Stanford | Accelerator for Learning

RESPONSIBLE ASSESSMENT IN THE AI ERA

Key Insights from a Future-Focused Conference

Nneka J. McGee
Candace Thille
Ikkyu Choi
Kadriye Ercikan
Isabelle C. Hau
Matthew S. Johnson (Reviewer)

Table of Contents

Acknowledgments

We would like to express our appreciation to the following individuals for their support and contributions to the Responsible Assessment in the AI Era convening:

Daniel L. Schwartz the I. James Quillen Dean, and Nomellini & Olivier Professor of Educational Technology at Stanford Graduate School of Education, and The Halper Family Faculty Director of the Stanford Accelerator for Learning.

Amit Sevak Chief Executive Officer, ETS.

Candace Thille Associate Professor, Stanford Graduate School of Education, Faculty Director, Adult and Workforce Learning Initiative, Stanford Accelerator for Learning.

Collaborating Organization

The convening was supported by Educational Testing Service (ETS), an organization whose mission focuses on "advancing the science of measurement to power human progress."

Recognition of Contributions

We would also like to extend our gratitude to panelists, moderators, poster session presenters, and breakout session contributors. Their discussions informed the themes, tensions, and future-forward moves reflected throughout this paper.

Panelists

Bryan Brown, Stanford University
Emma Brunskill, Stanford University
Sara Caldwell, Open AI
Tony Chan, Former President, KAUST*
Kristen DiCerbo, Khan Academy
Ben Domingue, Stanford Graduate School of Education*
Maureen Heymans, Google*
Matthew S. Johnson, ETS*
Victor Lee, Stanford Graduate School of Education and Stanford Accelerator for Learning
Lydia Liu, ETS
Temple Lovelace, Assessment for Good,AERDF
Nancy Otero, Gates Foundation
Kadriye Ercikan, ETS
Jesse R. Sparks, ETS*

Moderators

Mille Garcia, California State University
Kevin Gutherie, ITHAKA
Darienne Hudson, United Way, Southeastern MI
Beverly Daniel Tatum, Spelman College

*panelist or presenter and breakout session contributor

**graduate student at Stanford GSE and breakout session contributor

Poster Session Presenters

ETS
Ikkyu Choi*
Patrick Kyllonen*
Matthew S. Johnson*
Sandip Sinharay*
Diego Zapata-Rivera*

Stanford Graduate School of Education
Yunsung Kim**
Hansol Lee**
Eunjung Myoung
Rojas Pino
Marcos Santiago

*panelist or presenter and breakout session contributor
**graduate student at Stanford GSE and breakout session contributor

We thank all attendees who joined and contributed to the spirit of inquiry. Their engagement reminded us that responsible assessment in the AI era is not the work of one organization or discipline, but a shared effort across communities committed to learning.

Authors

Nneka J. McGee, J.D., Ed.D., Project lead and principal co-author for the "Responsible Assessment in the AI Era: Key Insights from a Future-Focused Conference" white paper.

Candace Thille, Ed.D., Associate Professor, Stanford Graduate School of Education, Faculty Director, Adult and Workforce Learning Initiative, Stanford Accelerator for Learning.

Ikkyu Choi, Ph.D., Senior Research Scientist, ETS

Kadriye Erkican, Ph.D., Senior Vice President of Global Research, ETS; President, ETS Canada Inc.

Isabelle C. Hau, Executive Director, Stanford Accelerator for Learning

Recommended Citation: Nneka J. McGee, Candace Thille, Ikkyu Choi, Kadriye Ercikan, and Isabelle C. Hau. "Responsible Assessment in the AI Era: Key Insights from a Future-Focused Convening," Stanford Accelerator for Learning, Stanford University, 2026.

Reviewer: Matthew S. Johnson, Ph.D., Principal Research Scientist, ETS Research Institute

Disclaimer: This paper reflects the perspectives of the authors and includes quotes from the "Responsible Assessment in the AI Era" convening. It does not necessarily represent the views of Stanford University, the Stanford Accelerator for Learning, ETS, participants or collaborating organizations.

