI system through retrieval augmented generation. Even more complex, organizations could locally deploy open-source models with custom API implementations and potentially fine-tune these open-source models.

Each of these example approaches requires a mix of practical, logistical and technical know-how. As another parallel example, the Attali et al. (2022) item generation study illustrates the demands of more sophisticated implementations, wherein the authors had to use multiple natural language processing (NLP) routines to produce their results.

More generally, current research has not yet established clear best practices regarding whether and how to implement these various kinds of technical infrastructure, what the tradeoffs are in various approaches, and how these approaches can interface with current item development processes and software.

Security

LSA programs must keep their content secure, as content exposure degrades the validity of test results. Therefore, requests to and responses from LLMs should also remain secure. As discussed above, building a custom LLM is beyond the scope of most companies, which means that test vendors interested in leveraging AIs for item development must do so either through special license agreements with AI providers or by adapting and locally deploying one of several available open-source LLMs. The first option requires legal and technical expertise; the second requires extensive computing resources and specialized programming skills. Thus, keeping items secure introduces cost and complexity for SEAs (or their vendors) hoping to leverage GenAI for item development.

Efficiency

Automated item generation and review, whether human-assisted or not, requires higher up-front costs than traditional item authoring, since a tool or tools must be built to support this. Considering the current state of AI-based item-generation research, expert human review will continue to be necessary for LSA programs. Therefore, the feasibility of GenAI-based item generation for large-scale assessments relies on whether the combined costs of initial investment and the diversion of human expert resources from authoring to review result in a more efficient item development process. Current guidance on automatic item generation, outside of and predating GenAI, notes that AIG is more effective than human item development when the number of items is large, e.g., more than about 250 items, and there is an extended maintenance period (Kosh et al., 2019; see also

Sommer & Arendasy, 2025). Similar guidance on when and how AI based item generation is more cost effective than human authoring has not yet emerged.

These considerations underscore that adopting GenAI in LSA contexts is not simply a matter of creating and using new tools; it requires reengineering the test development process to address accuracy, bias, legal, technical, security, and efficiency issues. Such integration implies substantial upfront investment, testing, and ongoing expert oversight to ensure the resulting content meets the standards of LSA programs.

FIELD TESTING AND EQUATING

In LSA programs, content that passes preliminary reviews is typically field-tested. Field testing involves administering items to students to "test" the items—that is, to determine if they exhibit sound measurement characteristics. The primary goal of field testing is to estimate item parameters, such as difficulty, but field testing also provides an empirical check on problems that may slip past expert review—for example, selected-response items with more than one plausible answer or scoring guides that fail to differentiate adjacent performance levels on constructed-response tasks.

Field testing is one of the most expensive and operationally challenging aspects of assessment programs. Stand-alone field testing, in particular, can suffer from student motivation issues, and managing the logistics of administering large numbers of unscored items places a significant burden on states.

In the context of item response theory (IRT, the predominant measurement framework for LSAs), item parameters are required for form construction to ensure that operational tests are equivalent in difficulty (in fixed-form administrations) or properly tailored to examinees (in adaptive testing). Ongoing LSAs also require *equating*—the process of placing parameters for new items or test forms on the test scale to maintain continuity over time.

As we explain below, AI has the potential to support field testing and equating by modeling item characteristics and student responses, thereby reducing the number of examinees needed to obtain accurate parameter estimates. If successful, this could ease the financial, psychometric, and operational challenges that field testing imposes on states. At present, however, limitations in prediction accuracy and security require caution. We recommend that state education agencies explore AI as a supplemental tool, one that may improve efficiency and reduce burdens but not fully replace empirical data from actual students.

Applications

AI can be used to predict item parameters, identify items likely to be flagged for differential item functioning (DIF, a common tool for exploring potential bias), and estimate equating relationships using fewer examinees. For example, a common problem in field testing for LSA programs is optimizing the allocation of items to scarce slots for field testing. In programs with more items than slots, algorithms for AI-enhanced field testing might recommend items that are most likely to meet

item-difficulty targets and less likely to be flagged for problems such as DIF. These approaches can help increase the number of items that are successfully field tested, leading to increased item pool size and quality.

Predicting item difficulty and response times

There has been some interest in exploring the degree to which AI can be used to reduce the number of students needed for field testing. One way to do this is to obtain initial estimates of item parameters prior to field testing, through item difficulty modeling (IDM). For example, AI could be used to predict item parameters directly from the content of test items. Recent studies that aim to predict item difficulty (and response time, in some cases) (e.g., Li et al., 2025; Razavi & Powers, 2025; Veeramani et al., 2024; Yaneva et al., 2024) show promise but also demonstrate that LLMs fail to meet the prediction accuracy required for operational deployment without student test takers. AI-assisted item difficulty prediction could improve field testing design by informing test developers of likely difficulty (and response time) ranges for items they author.

Similarly, a study by Maeda (2024) found that one approach to obtaining item parameters using GenAI-generated test-takers rather than humans fell short of the accuracy obtained when using real students—an approach that could lead to serious errors if implemented.

However, Hao et al. (2024), citing McCarthy et al. (2021), note that LLMs can estimate item parameters with fewer students than conventional field testing. Liu, Bhandari, and Pardos (2025) demonstrated how progress can be made by augmenting limited student responses using an ensemble of LLMs to produce student-like responses. The results were better than using a limited human sample, but again not as good as using a full set of human responses.

Predicting differential item functioning

Like the work involved in predicting item difficulty and response time, studies have also emerged on predicting differential item functioning (e.g., Maeda & Lu, 2025; Wolandt & Kraus, 2025). Notably, Maeda and Lu (2025) used data from approximately 42,000 items to predict DIF and then associate DIF with specific words and phrases. Like the results from the item-difficulty and item-response time work, this study suggests that AI-based methods can provide a rough indication of DIF, potentially providing a valuable check during the item-writing process.

In a slightly different vein, the DIF of GenAI-created items has also been a subject of research. Belzak, Naismith, and Burstein (2023) showed that the overall degree of DIF of GenAI items is not notably different than that of human-written items, but the amount of DIF by subgroup differed. Like the consensus that item writing should keep humans in the loop, these results also suggest that GenAI-written items need to be closely inspected for DIF.

Considerations

LSA programs have relied on field testing with actual students to ensure that test items are functioning as expected, to provide the item statistics needed for data reviews, and to estimate item parameters to build test scales. Similarly, item response data from students has allowed testing programs to conduct equating, linking new items or forms onto existing test scales.

Using fewer students to field-test and equate, however, should be supported by strong evidence that the quality-control and parameter-estimation functions of life cycle stages can be maintained and provide the same results as full samples. Thus, one consideration in using AI to achieve this is the collection of evidence that the methods are robust enough to justify relying on only a portion of the usual field-test sample.

When studies support using fewer students, they may make assumptions that are difficult to implement in practice. For example, it may not be feasible to administer secure items to a proprietary model like ChatGPT, Gemini or Claude. Additionally, the structure of the testing program may make it difficult to reduce field-testing sample sizes without compromising the representativeness of the test-taking population, which could introduce bias. Similarly, timelines might prevent a program from conducting all the research needed to determine approach best supports a reduction in field test sample sizes. Research on using LLMs to assist with field testing and equating needs to align with the practical realities of operational programs.

ADMINISTRATION

Test administration encompasses the activities required to deliver large-scale assessments securely, fairly, and efficiently to students. It bridges the transition from test development to scoring, encompassing logistical coordination, technology delivery, proctoring, accommodations, and monitoring. Thus, this stage plays a crucial role in maintaining the validity and credibility of LSAs. As testing moves increasingly online and adaptive designs become more common, administration processes have become data-rich and highly dependent on digital infrastructure. These characteristics make them particularly amenable to advances in AI.

Currently, AI use in test administration is mainly limited to rule-based algorithms that schedule test sessions, monitor test-taker engagement, or manage accessibility tools. However, GenAI introduces possibilities for adaptive proctoring, dynamic test supports, and more responsive communication with examinees and administrators. Because administration occurs at the intersection of technology, logistics, and human behavior, it also raises some of the most sensitive questions in the LSA life cycle about privacy, bias, and surveillance. As with all life-cycle stages, the responsible use of AI in administration must be grounded in the principles of fairness, transparency, and human oversight articulated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) and in emerging frameworks for responsible AI in education (Johnson, 2025; U.S. Department of Education, 2024).