A note on AI: This paper was developed and written by the principal authors, who reviewed and fact-checked all content. ChatGPT (GPT 5.5) was used sparingly for minor editorial refinement. Sembly AI was used to transcribe and summarize breakout session discussions.

1. Introduction: Assessment at a Moment of Change

"How do we shape the current moment?" The question, from Stanford Professor and Stanford Accelerator for Learning Faculty Co-Director Candace Thille, framed a broader discussion about how to approach assessment in the era of artificial intelligence (AI). In search of answers, the Stanford Accelerator for Learning, in collaboration and with support from ETS, hosted the "Responsible Assessment in the AI Era" convening on January 29, 2026.

"We want to have conversations that are rich and deep about how research, policy, technology and education come together," Amit Sevak, CEO of ETS, said in his opening remarks. The convening drew approximately 100 leaders from research, technology, K-12 schools, and higher education institutions seeking to advance the science and practice of responsible assessment in the age of AI. There was a consensus among them: Assessment is at a moment of change, and AI is a major driver of that transformation.

Panelists shared wide-ranging perspectives on evidence, rethinking readiness, socioculturally responsible assessment, and building trustworthy AI. Through poster sessions, scholars presented findings from their latest studies. Contributors expressed their viewpoints at breakout sessions on topics ranging from AI policy to extracting construct evidence. While this paper primarily reflects the insights and perspectives shared by speakers at the event, it also draws on relevant research and related work to contextualize the contributors' dialogue. It summarizes key takeaways from the convening and includes a glossary of terms related to assessment.

> *"We want to have conversations that are rich and deep about how research, policy, technology and education come together."*
>
> **— Amit Sevak, CEO of ETS**

1.1 Opening the Lens on Responsible Assessment

The conference took place at a critical juncture for assessment. There are growing calls for assessments to become more socioculturally responsive to learners' cultural, educational, and societal contexts in order to optimize engagement, motivation and performance. Through personalization, AI may help address some of these goals.

At the same time, the integration of AI into education and assessment raises questions about which competencies learning and assessment should prioritize, and how to leverage the potential of AI while minimizing the risks of exacerbating opportunity gaps for individuals from different societal and cultural backgrounds. The conference brought these concepts together by focusing on how assessment can become both responsive and responsible for

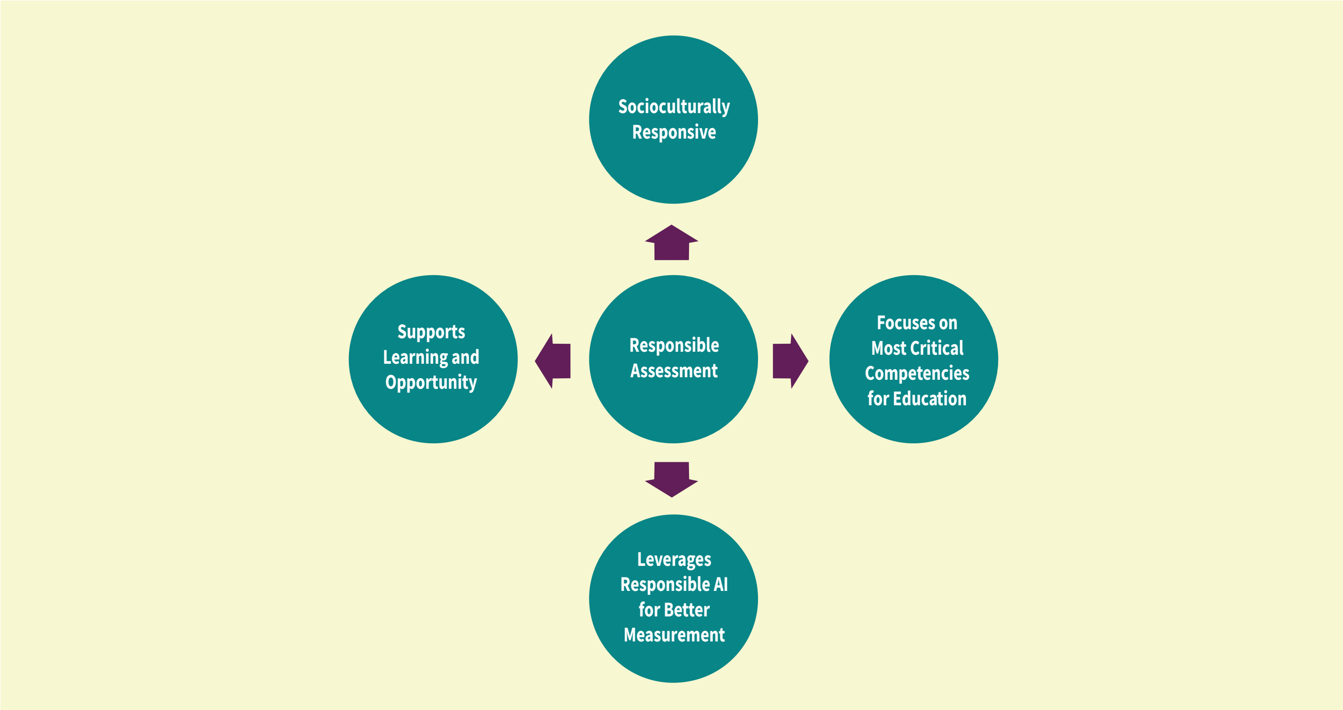

Figure 1: Responsible Assessment

meeting evolving educational goals. This approach promotes responsible AI by supporting valid and fair interpretations and uses of assessment (See Fig. 1).

In this paper, we use the term "responsible assessment in the AI era" to refer to assessment that is grounded in individuals' sociocultural contexts and designed to generate valid, trustworthy, and context-specific inferences from accumulated evidence to support learning and development. Responsible assessment is a call that represents a shift toward continuous, context-rich, and developmentally oriented assessment practices that leverage AI in responsible ways, whose principles and practices are detailed in Johnson (2025).[1] It emphasizes a focus on complex critical competencies, accumulation of multiple sources of evidence through socially situated performance over time, and the consideration of sociocultural context. It also seeks to support learning and growth while prioritizing transparency and interpretability in how evidence is used, elicited, and accumulated.

Videos of key sessions at the event can be watched here: https://acceleratelearning.stanford.edu/conference/responsible-assessment-in-the-ai-era/

2. Why Traditional Assessment is Evolving

Assessment is used to measure knowledge and skills, monitor growth over time, support learning, inform accountability, and guide decisions about admissions and program access.[2] In practice, traditional assessments frequently operate at the intersection of improvement and accountability, where the same measures are expected to support learning while also serving evaluative and comparative functions.[3] This section discusses three interconnected factors shaping the evolution of assessment amid multiple, sometimes competing demands.

The promise of AI depends on how assessment helps educators see and respond to learners' next steps.

2.1 Misalignment with How Learning Actually Occurs

"Traditional assessments may tell us if a student arrived at a correct answer," stated Millie García, Chancellor of the California State University. "But they often tell us little about how [students] got there. In today's collaborative digital and increasingly AI-supported communities, that gap matters." While learning unfolds as a process, assessment often remains largely event-based. This dynamic creates a mismatch between the realities of learning and the way assessment is commonly designed.

This mismatch is most evident when learning and assessment create varied conditions for students. Cindy Mazow, director of learning technology and design at the Stanford Graduate School of Business, emphasized that effective learning requires room for students to make mistakes or take risks because those actions can expose misconceptions or gaps in understanding which can then be corrected. In assessment mode, however, Mazow indicated that students "don't want to make a mistake because they know that there's a consequence to that assessment whether it's high-stakes or low-stakes." The conditions that support learning, then, are not always present when students are being assessed.

Temple Lovelace, executive director of assessment for good at Advanced Education Research & Development Fund (AERDF) described the value of providing space for learners to show up as their full selves in a learning environment. What students bring to learning follows them into assessment. "Full self includes family history, home context, and societal context," explained Kadriye Ercikan, senior vice president of global research at ETS. "This affects how they relate to the assessment…[and] how they engage and make sense of the assessment." Although standardization is often intended to even the playing field, Ercikan suggested that equivalent assessment conditions are lacking. Many current systems still emphasize end products over the pathways that produce them.