Applications

Intelligent scheduling and logistics

LSAs require optimizing test schedules and test form assignments within computer-based testing platforms. AI could extend current capabilities for accomplishing these tasks by dynamically modeling participation patterns, predicting where additional testing windows or accommodations will be needed, and generating real-time communications to district coordinators and administrators. For example, an AI system trained on historical participation data and district constraints could automatically propose rescheduling plans in response to weather-driven closures or network outages, reducing the burden on human coordinators. Similar scheduling systems are in use in other large-scale logistics contexts and could be adapted for LSA administration (NASEM, 2022).

Adaptive delivery and environmental control

Beyond logistics, AI can also support adaptive administration conditions. In computer-adaptive testing, algorithms already adjust item selection to match a test-taker's estimated proficiency. AI

could further adapt the *presentation* of the test environment. This might entail altering pacing, screen layout, or instructions in real time to support student engagement and reduce anxiety. An AI-driven interface might detect long pauses, repeated requests for clarification, or evidence of frustration and adjust accordingly, for example, by rephrasing directions or offering a short break. Such adaptivity, if properly validated and assigned (through personalized instructional plans, for example), could improve the consistency of testing conditions across populations, particularly for students with disabilities or language-learning needs (Laitusis, 2025; World Health Organization [WHO], 2024).

Proctoring and integrity monitoring

Automated proctoring systems are among the earliest AI applications in test administration. These systems typically rely on computer vision and rule-based detection of anomalies (e.g., multiple faces in frame). GenAI may enhance these systems by integrating multimodal signals—text, speech, and video—into a holistic analysis of testing behavior. For example, a model could integrate eye tracking, keystroke dynamics, and verbal utterances to determine whether a student is receiving outside assistance. It is worth noting that such applications raise significant concerns about privacy, equity, and potential bias, as false positives may disproportionately affect some groups. Research has shown that even well-intentioned AI proctoring systems can inadvertently penalize students based on skin tone, movement style, or background setting (Bulut et al., 2024). Therefore, GenAI-based proctoring should remain assistive, alerting human supervisors rather than making binding determinations about integrity.

GenAI-enhanced automated proctoring systems raise significant concerns about privacy, equity, and potential bias.

Accessibility and accommodations

AI-based assistive technologies are currently implemented to support students during assessments, particularly those with disabilities. Tools like chatbots and word-prediction programs can help all students engage more effectively with assessment tasks (Hollingsworth, 2024).

Laitusis (2025) notes that AI applications can function effectively as assistive technology during assessments. As she explains, these tools align with the WHO's definition of assistive technologies as products that "help maintain or improve an individual's functioning related to cognition, communication, hearing, mobility, self-care, and vision." (WHO, 2024, para. 2). Beyond traditional accommodations, GenAI can support cognitive tasks, Laitusis notes, such as pattern recognition, synthesis, and idea generation. This application of GenAI has the potential to ensure a more equitable testing environment.

However, although reviews of AI for students with learning disabilities document promising applications, they also underscore the scarcity of empirical work focused specifically on assessment contexts (Panjwani-Charania & Zhai, 2024; Marino et al., 2023). This makes it especially important, when designing such supports, to draw a clear delineation when an assistive function becomes a construct-irrelevant aid. For example, a paraphrase that simplifies a reading comprehension passage may change what is being measured. The design of such supports therefore must be anchored in the construct definitions established earlier in the LSA life cycle and validated accordingly.

Communication and support

GenAI can facilitate communication between students, administrators, and test coordinators. For instance, multilingual chatbots could provide immediate assistance to test administrators, explaining complex procedural steps or troubleshooting technical issues in real time. For examinees, GenAI-driven assistants could deliver consistent, scripted answers to questions during testing (within predefined constraints), reducing reliance on proctor discretion. Outside of testing windows, GenAI could help produce training materials, FAQs, and manuals tailored to different audiences, supporting scalability in statewide implementations.

Real-time monitoring and analytics

AI can be used to analyze test administration data as it is collected, tracking participation rates, detecting anomalous patterns (e.g., unexpected score distributions or rapid item responses; sometimes called "data forensics"), and flagging potential irregularities for follow-up. For example, Pan and Sinharay (2024) propose training an open-source LLM to detect response patterns and predict response times, which can be later used in an algorithm to flag potentially compromised (previously seen) test items. This form of continuous analytics could enhance existing test security and validity monitoring systems. In the long term, models such as these could also identify systemic access issues, such as persistent network bottlenecks in rural schools or scheduling disparities affecting specific student groups.

Considerations

Fairness, transparency, and explainability

The use of AI in administration requires careful consideration of fairness and explainability. As the *Standards* note, fairness is not limited to score interpretation but extends to every stage of the testing process. Automated proctoring or adaptive presentation systems must be transparent in their operations and open to audit. When AI systems make or inform real-time decisions about test delivery (such as pausing a session or flagging irregular behavior) examinees and administrators should have clear mechanisms to understand and contest those decisions. Explainable AI (XAI) methods (Li, Liu, & He, 2022) can support this by documenting the rationale for automated administrative actions.

Privacy and data governance

Administration systems capture sensitive multimodal data: video, audio, text logs, and behavioral metadata. When AI systems process these data, the risk of unauthorized access or misuse increases. SEAs must ensure that all data used for AI-enabled administration is governed by strict privacy and security protocols, consistent with FERPA (1974; 2024), COPPA (1998; 2023), and state data privacy laws. In recent years, some state data privacy laws have become increasingly specific and prescriptive and therefore merit close attention. Locally hosted or vendor-contained AI systems are preferable to public cloud-based models, which may store or reuse submitted data for model training. Transparency regarding data retention, audit trails, and deletion policies is essential.

Validity and construct representation

Adaptations of test content during administration can inadvertently alter what the test measures. For example, a GenAI system that rephrases questions to aid comprehension may blur the line between providing accessibility and changing the construct. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) emphasize that testing accommodations should remove irrelevant barriers without changing the intended meaning of scores. Therefore, every AI-based intervention in test administration should be reviewed through the lens of construct representation and empirically evaluated to ensure that it does not introduce construct-irrelevant variance.

Bias and surveillance risk

AI-driven monitoring and behavioral analytics can reproduce or amplify societal biases present in training data. An AI proctoring system might misinterpret cultural communication styles or environmental differences as suspicious behavior. As Bulut et al. (2024) observe, bias mitigation must be an explicit design goal rather than an afterthought. SEAs should ensure that any AI-based administrative tools are tested for disparate impact and that humans retain ultimate authority over enforcement decisions. Overreliance on AI surveillance may also erode public trust in LSAs; maintaining transparency about how monitoring systems operate is crucial for legitimacy.

SCORING

Automated scoring involves replicating the numerical scores that expert humans would assign to open-ended responses. Scoring student essays on a rubric from 0 to 4 is a prototypical example. Typically, expert human raters score a set of responses that are then used to train a larger set of human raters. The goal of this training is to have all the raters replicate the judgments made by the expert raters. The same criterion applies when an automated scoring engine assigns a rating. Human scoring, like content development, has been a major cost driver in LSAs,[7] and as such, using technology to reduce costs and increase efficiency has long been an area of active research.

> *For these innovations to be adopted in LSA settings, SEAs and vendors must attend to concerns about replicability, data privacy, and model validation before operational use.*

The automated scoring of long and short text responses is a relatively mature area in LSAs, with operational applications dating back at least two decades (e.g., Hao et al., 2024)—well before the advent of GenAI. Outside of GenAI, automated scoring has recently been applied to open-ended responses in mathematics (e.g., Fife, 2017) and spoken responses (e.g., Bernstein & Cheng, 2023; Lottridge et al., 2022). Studies have consistently found that well-developed automated scoring engines predating LLMs can achieve a level of agreement with human scores as high as the agreement between two human raters (Shermis & Hamner, 2013).

AI has the potential to further enhance automated scoring by improving the accuracy and flexibility of scoring models, reducing reliance on large volumes of training data, and enabling new approaches such as zero- or few-shot prompting.[8] Other applications of AI in automated scoring include modeling interrater variability in scores, refining rubrics, and building AI agents that can

[7] The cost for scoring the NAEP state samples in reading and mathematics is $2.5 million or about $8 per student (NASEM, 2022).

[8] "Few-shot" prompting in content development refers to prompting a GenAI system by providing several correct examples of the kind of output sought. This stands in contrast to "zero-shot" prompting, where no examples are provided.

score across a variety of contexts. For these innovations to be adopted in LSA settings, SEAs and vendors must attend to concerns about replicability, data privacy, and model validation before operational use.

Applications

Replicating human ratings

Automated scoring engines have traditionally been trained with human-rated data to replicate human expert scores. Using natural language processing (NLP) techniques, a scoring engine can extract linguistic features of essays or other open-ended responses and use the features most predictive of scores in its scoring algorithm. Today's scoring systems can now leverage the power of LLMs, which model language at a global level. By doing so, these systems can potentially improve the accuracy of their predictions, using less human-scored data in the process.