AI-supported learning environments may create new opportunities to address this misalignment, but only if they are designed around sound learning principles. Maureen Heymans, vice president of learning at Google, suggested that "learning is most effective when it is active, engaging, and personalized." When connecting learning to Vygotsky's Zone of Proximal Development, Josh Arnold, co-founder of Impacter Pathway, suggested that learners can do "hard and gritty things" when those things are relevant and right-sized to learners' current abilities.[4]

In practice, the promise of AI depends on how assessment helps educators see and respond to learners' next steps. "What do we change on the ground when it comes to assessment and learning?" Lovelace asked, connecting the broader discussion to her experiences as a special education teacher. Her question points toward a future where AI-supported assessment is more integrated with learning. The challenge is to strengthen learner support without turning every moment of learning into another layer of assessment.

2.2 The Increasing Influence of AI

AI affects multiple points of the assessment process, including how student work is produced and how that work is evaluated. As the use of AI increases among learners, there are potential disconnects between what assessments are designed to measure and what they actually capture. The technology also impacts the skills and competencies deemed important at schools and in the workforce.

2.2.a Validity

Generative AI (GenAI) has exposed the limitations of traditional assessments based primarily on final outputs.[5] Learners can now use GenAI to generate high quality products without fully engaging in the learning processes those products are meant to represent. As a result, outputs alone can no longer be assumed to serve as reliable indicators of human capability. In these cases, assessment "risks measuring technological proficiency rather than human skill or understanding"[6] As Thille of the Stanford Accelerator for Learning said, when assessment no longer measures what it was meant to measure, the validity of inferences based on assessment outcomes may be compromised.

AI is also shaping how assessment evidence is scored and interpreted. Ercikan of ETS specified that scoring is the most widely used application of AI in assessment. This creates validity concerns when AI-based scoring, such as automated scoring systems, captures features of performance unrelated to the intended construct. Ercikan offered an example of a science assessment in which AI scoring rewards language competency rather than scientific reasoning to describe construct-irrelevant variance.

When such irrelevant variance systematically advantages or disadvantages particular learners, it becomes a fairness concern and, as Ercikan noted, "is at the core of bias in assessment."[7] Stanford Graduate School of Education Professor Guillermo Salano-Flores has identified three areas where bias can be introduced: language elements (e.g. words), visual components (e.g. charts), and contextual scenarios (e.g. word problems).[8] Bias threatens validity when parts of the assessment are not central to what is being measured.[9]

Contributors shared additional AI-influenced assessment practices that impact validity, including:

- **Construct underrepresentation**: Ercikan explained this can occur when an AI scoring system only partially captures what the assessment is intended to measure.[10]

- **Generalization**: Andrew McEachin, senior research director at ETS, described how AI systems can fail to generalize across contexts and how a human's ability to recognize when AI systems generalize is not well-developed.[11]

- **Training data**: McEachin also expressed concern about AI models that are not trained on data representing authentic child learning, particularly in K–12 education.[12]

- **Calibration**: Matthew S. Johnson, managing director of innovative research at ETS, used the example of rubrics to highlight how what humans and AI systems take into account can differ on the same assessment.[13]

This list of validity threats is not exhaustive, but it shows how AI can influence assessment at multiple points, from the production of student work and the scoring and interpretation of that work.

2.2.b AI literacy and essential competencies

During a panel discussion on measuring essential competencies, Jesse R. Sparks, research director at ETS Research Institute, pointed to AI literacy as an emerging construct. Its rise reflects AI's increasing influence across education and work. Sparks emphasized that redesigned assessments are needed to capture key abilities associated with AI literacy, including collaborative interaction and critical evaluation.[14]

The conversation also moved beyond AI literacy to consider additional competencies learners will need in the age of AI. Sparks, Sara Caldwell, head of GTM Readiness at OpenAI, and Victor Lee, associate professor at the Stanford Graduate School of Education and faculty lead for AI+Education at the Stanford Accelerator for Learning, identified competencies such as systems thinking and deep domain expertise as especially important. Caldwell suggested that people who combine their own competencies with AI will be better positioned in the workforce than those who rely on AI alone.

As AI systems take on more routine cognitive tasks, assessment will need to focus more deliberately on capabilities that are uniquely human.

Together, these examples show that assessment priorities are changing. AI literacy is not sufficient by itself. As AI systems take on more routine cognitive tasks, assessment will need to focus more deliberately on capabilities that are uniquely human. They are often described as "durable skills" and include critical thinking, creativity, curiosity, collaboration, agency, and adaptability.[15] Emotional intelligence, relational intelligence, ethical judgment, and challenging assumptions have also been identified as uniquely human skills.[16] The Stanford Accelerator for Learning examined these shifting priorities in Future-Ready Voices, a November 2025 interview series featuring researchers, educators, and thought leaders.[17]

2.2.c AI and the Workplace

Job availability and workforce displacement remain pressing challenges in the AI era. Citing a McKinsey report, Darienne Hudson, president and CEO of United Way for Southeastern Michigan, predicted that 30% of jobs will become automated within four years.[18] According to the World Economic Forum, employers expect 39% of workers' core skills to change by 2030.[19] Regarding specific professions, Caldwell said, "It's much less about that an engineer won't exist, but what aspects of that role are going to fundamentally change." In other words, assessment should focus on which skills workers will need to succeed in changing work environments.

Sparks envisioned a workplace where "we interact with AI tools in ways that really enhance the strengths that we're bringing to the situation." Caldwell used as an example the ability to orchestrate a set of AI agents. Lydia Liu, associate vice president of ETS, highlighted a collaboration between ETS and OECD named PISA VET (Vocational Education and Training).[20] The international initiative aims to benchmark performance across countries to identify factors that are associated with positive outcomes in vocational education. Some of the AI-related innovations Liu described include:

- **VR glasses** to observe if automotive technician students can perform critical tasks related to the job such as inspections and diagnoses, and maintenance and repair, and

- **Pressure sensing gloves** to determine if trainees can apply the right amount of pressure to complete a task.

Liu's descriptions of immersive technologies reveal a gap between what many traditional assessments measure and the types of skills increasingly valued at work. Applications of new technologies and tools opened the door for more comprehensive and authentic assessment. As expectations change, so do signals of readiness, as evidenced by the proliferation of alternative certifications, skill verifications, and micro-credentials.[21] These signals reflect rising demand for evidence that is more closely tied to demonstrable capability.[22]

2.3 A Growing Gap Between What We Measure and What Matters

"Traditional assessment is misleading," said Daniel L. Schwartz, dean of the Stanford Graduate School of Education and the Halper Family Faculty Director of the Stanford Accelerator for Learning. He described the implications of a study on how students adapt knowledge to a novel situation.[23] "We need instruction that produces adaptive learners and assessment that can tell."

Beneath the immediate concerns about AI and assessment is a more enduring issue expressed by OpenAI's Caldwell: Measurement signals what systems value. That relationship is not neutral. Assessment shapes not only how learning is measured, but also what educational systems prioritize.[24] A large-scale assessment used for comparability and accountability can influence curriculum, instruction, and public judgments of quality.[25]

Over time, systems establish priorities, assessments operationalize them, and repeated use reinforces what counts as evidence. While these approaches enable common metrics across large populations, they can also narrow how and what learning is represented. If educational systems place more value on skills and learning processes, then the data gathered through assessment must broaden to show how capabilities develop and appear in practice.

2.3.a Advances in Data Availability

AI is expanding what can be observed and measured about learning and performance. Throughout the panels and breakout sessions, stealth assessment was discussed as one example of how AI is advancing existing approaches to gathering evidence.
Sparks described stealth assessment as a process methodologically grounded in evidence-centered design (ECD) where educators and researchers "continuously extract signals from a learner's everyday learning environment." Contributors also discussed how systems can draw on process data, ambient data, longitudinal data, and linked systems that connect assessment results with broader patterns of learning and performance.