Indeed, as LLMs have become more widespread, popular, and accessible, those developing automated scoring systems have incorporated them into those systems. Many scoring systems are modifying some of the parameters from LLMs (that is, fine-tuning) to train prediction models on specific test items and subsequently use those prediction models. These models can, in turn, be part of a larger collection (often referred to as an "ensemble") of models for scoring, including traditional models. This means that models trained using LLMs and traditional models that predate LLMs are being used together to improve scoring.

For example, Cambium has been using LLMs in automated scoring since 2020, with open-source LLMs configured for classification (rather than text generation) (Lottridge, Ormerod, & Patel, 2024; Texas Education Agency, 2023). Ormerod and Kwako (2024) provide a recent example of the approach at the forefront of research in this area. The researchers fine-tuned several open-source models (e.g., Meta's LLaMA 2) on public essay and short-answer data, obtaining Quadratic Weighted Kappas (QWKs)[9] around 0.78 across eight essay prompts; a level on par with many commercial scoring engines. A more recent study by Atkinson and Palma (2025) incorporated traditional linguistic-features analysis with LLMs to achieve average QWKs of 0.83 across the same set of essays.

The use of LLMs in scoring has been very successful and is quickly becoming a permanent fixture of the state of the art in this life cycle stage of LSA programs. As Hao et al. (2024) explain, the architecture of LLMs has demonstrated excellent performance in automated scoring. For example, Hao et al. (2024) note that the winners of the 2021 NAEP Reading automated scoring competition all used this technology. Similarly, all but one of the 2023 NAEP Math automated-scoring challenge entries used an LLM (Whitmer et al., 2023).

The use of LLMs in scoring has been very successful and is quickly becoming a permanent fixture of the state of the art in this life cycle stage of LSA programs.

As LLMs continue to be adopted in automated scoring research and development, we may

[9] In evaluations of automated scoring, the primary metric is Quadratic Weighted Kappa (QWK), which measures score agreement (between humans and automated scoring models) adjusted for chance. Top-performing neural and feature-based automated scoring systems between 2015 and 2017 reported QWK values in the 0.7–0.85 range (Dong & Zhang, 2016; Riordan et al., 2017; Taghipour & Ng, 2016). (For context, human-to-human QWK on these tasks is typically around 0.7-0.8, indicating that the best models approached parity with human scoring consistency.)

see systems that push the boundary of how many human-scored responses are required to train a scoring engine. For example, an alternative that has captured much interest is accessing LLMs directly through prompting. In one scenario, LLMs can be provided with the same scoring guides and training materials given to human raters. The LLMs can then be instructed to apply these directly to student responses. Here it is important to note that LLMs can be used with little to no training data, allowing them to derive the scoring of responses in ways that previous scoring technologies could not. However, programs will still need validation data (from actual test takers) to assess the accuracy of automated scoring models. Moreover, the lower bounds for the number of human test takers will also be determined by an LSA program's requirements for item parameter estimation (see the section on field testing and equating).

A recent study showed that the most sophisticated LLM models (such as OpenAI's GPT 4) do not (yet) perform consistently well across different open-ended tasks to be relied upon for either of zero- or few-shot prompting (Chamieh, Zesch, & Giebermann, 2024; see footnote 8). Relatedly, GenAI can be used to identify and refine range-finding (see footnote 3) data for human review (S. Lottridge, personal communication, June 16, 2025), potentially improving on the effectiveness of scoring through few-shot prompting or other approaches that use smaller training samples.

By processing scoring instructions and the same sample responses that human raters receive during training, future AI raters might conduct scoring more like human experts and less like traditional automated systems trained on thousands of cases.

Identifying reasons for lack of agreement among raters

Trained raters often disagree. And some items have higher rates of disagreement than others. High interrater disagreement is a problem for the validity of scores because it means that scores depend on who is scoring. AI has the potential to model and subsequently predict this kind of disagreement, thereby improving item development (by identifying aspects of items or rubrics that lead to disagreement) and scoring processes (by identifying areas of the training materials or rubrics that need refinement).

In one potential application, multiple AI "raters" are developed that exhibit the same degree of interrater variability as human scorers. *Predicting* the likely human interrater agreement of newly authored candidate tasks could assist item developers in refining such open-ended tasks (or their scoring guides) to maximize predicted interrater agreement. (This application crosses over into item development.)

Refining rubrics

AI agents are instances of AI that have been trained to do specific tasks, for example, scoring middle-school English language arts essays from a particular testing program. AI agents can help refine the rubrics (or support range-finding) associated with open-ended tasks. When sample student responses are reviewed for scoring, an AI scoring agent might also be part of that exercise. It would generate explanations for its scores, be able to ask questions of other human scorers about their scores, and answer their questions about its scores (and its explanations). The agent could learn from the human scorers, and vice versa, and rubrics could be refined during this process to address edge cases and sticking points. Ideally, this rubric refinement improves both human-human interrater agreement and human-AI agreement, resulting in better items and scoring engines.

AI scoring agents

Taking the possibilities of AI scoring agents even further, we speculate that agents-in-training can become automated raters that develop a general ability to score new tasks by learning to apply

rubrics to responses across a range of open-ended tasks. In other words, if such automated raters are allowed to retain a memory of past scoring activity and feedback on that scoring, it may be possible for them to approximate the expertise of professional human scorers who rate responses across varied contexts. This kind of general-purpose agent would be extremely useful, as typical automated scoring applications are trained to score on specific prompts and are therefore not generalizable. The monitoring and evaluation of such agents' scoring behavior would be like that for human experts—checking for drift, internal consistency, strictness/leniency, etc.

Considerations

Balancing flexibility and consistency

Although generative AI systems are inherently probabilistic (producing slightly different outputs for the same input depending on model settings), this behavior is controlled or eliminated in LSA scoring contexts. When used as scoring engines, models are configured to operate deterministically, ensuring that identical responses consistently receive identical scores. However, in AI scoring agent applications, models interact with human raters, generate rationales, or ask clarifying questions. And in these applications, some degree of generative flexibility is both inevitable and potentially valuable. During rubric refinement or range-finding exercises, for instance, allowing the agent to produce varied explanations or alternative phrasings can surface ambiguities in rubrics and promote shared understanding among human scorers.

In such interactive phases, variability in the agent's responses can be constructive, supporting dialogue and rubric calibration. But once a scoring model or agent transitions from training or collaboration into operational scoring, determinism becomes paramount. At that point, its scoring output must be stable, auditable, and replicable. Thus, the challenge for AI scoring agents is to balance flexibility during learning with consistency during scoring.

Data security and privacy

LSA programs that employ automated scoring need full control over the software used to score responses. It would not be appropriate for an LSA to submit student responses to a publicly available, proprietary AI system such as ChatGPT, Gemini or Claude since those systems typically do not guarantee the security of user-submitted prompts (which in this case would include student data). As Hao, et al. (2024) note, use of LLMs for scoring must (1) adhere to United States federal and state regulations (such as FERPA and COPPA) dealing with data privacy for children and (2) protect test items as intellectual property (the content of items can be indirectly disclosed through student response). These data security and privacy considerations point to the development and deployment of an AI system under the scoring agency's control.

Validation of models before operational deployment

Madnani, Cahill, and Loukina (2023) argue that, to scale up and support the validity of scores, automated scoring engines must be robust, which they describe as "well-developed, well-tested, and well-documented" (p. 3). Robustness includes the scoring model and the whole pipeline of software used for training, evaluating, and delivering automated scores. The authors stress the importance of comprehensive testing—including unit and functional tests—to catch implementation errors and using version control to ensure software consistency across updates. A scoring model can be safely put into operational use only after piloting and testing, as Shermis and Lottridge (2019) argued before LLMs were widespread.

SEAs looking to leverage AI scoring should implement such systems gradually, ensuring thorough validation against human ratings, and testing automated scoring systems for possible bias (e.g.,

against language learners or students with disabilities). Shifts in scoring should be investigated for their potential to impact year-to-year comparisons.

These requirements help explain why many impressive novel features of public-facing LLMs, which appear promising for automated scoring, cannot be immediately incorporated into operational scoring in LSAs.

REPORTING

Reporting involves supporting the appropriate interpretation of assessment results and guiding their use. Restated, effective reporting translates results into accessible visualizations and language that help various audiences (students, families, educators, policymakers) understand the results and then act on those results appropriately.