Patrick Kyllonen, distinguished presidential appointee at ETS, highlighted continuous assessment, which collects data to measure student progress continuously over time. He suggested that this approach may reduce assessment fatigue and test anxiety. Advanced analytic methods are making it possible to interpret patterns from those data in new ways.

However, simply increasing the amount of data is not sufficient. To paraphrase Julia Wilkowski, who directs learning sciences at Google, more evidence does not automatically produce better insight.

Assessments such as stealth assessment and continuous assessment raise concerns about how learner data are collected, interpreted, and protected. Vikas Wadhwani, a leader of the learning and certifications team at Meta, identified privacy as one challenge. Another is learner awareness, since students may change their behavior when they know they are being assessed. Synthetic data and simulated learner models, sometimes described as "simulated students," could offer a way to test assessment designs, study possible response patterns, and examine item performance while reducing reliance on identifiable student data.[26] Simulated students may support faster experimentation, but validation would still be needed to ensure simulated patterns do not distort assumptions about real learners.

2.3.b Measuring Human-Centered Skills

Human-centered skills, such as durable skills, were described briefly earlier in this paper. There is some consensus that human-centered skills can be assessed. However, methods for measuring them remain less established, particularly because they are "interpersonal, intrapersonal, or behavioral in nature."[27]

2.3.b.i. Underdefined meanings for AI literacy and human-centered skills

Stanford Accelerator for Learning's Faculty Lead for AI+Education Victor Lee recognized that AI literacy is one of the core competencies for the AI era, but stresses that "we don't really have a clear articulation for what that is." The AILit Framework, a joint initiative from OECD and the European Commission, includes a comprehensive definition of AI literacy and will contribute to the 2029 PISA Media and Artificial Intelligence Literacy assessment.[28] Even so, definitions of AI literacy continue to vary across recognized institutions and organizations.[29]

A similar issue appears with adaptability. Although widely named as essential in future-focused, AI-rich environments, adaptability lacks a standard definition with operational indicators across AI-use scenarios.[30] These constructs must be carefully defined in operational terms if assessments are to collect meaningful data to measure.[31]

2.3.b.ii. Existing and emerging tools to measure human-centered skills

Contributors characterized the measurement of human-centered skills as part of a larger shift from controlled assessment environments toward more authentic, open-ended interactions. Some of the tools discussed are already established in practice, while others are newer or remain in varying stages of research and development. Here is sampling of the tools referenced during the convening:

- **Cognitive interviews**: Lydia Liu of ETS suggested this format to gain a better understanding of students' cognitive process when interacting with assessment or learning tasks, particularly when considering cultural or behavioral cues.[32]

- **Live demonstrations:** In addition, Liu described this format as the most authentic way to collect evidence about what learners know and can do.

  Wide applications of live demonstrations will require operational efficiency for feasibility.

- **Role-playing**: ETS's Patrick Kyllonen outlined an initiative where students interact with AI as a role-playing partner to measure social skills.[33]

- **Scenario-based assessments**: Kyllonen also explained how AI can be used in assessment environments to generate scenarios that reflect non-traditional, real-world situations.[34]

- **Portfolio-based assessments:** Contributors discussed evaluations that allow learners to demonstrate growth and real-world application of knowledge through a curated collection of work over time, rather than through a single test or exam.

- **Conversation-based assessments**: Diego Zapata-Rivera, distinguished president appointee at the ETS Research Institute, described an evidence-centered design (ECD) based approach that uses AI agents to help learners demonstrate knowledge and skills through dialogue.[35]

- **Virtual robotics programming tasks**: Sparks, of ETS, detailed emerging work that examines the potential for identifying signals of persistence and resilience from learners' interactions with virtual robots in block-based coding environments.[36]

- **Game-based assessment in naturalistic settings**: Guess What?, a smartphone app created by Stanford Professor and Stanford Accelerator for Learning faculty affiliate Dennis Wall, uses game-based interactive computer vision to identify early signs of autism in naturalistic settings.[37]

3. What System Change Requires: Rethinking Assessment

System change presents a challenge for how assessment can support the complexities of modern learning. Learning is continuous, adaptive, contextual, and increasingly shaped by human and technological interaction. That complexity makes it difficult to rely on assessment models designed primarily to capture performance at a single point in time. There are multiple pathways to rethinking assessment, and AI will play an important role in expanding the evidence available to guide interpretations, noted Karin Forssell, senior lecturer with the Stanford Graduate School of Education and director of the AI Tinkery at the Stanford Accelerator for Learning.

3.1 Testing Events to Systems of Inference

Summative assessment is designed to evaluate learning at the culmination of an instruction period. Despite its importance, ETS' Sparks highlighted a limitation: the results of some summative assessments arrive too late to be useful to learners and educators. Johnson, also of ETS, said motivation is another limiting factor because learners question the reasons for summative assessments for which they may never see the results.

A different orientation places learner development at the center of assessment. This orientation focuses on formative assessment that provides ongoing feedback rather than a snapshot of performance and draws on evidence that can illuminate "the course of development and the status of skills."[38] Once the attention on assessment shifts from summative assessment to learner development, the process of gathering and utilizing evidence must be reconsidered.

AI may increase the feasibility of formative assessment by synthesizing evidence and identifying patterns within large and complex data streams, which in turn can contribute to more refined inferences about learning. Zapata-Rivera used the principles of ECD and the Autotutor dialogue framework to describe how AI agents could support this type of assessment. "In evidence centered design, each piece of each task is designed for a particular purpose," explained Zapata-Rivera. "With AI, we could create expert agents to generate questions or create opportunities even if we are not there."

3.2 Building Infrastructure and Validity for AI-Powered Assessment

Christian Pinedo, vice president of external affairs and advocacy at aiEDU, emphasized that building comprehensive infrastructure "requires a lot of human investment." Systems that incorporate AI, including data architectures and interoperable platforms, must have the infrastructure to support the collection, integration, interpretation, and storage of large amounts of data. In response, LeeAnn Lindsey, director of edtech and innovation at Northern Arizona University's Arizona Institute for Education and the Economy, added that many state and district budgets are stretched while acknowledging the "lack of basic infrastructure for implementing good decision-making through assessment." Their exchange points to the need for sustained investment in assessment infrastructure, even in constrained fiscal environments. Without such infrastructure, even the most advanced AI-supported assessment cannot function effectively.

The value of robust digital data systems depends on whether the claims, interpretations, and decisions they support remain valid.[39] Therefore, investment in assessment infrastructure should be paired with investment in collecting ongoing validity evidence in AI-mediated environments. For example, large language model (LLM) scoring of constructed responses may require more extensive validity evidence than traditional approaches due to concerns about consistency and interpretability.[40] In addition, using LLMs as opposed to traditional scoring methods could be less reliable and more costly due, in part, to the expense involved in curating validity evidence.[41]

Without infrastructure, even the most advanced AI-supported assessment cannot function effectively.

3.3 Expanding What Counts as Evidence

The transformation of assessment requires a broader evidentiary base. For example, relational learning builds on Vygotsky's social constructivism by emphasizing strong teacher-student relationships in classroom settings.[42] In a panel on rethinking measurement, Liu of ETS indicated that traditional assessment systems were not designed to capture these dimensions of learning well.[43] Social and relational learning often involve complex skills that have cognitive, affective, and behavioral components, but also depend on capacities for dialogue, interactions, building trust, and collective problem-solving. A future-oriented taxonomy from Liu and her colleagues illustrate this breadth by outlining 30 skills ranging from adaptability to transformative competencies.[44] Liu also emphasized the importance of gathering evidence from multiple sources and creating an inclusive system for students to demonstrate their skills. Leadership, for instance, may be shown not only from being a captain on varsity sports teams, but also from taking care of younger siblings or leading community activities.

Emerging research suggests that a broader evidentiary base would be beneficial in measuring social learning, even if the work remains preliminary. Positive outcomes were observed when students engage in inquiry and constructive collaboration suggest that social processes can be documented and interpreted in analytically useful ways.[45] Contributors also discussed developing storylines about future-focused scenarios, such as Participatory Scenario Planning, which can be used to enable the social learning process.[46] Researchers acknowledge limitations in scope, but they point toward a potentially important direction for assessment by expanding the evidentiary base to include social and relational learning.