Generally, LSA programs produce a constrained number of reports at varying levels of aggregation. For example, all LSA programs produce individual student reports (ISRs) and often produce classroom, school and district reports, as well as various spreadsheets for use by the field. These reports generally have a limited amount of customized content; the only text fields that might change in an ISR are the student's name and the description of their achievement level. In addition, the primary way student performance is communicated is through scale scores and achievement levels. These kinds of limitations are, in part, due to the need to ensure that there is a finite number of unique reports that can be generated and subjected to quality control. These reports are usually accompanied by an interpretative guide and, in some cases, additional supporting materials like one-page briefs, short videos, etc.

GenAI has the potential to reshape both reporting and the interpretive materials that support reporting, both during and after testing. These approaches challenge typical reporting practices, which often frame reporting as a post-administration process that produces a limited number of reports, each with a limited set of elements that are customized in a limited number of ways. Moving to reporting that incorporates GenAI means revisiting this framing and, potentially, adopting new approaches to providing results and ensuring their quality.

Applications

We propose that AI can be used (1) to support the development of more robust, detailed, or customized score reports and interpretive materials, and (2) to provide access to this information through immediate feedback and chat-style interfaces.

Score reports and interpretive materials

We suggest that score reports and interpretive materials can be improved in at least three ways with AI. First, GenAI could be used to develop more fine-grained descriptions of students, such as more detailed performance level descriptors, by applying AI methods to scores and assessment metadata. Such descriptions could then be used in future reporting, in much the same way current descriptors for performance or achievement levels are used. Second, AI systems might access not only scores and assessment metadata but also data elements not typically included in reporting,

such as the time students spend on items, item-response patterns, and tool use, to develop and validate student profiles. Third, GenAI could be used in a less constrained way, providing fully individualized feedback to students, without the use of predetermined descriptions or profiles.

Under the first approach, more detailed categories could be developed based on the achievement level descriptors, test specifications, item difficulties, or other design documentation. These categories could then be mapped to scale ranges. Critically, this first approach is designed to improve upon achievement-level reporting by relying on derived scores without re-analyzing student response data. LLMs can enhance score reporting by providing detailed narrative insights tailored to different audiences. These narratives can include specific skill strengths, examples of student strategies, and motivational language, turning a test result into a richer, more actionable account of performance.

A second approach employs AI to derive and validate performance profiles that can supplement typical achievement-level reporting. GenAI could derive those descriptions from patterns in student responses, without restricting itself to score-based categories, yielding a fixed set of score-independent reports of results that can then be applied to future students with similar patterns. Such systems could draw on item-level data and metadata, response process data, contextual variables (such as the student's curriculum), and other sources to offer information relevant to a reporting use case. Moving in that direction, Guo et al. (2024) use NAEP process data (such as timings and tool use) to produce "process profiles" that provide rich context for score interpretations, indicating whether a student is engaged or struggling and if they exhibit regulated tool use. Similarly, Bruno & Becker (2025) show how LLMs can be used to identify easy-to-explain features that predict student scores on open-ended items. These features can be communicated in reports to students, teachers, and parents.

Under the third approach, GenAI could create more individualized reports, with relatively unique reports (and accompanying interpretive materials) for each student. These individualized narratives can explain performance patterns at the item level and describe misconceptions, mirroring capabilities shown to improve outcomes in formative and tutoring settings (e.g., Pardos & Bhandari, 2024). In the context of LSAs, this third approach to reporting would require tests to provide this information by design, highlighting how reporting is intricately tied to construct definition and assessment design.

Cutting across these approaches are the amount and types of information provided to the LLM. At the most constrained, these approaches would be limited just to information from the assessment program, including item responses, process data, and test metadata, or more expansive information. At the least constrained, the LLM could draw on a variety of supporting materials through retrial augmented generation, including various instructional materials. These instructional materials could be intentionally agnostic to curriculum or tied directly to multiple curricula. The latter is more challenging, as the assessment program would need to both keep track of which curriculum students are enrolled in and also navigate the complexities involved when a state program connects to local curriculum choices. However, this kind of work is being explored in the interim-assessment space as a way to better connect results to curriculum-based next steps (e.g., Wertheim, Heck & Pruitt, 2025).

Finally, although this section has been focused mostly on reports, interpretive materials are an often-overlooked area of support that is ripe for improvement using GenAI. Typical practice involves an interpretive guide and, potentially, a limited set of supporting resources like one-page briefs,

short videos, and professional learning modules. GenAI can potentially be used to generate a much broader body of resources, in multiple formats, for a variety of audiences. Pushed to the extreme, this kind of development could result in highly differentiated materials that are directly connected to reports.

Immediate feedback

There has been considerable work on using GenAI to provide feedback in the context of formative assessment (e.g., Hopfenbeck et al., 2023, Kaliisa et al., 2025). These approaches often provide student feedback as students answer questions, with some systems going so far as to structure assessment and feedback in the form of dialogue with a chatbot (i.e., conversation-based assessment, see Yildirim-Erbasli & Bulut, 2023; also Pan et al., 2025). This kind of feedback cannot be provided while students are answering items in typical LSA contexts, where testing aims to pinpoint student achievement rather than enhance learning. However, we think GenAI-assisted feedback *could* be provided after students test, especially in computer-based administrations. The benefits of immediate feedback are well established (e.g., Black & Wiliam, 1998; Hattie & Timperley, 2007), helping students understand how they did and where they could improve. An added benefit of immediate feedback is testing program buy-in, as it provides an additional value-add to the testing experience.

Chat-style interfaces

All of the kinds of materials mentioned to date presume that reports and materials are provided through usual means (e.g., through web-based platforms, distributed to test coordinators, provided online). However, the materials themselves could also be made accessible via a chatbot interface, allowing users to query results and make sense through dialogue. The risk of this application of GenAI in reporting is that LLM-based chatbots are notorious for hallucinations (i.e., making up content), so such chatbots would need to be carefully designed to ensure users are not misled.

Considerations

The primary barrier to using AI to produce more detailed, personalized reporting is that LSA programs are not generally designed to support inferences at levels of specificity deeper than broad reporting categories. Therefore, customized reporting requires validation support beyond the psychometrics involved in creating scores and subscores.

The current *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) predate GenAI, but address automatically generated reporting:

> Standard 6.11 When automatically generated interpretations of test response protocols or test performance are reported, the sources, rationale, and empirical basis for these interpretations should be available, and their limitations should be described. (p.119)

This standard is particularly relevant as GenAI systems become more capable of generating interpretive narratives based on test results, as well as directly from test responses. The challenge for assessment programs seeking to leverage these capabilities will be to ensure that the automatically generated interpretations are empirically supported, transparent in their rationale, and clear about their limitations.

The two reporting-system design dimensions relevant for GenAI reporting applications

To understand how this can happen—that GenAI based reporting can be empirically supported, transparent and clearly delimited—we propose two dimensions along which reporting systems can be designed. One dimension is the data that the system uses to produce its reports. As mentioned,

traditional large-scale assessment programs use scores and subscores. However, a reporting system might go deeper, using responses to specific test items and the interrelationships among responses to all items. Additionally, a system might also utilize external data, such as information about the student's curricular experiences or past scores, in generating reports.

The second dimension describing a reporting system's design is the extent to which the language used in reports is predetermined before testing or generated on the fly. The above-mentioned Guo et al. (2024) application exemplifies the former, because the authors validate their process profiles and use them in operational testing. (That is, these profiles are not generated anew with each new administration.)

Taking the first dimension first (the data used), the more a system draws upon non-score information (such as item responses or external data about a student), the more difficult it is to account for, as the *Standards* put it, "the sources, rationale, and empirical bas[e]s" of the reports.

AI-generated reporting based on raw response data may employ machine learning models that do not share the same construct representation as the psychometric model used to generate scores. Research highlights the importance of model consistency and warns of risks when AI-based inferences contradict or bypass validated scoring models (Bulut et al., 2024).

Although a test can support inferences through more than one model, the models involved should be consistent. For example, suppose a traditional IRT model classifies students into performance levels, while an AI-based model describes students based on their raw responses to specific items using the same test. Furthermore, suppose many students classified as proficient by one model also require significant support on key learning standards according to the AI-based model. In that case, it is difficult to understand how these two models can operate in tandem.

Turning to the second dimension (report language predetermination), systems that generate report language on the fly are significantly more challenging to validate than those where the reporting language is predetermined or restricted in some manner. One would need to ensure that the system rarely (if ever) makes an inference or recommendation that would not have been endorsed by those who designed the test. Consequently, such systems are unlikely to be feasible for today's LSA programs. We believe that in the foreseeable future, leveraging AI at the reporting stage of the LSA life cycle may be limited to generating and validating a pre-defined set of interpretations prior to operational administration.

The role of explainability

Explainable AI (XAI; Li, Liu, & He, 2022) emphasizes transparency and provides mechanisms for users to interrogate and understand system outputs, making it a natural organizing principle for AI-enhanced reporting systems in large-scale assessment.

XAI is essential for accountability, especially when natural language is generated to convey scores, feedback, or learning recommendations. Reports must include mechanisms to trace generated text back to the evidence or student responses that support it (Zapata-Rivera & Katz, 2014). Moreover, integrating guardrails, such as reporting templates vetted during the test development phase, can constrain generative systems to produce outputs only within the scope of validated interpretations. For AI-enhanced reporting to support valid score interpretation and use, the reporting system design must uphold the same evidentiary standards that apply to score production, ensuring that the report remains a valid and trustworthy communication tool.

OTHER APPLICATIONS

LSA programs require ongoing maintenance and validation. AI is playing an increasingly prominent role in processes that cut across the lifecycle stages. In this section, we cover AI applications in translation, alignment, standard setting, program operations, and documentation.

Translation

In LSAs, items, test reports, and other material often need to be translated. Translations through ChatGPT with the GPT-4 engine have been found to be comparable to commercial translation products, including for low-resource or distant languages (Jiao et al., 2023). In the future, this kind of application may provide greater access to translated assessments. Related to this, Jung, Tyack, and von Davier (2024) report on a creative application of automated translation in a multilingual context, where students responded to math and science prompts in their native language. In the study, open-ended responses in various non-English languages were translated to English through ChatGPT and then scored with an automated engine. The researchers found that the automated scoring was comparable to human scoring of the native language responses, and that, moreover, the psychometric characteristics derived from those automated scores were similar to those obtained from human scores.

Considerations

The above-referenced Jung, Tyack, and von Davier (2024) study illustrates an application of translation in which the generated text is not read by humans, but used as input to a scoring engine. In general, however, translations presume a human reader. Thus, GenAI translations should be subject to the same standards of accuracy and quality as human translations. The *International Test Commission Guidelines for Translating and Adapting Tests* (ITC, 2017) articulate best practices including forward and backward translation, independent committee review, reconciliation procedures, and evaluation of semantic, functional, and cultural equivalence. These guidelines have been widely adopted in international and national assessment programs.

Alignment

Aside from automated scoring, several large-scale assessment processes require human experts to undertake judgmental tasks that involve reviewing a significant amount of textual information and result in ratings of various forms. These ratings might pertain to the relative difficulties of items, the alignment of items with standards or performance descriptors, or other mappings that provide evidence to support test validation. The mapping of items to standards is a core component of many approaches to alignment, for example the Webb (1999) approach to alignment, and a key part of test development generally.

Butterfuss and Doran (2024) and Camili (2025) describe how AIs can be deployed to sort pairs of statements from different content standards using similarities computed by LLMs, effectively reducing the time and cognitive load of aligning standards between different frameworks. Karimi-Malekabadi, Razavi, and Powers (2025) and Xu et al. (2025) show that pre-trained language models can effectively detect alignment between items and skill statements.

These applications leverage the power of AI without generating text, but it is easy to imagine other kinds of alignment studies where GenAI systems are prompted to provide rationales to accompany ratings or similarity computations. When investigating alignment as part of test development, GenAI systems can also be prompted to recommend item revisions to improve the alignment between a content standard and an item written to that standard.

Considerations

Alignment is a critical component of test validation in LSA programs. The U.S. Department of Education's peer review guidance (2018) calls for evidence of alignment between academic content standards and college-level credit-bearing coursework / career and technical education standards (p. 30), required assessments and the state's academic content standards (pp. 31, 36), English-language proficiency (ELP) assessments and a state's ELP standards (p. 48), and academic achievement standards and content standards (p. 68). Given the high-stakes nature of these assessments, it is difficult to see how evidence derived solely from AI systems will suffice. For the foreseeable future, applications of AI to alignment in LSA contexts are likely to take the form of supports for test developers and alignment study participants.

Standard setting

Standard setting is the process whereby representative panels of individuals with knowledge of the tested population and content expertise render judgments that result in cut scores on tests. Those cut scores are used to classify students into performance or achievement levels. Lewis et al. (2025) used LLMs to provide ratings for the item-descriptor match (IDM) standard-setting method (Ferrara, Perie, & Johnson, 2014), comparing these to human ratings. The researchers found that when ChatGPT-5 was prompted either to simulate human standard-setting "personas" or to act directly as an expert panelist, the LLM-derived cut scores approximated human-derived cuts to varying degrees: an "AI as expert panelist" method produced cut scores close to the original human panel (within the standard error for one of the two cut points), whereas an "AI personas" method produced somewhat more variable results. Overall, the study concluded that although AI cannot yet replace human panelists, LLMs show promising accuracy in reproducing human IDM judgments. The matching task in IDM is structurally similar to alignment judgments, where LLMs show promise, as noted previously.

Another standard-setting application of AI—in this case, GenAI—is the creation of performance level descriptors (PLDs). This application has recently been discussed by Amoateng and Ricker-Pedley (2025), who investigated whether large language models could draft PLDs and borderline performance descriptors (BPDs), and render item-level Angoff judgments for state summative assessments in Grades 3–8 ELA and mathematics. Using Claude 4.0 Sonnet within Pearson's internal AI content ecosystem, they found that AI-generated PLDs were generally reasonable in content but differed enough in organization and focus that human experts would still need to refine them. AI-generated BPDs were similarly plausible but often not sufficiently distinct from PLDs, indicating that human editing remains essential. For item-level judgments, the AI was "hit or miss," tending to be more stringent than human panelists, with somewhat better alignment in mathematics and at higher grade levels. The authors emphasized that AI-generated PLDs and BPDs may serve as useful starting points for human reviewers, potentially increasing efficiency, but cannot replace expert judgment.

Considerations

As emphasized in the most recent comprehensive treatment of standard setting, the process is inherently both cognitive and social (Ferrara et al., 2025). Its political aspects are encapsulated in the opening sentence of Ferrara et al.: "Standard setting is policymaking" (p. 822). Although standard-setting methods are often highly structured, nearly all involve group processes through which participants share their perspectives on rigor, expectations, experience of student performance, and the reasonableness of different cut scores. Moreover, the validity of a particular standard setting study, which results in consequential cut scores, relies to a great degree on its fidelity to a pre-

determined procedure. These aspects of standard setting are very difficult to account for in a simulated process, meaning that the defensibility of the results must be stringent attend to, almost invariably through a low-dependency HITL application of AI.

Program operations and documentation

LSA programs rely heavily on high-quality documentation to document procedures and support their implementation. GenAI has the potential to provide initial drafts of assessment analysis specifications, procedures manuals, business rule documents, technical reports and the like. It can draw on extant code, examples from other programs, or prior versions of the document, saving time in initial development. Related to this, GenAI could be used to review current documentation and suggest improvements, and could even do so by comparing code to current documentation for discrepancies.

GenAI can also increase the speed and quality of coding, particularly coding that does not involve proprietary software. This could in turn lead to better and more robust analysis, vetting and exploration of various aspects of the program, including item bank analysis and investigations of assessment results.

Evidence from software engineering and technical writing suggests GenAI is particularly well-suited to "first-draft" and "maintenance" work for technical documentation (e.g., summarizing code and generating structured documentation), provided outputs are treated as drafts and verified by staff (Hou et al., 2024).

Considerations

LSA program documentation often includes proprietary information, and program data certainly needs to remain secure. This requires that organizations producing LSA program documentation (including code) address AI policy and standard operating procedures for AI use. This will in turn determine which AI tools one should incorporate in their program operations and documentation processes. One option is to have a tiered approach to AI use depending on context.

At the lower tier, the use of publicly available, proprietary GenAI tools may be permitted for low-risk tasks, such as producing generic documentation templates or summarizing non-confidential material. At the higher tier, working directly with program documentation and code might require either (1) a (technology) vendor-provided "enterprise" offering or secure API-based access with contractual assurances about data handling, retention, and non-training on submitted content or (2) locally-hosted or private-cloud instances of open-source LLMs that are then adapted for program-specific documentation tasks. The second of these offers the greatest degree of control over data security and model behavior. However, as mentioned earlier in this paper, adapting open-source LLMs requires substantial upfront investment in infrastructure, engineering expertise, and ongoing maintenance. Organizations must also consider the costs associated with calls to AIs.

Regardless of the technical approach, several considerations apply. First, documentation generated or revised with GenAI should be versioned, traceable, and auditable, with clear attribution of human review and approval. Second, organizations should explicitly define which kinds of documentation are appropriate for which tier of AI assistance. Third, staff must be trained not only in how to use GenAI tools, but also in how to critically evaluate their outputs, recognizing that fluent prose can mask omissions, inconsistencies, or errors.

CONCLUSION

AI is reshaping the practical and conceptual foundations of large-scale assessment. Across the life cycle, its most immediate value lies in augmenting—not replacing—expert human judgment: assisting framework writers, generating and reviewing items, predicting field-test outcomes, supporting adaptive administration, enhancing scoring consistency, improving reporting, and assisting with translation, alignment, standard setting, and program operations and documentations. Each of these AI application areas can increase efficiency and insight, but each also carries risks that must be managed through transparency, validation, and governance.

This paper cited findings in some of the more-researched applications of AI in the LSA life cycle, such as content development, field testing, and scoring. But, unlike most innovations in large-scale assessment, AI development is occurring on week-to-week and month-to-month timelines. New models, architectures, and deployment strategies are likely to alter what is feasible in all LSA stages. Although the findings we present here will become dated as research advances, the guiding principle should stay the same: AI can expand human capacity in assessment only if it remains accountable to human expertise and the public purposes these programs serve.

AI can expand human capacity in assessment only if it remains accountable to human expertise and the public purposes these programs serve.

It is also important to acknowledge that much of the most advanced AI development in assessment is likely occurring behind closed doors, as vendors pursue proprietary tools and workflows to gain competitive advantage. As a result, publicly available research and guidance may lag behind operational practice. This reality heightens the importance of principled evaluation by SEAs: procurement decisions, RFP language, and change requests will increasingly determine how AI is used in practice.

For SEAs, the central challenge lies in when, where, and how to integrate AI into established LSA life-cycle stages. Successful implementation will depend on clearly defined human oversight, locally controlled technical environments, and evidence that AI-assisted processes meet the same standards of fairness, accuracy, and security as traditional methods. Across all stages of the lifecycle, maintaining low-dependency human-in-the-loop relationships remains critical for preserving validity and public trust.

AI offers real opportunities to improve efficiency, coherence, and responsiveness across assessment systems. Our aim is to encourage principled, measured adoption—one that aligns innovation with the public purposes of assessment and the longstanding professional standards that govern it. Used thoughtfully, AI can strengthen large-scale assessment programs. Used uncritically, it risks undermining the very validity, fairness, and credibility that those programs exist to ensure.

IMPLICATIONS FOR STATE EDUCATION AGENCIES

For state education agencies, the implications of AI in large-scale assessment are practical and immediate. SEAs need not become AI developers, but they do need sufficient technical and conceptual fluency to evaluate AI offerings in proposals, set expectations in procurement, and

oversee implementation responsibly. This includes articulating where AI use is permissible across the assessment life cycle, requiring evidence that AI-assisted processes preserve validity and fairness, and ensuring that human oversight is clearly defined and documented.

SEAs should also attend explicitly to return on investment, prioritizing applications where AI demonstrably reduces cost, improves quality, or enhances interpretability without introducing undue risk. Finally, as AI capabilities evolve rapidly and unevenly across vendors, SEAs are well positioned to serve as stewards of transparency and public trust—by insisting that innovation in assessment remains aligned with professional standards, legal requirements, and the educational purposes these programs are designed to serve.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Amoateng, E, & Ricker-Pedley, K. (2025, October 27–29). *AI-enhanced standard setting: Leveraging AI to support human experts*. [Conference presentation]. Artificial Intelligence in Measurement and Education Conference (AIME-Con), Philadelphia, PA.

Anthropic. (n.d.). https://www.anthropic.com

Artsi, Y., Sorin, V., Konen, E., Glicksberg, B. S., Nadkarni, G., & Klang, E. (2024). Large language models for generating medical examinations: Systematic review. *BMC Medical Education, 24*, Article 354. https://doi.org/10.1186/s12909-024-05239-y

Atkinson, J., Palma, D. An LLM-based hybrid approach for enhanced automated essay scoring. *Sci Rep* 15, 14551 (2025). https://doi.org/10.1038/s41598-025-87862-3

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, Article 903077. https://doi.org/10.3389/frai.2022.903077

Baudin, J. S. P. (2025). Assessing the psychometric properties of AI-generated multiple-choice exams in a psychology subject. *Journal of Pedagogical Sociology and Psychology, 7*(3), 18-34. https://doi.org/10.33902/jpsp.202536891

Bedi, S., Fleming, S. L., Chiang, C.-C., Morse, K., Kumar, A., Patel, B., Jindal, J. A., Davenport, C., Yamaguchi, C., & Shah, N. H. (2024). QUEST-AI: A system for question generation, verification, and refinement using AI for USMLE-style exams. *medRxiv.* https://doi.org/10.1101/2023.04.25.23288588

Belzak, W. C. M., Naismith, B., & Burstein, J. (2023). *Ensuring fairness of human- and AI-generated test items.* In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (pp. 701–707). Springer. https://doi.org/10.1007/978-3-031-36336-8_108

Bernstein, J. C., & Cheng, J. (2023). Speech analysis in assessment. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 31–57). Routledge. https://doi.org/10.4324/9781003278658-4

Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of ChatGPT-generated questions. *Computers and Education: Artificial Intelligence, 7*, 100284. https://doi.org/10.1016/j.caeai.2024.100284

Black, P., & Wiliam, D. (1998). *Assessment and classroom learning.* Assessment in Education, 5(1), 7–74. https://doi.org/10.1080/0969595980050102

Bruno, J. V., & Becker, L. (2025). *Explainable writing scores via fine-grained, LLM-generated features.* In Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) (Vol. 2, *Works in Progress*, pp. 155–165). National Council on Measurement in Education. https://aclanthology.org/2025.aimecon-wip.19/

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). *The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges.* arXiv. https://arxiv.org/abs/2406.18900

Burstein, J. (2025, April 17). *The Duolingo English Test Responsible AI Standards* (Duolingo Research Report DRR-25-05). Duolingo. https://drive.google.com/file/d/1a39zAJvka-cRGcked4Da0sEYd0yN_GSX/view

Butterfuss, R., & Doran, H. (2024). An application of text embeddings to support alignment of educational content standards. *Educational Measurement: Issues and Practice.* Advance online publication. https://doi.org/10.1111/emip.12641

Camilli, G. (2024). An NLP crosswalk between the Common Core State Standards and NAEP item specifications. *arXiv preprint arXiv:2405.17284.* https://doi.org/10.48550/arXiv.2405.17284

Chamieh, I., Zesch, T., & Giebermann, K. (2024). LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 309–315). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.25/

Chan, K. W., Ali, F., Park, J., Sham, K. S. B., Tan, E. Y. T., Chong, F. W. C., Qian, K., & Sze, G. K. (2025). Automatic item generation in various STEM subjects using large language model prompting. *Computers and Education: Artificial Intelligence,* 8(100344). https://doi.org/10.1016/j.caeai.2024.100344

Chauhan, A., Khaliq, F. & Nayak, K.R. (2025) Assessing quality of scenario-based multiple-choice questions in physiology: Faculty-generated vs. ChatGPT-generated questions among phase I medical students. *International Journal of Artificial Intelligence in Education,* 35, 2315–2344. https://doi.org/10.1007/s40593-025-00471-z

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PloS one, 18*(8), e0290691. https://doi.org/10.1371/journal.pone.0290691

Children's Online Privacy Protection Act of 1998, 15 U.S.C. §§ 6501–6506 (1998). https://www.law.cornell.edu/uscode/text/15/chapter-91

Children's Online Privacy Protection Rule, 16 C.F.R. pt. 312 (2023). https://www.ecfr.gov/current/title-16/chapter-I/subchapter-C/part-312

Chuks-Okeke, E., Linero, N., & Leong, B. (2024, August 7). *Generative AI and intellectual property: Copyright implications for AI inputs, outputs.* International Association of Privacy Professionals (IAPP). https://iapp.org/news/a/generative-ai-and-intellectual-property-copyright-implications-for-ai-inputs-outputs

Coursera Staff. (2024, April 3). What is artificial intelligence? Definition, uses, and types. *Coursera.* https://www.coursera.org/articles/what-is-artificial-intelligence

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://arxiv.org/abs/1810.04805

Dong, F., & Zhang, Y. (2016). Automatic features for essay scoring – An empirical study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1072–1077). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1101

European Commission & Organisation for Economic Co-operation and Development. (2025). *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education (Draft).* AILit Framework. https://ailiteracyframework.org/wp-content/uploads/2025/05/AILitFramework_ReviewDraft.pdf

Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g (1974). https://www.law.cornell.edu/uscode/text/20/1232g

Family Educational Rights and Privacy Act, 34 C.F.R. pt. 99 (2024). https://www.ecfr.gov/current/title-34/subtitle-A/part-99

Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., Otero Ornelas, N., & Lan, A. (2024). Exploring automated distractor generation for math multiple-choice questions via large language models. *arXiv preprint arXiv:2404.02124.* https://arxiv.org/abs/2404.02124

Ferrara, S., Davis-Becker, S., Kannan, P., & Reynolds, K. (2025). Standard setting: A cognitive and social model. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 821–894). Oxford University Press. DOI: 10.1093/oso/9780197654965.003.0012

Ferrara, S., Perie, M., & Johnson, E. (2014). Matching the Judgmental Task with Standard Setting Panelist Expertise: the Item-descriptor (id) Matching Method. *Journal of Applied Testing Technology, 9*(1), 1–20. https://www.jattjournal.net/index.php/atp/article/view/48346

Fife, J. H. (2017). The m-rater Engine: Introduction to the automated scoring of mathematics items. *Research Memorandum, ETS RM-17-02.* https://www.ets.org/Media/Research/pdf/RM-17-02.pdf

Goodwin, S., Poe, M., Cardwell, R., Runge, A., Attali, Y., Mulcaire, P., Lo, K.-L., & LaFlair, G. T. (2024, May 21). *Facilitating the writing process on the DET: The interactive writing task* [White paper]. Duolingo. https://duolingo-papers.s3.amazonaws.com/other/interactive-writing-whitepaper.pdf

Google DeepMind. (n.d.). https://deepmind.google

Gorgun, G., & Bulut, O. (2024). Instruction-tuned large-language models for quality control in automatic item generation: A feasibility study. *Educational Measurement: Issues and Practice*, 44(1), 96–107. https://doi.org/10.1111/emip.12663

Guo, H., Johnson, M. S., Ercikan, K., Saldivia, L., & Worthington, M. (2024). *Large-scale assessments for learning: A human-centred AI approach to contextualizing test performance. Journal of Learning Analytics, 11*(2), 229–245. https://doi.org/10.18608/jla.2024.8007

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice, 43*(2), 16-29. https://doi.org/10.1111/emip.12602

Hattie, J., & Timperley, H. (2007). *The power of feedback. Review of Educational Research,* 77(1), 81–112. https://doi.org/10.3102/003465430298487

Hollingsworth, H. (2024, December 26). AI is a game changer for students with disabilities. Schools are still learning to harness it. *Associated Press.* https://apnews.com/article/ff1f51379b3861978efb0c1334a2a953

Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). *Challenges and opportunities for classroom-based formative assessment and AI: A perspective article.* Frontiers in Education, 8, Article 1270700. https://doi.org/10.3389/feduc.2023.1270700

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., & Wang, H. (2024). *Large language models for software engineering: A systematic literature review.* ACM Transactions on Software Engineering and Methodology, 33(8), Article 220. https://doi.org/10.1145/3695988

International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests (Second edition).* [www.InTestCom.org]. intestcom.org/files/guideline_test_adaptation_2ed.pdf

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). *Survey of hallucination in natural language generation. ACM Computing Surveys, 55*(12), Article 271. https://doi.org/10.1145/3571730

Jiao, W., Wang, W., Huang, J.-t., Wang, X., Shi, S., & Tu, Z. (2023). *Is ChatGPT a good translator? Yes with GPT-4 as the engine* (arXiv preprint arXiv:2301.08745). https://doi.org/10.48550/arXiv.2301.08745

Johnson, M. S. (2025). *Responsible AI for measurement and learning: Principles and practices* (Research Report No. RR-25-03). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-25-03.pdf

Jung, J.Y., Tyack, L. & von Davier, M. Combining machine translation and automated scoring in international large-scale assessments. *Large-scale Assess Educ* 12, 10 (2024). https://doi.org/10.1186/s40536-024-00199-7

Karimi-Malekabadi, F., Razavi, P., & Powers, S. (2025, April 25). *ChatGPT is an Effective Tool for Evaluating Educational Content Alignment.* [Conference paper]. Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Kaya, M., Sonmez, E., Halici, A., Yildirim, H., & Coskun, A. (2025). Comparison of AI-generated and clinician-designed multiple-choice questions in emergency medicine exam: a psychometric analysis. *BMC medical education, 25*(1), 949. https://doi.org/10.1186/s12909-025-07528-6

Kıyak, Y. S., Soylu, A., Coşkun, Ö., Budakoğlu, I. İ., & Peker, T. V. (2025). Can ChatGPT Generate Acceptable Case-Based Multiple-Choice Questions for Medical School Anatomy Exams? A Pilot Study on Item Difficulty and Discrimination. *Clinical anatomy (New York, N.Y.), 38*(4), 505–510. https://doi.org/10.1002/ca.24271

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). *A Cost–Benefit Analysis of Automatic Item Generation.* Educational Measurement: Issues and Practice, 38(1), 48–53. https://doi.org/10.1111/emip.12237

Kūkea-Shultz, P., & Brockmann, F. (2025). *Bridging psychometric and content development practices with AI: A community-based workflow for augmenting Hawaiian language assessments* (arXiv preprint arXiv:2512.17140. https://doi.org/10.48550/arXiv.2512.17140

Laitusis, C. (2025, April 9). Generative AI and Cara's other favorite accessibility things. The National Center for the Improvement of Educational Assessment. https://www.nciea.org/blog/caras-favorite-accessibility-things/

Law, A. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Kei-Ching Hung, K., & Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC medical education, 25*(1), 208. https://doi.org/10.1186/s12909-025-06796-6

Lee, K., Mills, K., Ruiz,P., Coenraad, M., Fusco, J., Roschelle, J., & Weisgrau, J. (2024, June). *AI Literacy: A Framework to Understand, Evaluate, and Use Emerging Technology.* Digital Promise. https://doi.org/10.51388/20.500.12265/218

Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice, 39*(1), 8–21. https://doi.org/10.1111/emip.12318

Lewis, J., Sireci, S. G., Tran, P., Zenisky, A. L., & Ketan. (2025, October 28). *Comparing human and AI standard setting results: Are standard setting panelists obsolete?* Paper presented at the Artificial Intelligence in Measurement & Education Conference, Pittsburgh, PA.

Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., & Lissitz, R. W. (2025). Item Difficulty Modeling Using Fine-tuned Small and Large Language Models. *Educational and psychological measurement,* 00131644251344973. Advance online publication. https://doi.org/10.1177/00131644251344973

Li, J., Liu, X., & He, J. (2022). Explainable artificial intelligence (XAI) for education: A review. *Computers and Education: Artificial Intelligence, 3,* 100074. https://doi.org/10.1016/j.caeai.2022.10007

Liu, Y., Bhandari, S., & Pardos, Z. A. (2025). Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology,* 56, 1028–1052. https://doi.org/10.1111/bjet.13570

Lottridge, S., Ormerod, C., Jafari, A., & Godek, B. (2022). Automated speech scoring methods and results. Cambium Assessment, Inc.

Lottridge, S., Ormerod, C., & Patel, M. (2024). Redesigning automated scoring engines to include deep learning models. In Shermis, M., & Wilson, J. (Eds.), *The Routledge International Handbook of Automated Essay Evaluation.* New York, NY: Taylor and Francis.

Madnani, N., Cahill, A., & Loukina, A. (2023). The role of robust software in automated scoring. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 3–14). Routledge. https://doi.org/10.4324/9781003278658-2

Maeda, H. (2024). Field-testing multiple-choice questions with AI examinees. *Research Square.* https://doi.org/10.21203/rs.3.rs-3858355/v1

Maeda, H., & Lu, Y. (2025). *Finding words associated with DIF: Predicting differential item functioning using LLMs and explainable AI* (arXiv preprint arXiv:2502.07017). https://doi.org/10.48550/arXiv.2502.07017

Marino, M. T., Hayes, L., Black, A. C., & Beecher, C. C. (2023). *The future of artificial intelligence in special education: Opportunities, challenges, and policy considerations.* Journal of Special Education Technology, 38(4), 271–284. https://files.eric.ed.gov/fulltext/EJ1387002.pdf

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021, November). Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 883–899). 10.18653/v1/2021.emnlp-main.67

Meta AI. (n.d.). https://ai.meta.com/llama

Midjourney. (n.d.). https://www.midjourney.com/home

Mistral AI. (n.d.). https://mistral.ai

Mittal, A. (2024, January 9). *The plagiarism problem: How generative AI models reproduce copyrighted content.* Unite.AI. https://www.unite.ai/the-plagiarism-problem-how-generative-ai-models-reproduce-copyrighted-content/

National Academies of Sciences, Engineering, and Medicine. (2022). *Modernizing the National Assessment of Educational Progress: Getting it right.* The National Academies Press. https://doi.org/10.17226/26389

National Assessment Governing Board. (2023). *Science framework for the 2028 National Assessment of Educational Progress.* U.S. Department of Education. https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/science/2028-naep-science-framework.pdf

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework

OpenAI. (2022, November 30). "ChatGPT: Optimizing Language Models for Dialogue." OpenAI. Retrieved from https://openai.com/blog/chatgpt.

OpenAI. (n.d.). DALL·E 3. https://openai.com/index/dall-e-3/

Ormerod, C., & Kwako, A. (2024). *Automated text scoring in the age of generative AI for the GPU-poor.* arXiv. https://arxiv.org/abs/2407.01873

Owen, N. (n.d.). *Evolving test specifications: From paper to AI.* LinkedIn. https://www.linkedin.com/pulse/evolving-test-specifications-from-paper-ai-nathaniel-owen-nx8de

Pan, Z., Ba, S., Jiang, Z., & Li, C. (2025). *Patterns of inquiry, scaffolding, and interaction profiles in learner–AI collaborative math problem-solving.* In Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) (Vol. 1, pp. 297–305). National Council on Measurement in Education. https://aclanthology.org/2025.aimecon-main.32.pdf

Pan, Y., & Sinharay, S. (2024). Detecting Compromised Items in Computerized Linear Testing: A Novel Approach Using Autoencoders and BERT, *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语期刊*: Vol. 5: Iss. 3, Art. 7. https://doi.org/10.59863/XSSZ8498

Panjwani-Charani, S. & Zhai, X. (2024). AI for Students with Learning Disabilities: A Systematic Review. In X. Zhai & J. Krajcik (Eds.), Uses of Artificial Intelligence in STEM Education (pp. 469-493). Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780198882077.003.0021

Pardos, Z. A., & Bhandari, S. (2024). *ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills.* PLoS ONE, 19(5), e0304013. https://doi.org/10.1371/journal.pone.0304013

Razavi, P., & Powers, S. J. (2025). *Estimating item difficulty using large language models and tree-based machine learning algorithms.* arXiv. https://doi.org/10.48550/arXiv.2504.08804

Riordan, B., Horbach, A., Cahill, A., & Yannakoudakis, H. (2017). Investigating neural architectures for short answer scoring. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 159–168). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5018

Runge, A., Attali, Y., LaFlair, G. T., Park, Y., & Church, J. (2024). A generative AI-driven interactive listening assessment task. *Frontiers in Artificial Intelligence, 7,* 1474019. https://doi.org/10.3389/frai.2024.1474019

Rupp, A. & Lorié, W. (2023, April 19). Ready or Not: AI Is Changing Assessment and Accountability. [blog]. nciea.org/blog/ready-or-not-ai-is-changing-assessment-and-accountability

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions.* (pp. 298–312). New York.

Shermis, M. D., & Lottridge, S. (2019, April 7). *Communicating to the public about machine scoring: What works, what doesn't.* Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada. https://files.portal.cambiumast.com/corporate-site/documents/CAI-Cambium-CommunicatingPublicMachineScoring-WhitePaper.pdf

Shultz, B., DiDomenico, R. J., Goliak, K., & Mucksavage, J. (2025). Exploratory Assessment of GPT-4's Effectiveness in Generating Valid Exam Items in Pharmacy Education. *American journal of pharmaceutical education,* 89(5), 101405. https://doi.org/10.1016/j.ajpe.2025.101405

Sommer, M., & Arendasy, M. (2025). Automatic- and Transformer-Based Automatic Item Generation: A Critical Review. *Journal of Intelligence, 13*(8), 102. https://doi.org/10.3390/jintelligence13080102

Song, Y., Du, J., & Zheng, Q. (2025). Automatic item generation for educational assessments: a systematic literature review. *Interactive Learning Environments, 33*(9), 5386–5405. https://doi.org/10.1080/10494820.2025.2482588

Stanford Institute for Human-Centered AI. (2021, October 14). *AI in the loop: Humans must remain in charge.* Stanford HAI. https://hai.stanford.edu/news/ai-loop-humans-must-remain-charge

Sykora, C. (2024, August 2). *A new approach to updating the ISTE Standards.* International Society for Technology in Education (ISTE). https://iste.org/blog/a-new-approach-to-updating-the-iste-standards

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1191

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., et al. (2025). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education, 12*(2), 317-340. https://doi.org/10.21449/ijate.1602294

Texas Education Agency. (2023). *The State of Texas Assessments of Academic Readiness (STAAR®) Hybrid Scoring Study: Methods and Results, Spring 2023 Items.* https://tea.texas.gov/student-assessment/reports-and-studies/2023-staar-hybrid-scoring-study.pdf

Trump, D. J. (2025, April 23). *Advancing artificial intelligence education for American youth* [Executive order]. The White House. https://www.whitehouse.gov/presidential-actions/2025/04/advancing-artificial-intelligence-education-for-american-youth/

U.S. Copyright Office. (2025, January 29). *Copyright and artificial intelligence: Part 2 – Copyrightability*. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf

U.S. Department of Education, Office of Educational Technology. (2024). *Designing for Education with Artificial Intelligence: An Essential Guide for Developers.* Washington, D.C. https://tech.ed.gov/designing-for-education-with-artificial-intelligence/

U.S. Department of Education, Office of Elementary and Secondary Education. (2018). *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process.* Washington, D.C. ed.gov/sites/ed/files/2023/11/assessmentpeerreview.pdf

Veeramani, H., Thapa, S., Shankar, N. B., & Alwan, A. (2024). *Large language model-based pipeline for item difficulty and response time estimation for educational assessments. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 561–566). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.49.pdf

Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. *CCSSO Monograph No. 18.*

Wertheim, J, Heck, T., Pruit, S. (2025). *Assessing brilliance: Feedforward Approaches to Next Generation Learners.* Presentation given at the National Forum on the Future of Assessment & Accountability: Dallas, TX. https://drive.google.com/file/d/1dHFNS4qLZDoxZw5-diExzp1fU26o7On4/view

Weissburg, I., Anand, S., Levy, S., & Jeong, H. (2024). *LLMs are biased teachers: Evaluating LLM bias in personalized education. arXiv preprint* arXiv:2410.14012. https://arxiv.org/abs/2410.14012

Whitmer, J., Beiting-Parrish, M., Blankenship, C., Folwer-Dawson, A., & Pitcher, M. (2023). NAEP Math Item Automated Scoring Data Challenge Results: High Accuracy and Potential for Additional Insights. https://doi.org/10.35542/osf.io/eyzgd

Wolandt, K., & Kraus, E. B. (2025). *Detecting differential item functioning (DIF) in multidimensional item response theory (MIRT) models using explainable artificial intelligence (XAI).* Manuscript in preparation. LMU Munich & University of Tübingen. https://osf.io/preprints/psyarxiv/vf8mt_v1

World Health Organization. (2024, February 1). *Assistive technology.* https://www.who.int/news-room/fact-sheets/detail/assistive-technology

Xu, Q., Jiao, H., Zhou, T., Li, M., Zhang, N., Peters, S., & Fu, Y. (2025). *Automated alignment of math items to content standards in large-scale assessments using language models* (arXiv:2510.05129v2). arXiv. https://arxiv.org/abs/2510.05129

Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2021). *Handbook of automated scoring: Theory into practice.* Chapman & Hall/CRC.

Yaneva, V., North, K., Baldwin, P., Ha, L. A., Rezayi, S., Zhou, Y., Ray Choudhury, S., Harik, P., & Clauser, B. (2024). Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Questions. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA 2024) (pp. 470–482). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.39/

Yaneva, V., & von Davier, M. (Eds.). (2023). *Advancing natural language processing in educational assessment.* Routledge. https://doi.org/10.4324/9781003278658

Yildirim-Erbasli, S. N., & Bulut, O. (2023). *Conversation-based assessment: A novel approach to boosting test-taking effort in digital formative assessment.* Computers and Education: Artificial Intelligence, 4, Article 100135. https://doi.org/10.1016/j.caeai.2023.100135

Zapata-Rivera, D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to score reporting research. *Educational Measurement: Issues and Practice, 33*(1), 25–33. https://doi.org/10.1111/emip.12028

Zhao, R., Singh, J., & Li, X. (2024). *The life cycle of large language models: A review of biases in data, algorithms, and deployment for education. arXiv preprint* arXiv:2407.11203. https://arxiv.org/abs/2407.11203

National Center for the Improvement
of Educational Assessment, Inc.
Dover, New Hampshire

www.nciea.org