3.4 Extending the Role of Formative Assessment

Formative assessment embeds assessment in instruction to guide next steps. Teaching and learning must be interactive, with teachers responding to student progress and difficulties as they adapt instruction.[47] Contributors noted that this feedback loop can also include peer formative assessment, in which students exchange feedback and compare it with AI-generated feedback.[48]

In a panel discussion on the potential of AI to help or harm socioculturally responsible assessment, Bryan A. Brown, professor of science education at Stanford Graduate School of Education, emphasized that formative assessment should meet students where they are. To support that goal, Brown is working on Multicultural Oriented Science Adaptive Intelligence for Classrooms (MOSAIC), a culturally responsive AI-powered app that connects science to students' lived experiences.[49] Contributors also discussed a five-step framework for integrating AI into formative assessment that builds on earlier work related to digital technologies.[50]

Even researchers advocating for formative assessment acknowledge the risks of AI use. AI can exhibit bias resulting in outcomes that minimize or exclude learners from the learning process.[51] The technology still generates inaccurate information, which can skew evidence for learning.[52] Ben Domingue, associate professor at Stanford Graduate School of Education and faculty affiliate at the Stanford Accelerator for Learning, cautioned that passive data collection in some types of formative assessment raises increasingly difficult ethical, privacy, and transparency questions. "The more data we're able to collect passively," Domingue said, "the harder those questions are going to become."

A stronger role for formative assessment does not diminish the need for summative assessment. The goal is to ensure that each serves a clear and complementary purpose.[53] Instead of one assessment having the final judgment on learners' success, multiple assessments can be designed to collect and synthesize accumulated evidence of learning, development, and capability over time.[54]

4. Advancing a Focus on Responsible Assessment

The work ahead includes reflecting on what assessment should make possible for learners. Despite advances in AI, the foundations of assessment still rest on why it exists and who it is meant to serve. This section invites readers to explore assessment design and implementation in the AI era.

Responsible assessment should be co-designed with the people it affects and interpreted with careful attention to context.

4.1 Reflecting on Responsible Assessment

In keeping with the theme of the convening, panelists were asked to share their thoughts on the concept of responsible assessment. Their responses suggested that responsibility begins with the people assessment is meant to serve. Several considerations emerged from their discussions.

Centering humans: For Sparks of ETS, responsible assessment includes "humanizing assessment" in ways that support individuals in growing and developing. When discussing AI tools, Sparks recommended keeping humans in the loop to ensure technical quality, to minimize harm, and to maximize assessment value.

Opening participation: Stanford Accelerator for Learning's Lee stressed the importance of co-designing and co-developing assessments with diverse and representative stakeholders. As Lee put it, "There's amazing brilliance that exists in so many communities that may not always appear in assessments." He added that two of the field's strongest tools are "to open participation and to be inquisitive."

Recognizing uniqueness: When considering socioculturally responsible assessment, Nancy Otero, senior program officer for AI assessment at the Gates Foundation, asked how we could measure what makes learners unique and how AI could help capture that uniqueness. Lovelace, from AERDF, described success as learners becoming self-regulated enough to chart their own learning pathways. Lovelace also noted AI's potential in helping measure what learners are experiencing in the moment so that learning can continue.

Contextualizing design: "Context and content matter," Brown emphasized, including "where young people have learned particular experiences." He described an assessment that can point to where students know what they know as a source of empowerment. There is potential for AI-powered personalized assessments that align tasks with an individual's cultural, linguistic, and social context as well as learner preferences.

These comments suggest that responsible assessment should be co-designed with the people it affects and interpreted with careful attention to context.

4.2 Establishing Trust through Transparency

As AI becomes more embedded in assessment design and delivery, trust is essential for its meaningful use. Contributors emphasized that full technical visibility is not always necessary, but transparency helps establish the conditions under which people can responsibly rely on a system.

Clarifying trust: "The reason we often want transparency is really an issue of trust," noted Emma Brunskill, professor of computer science at Stanford and faculty affiliate of the Stanford Accelerator for Learning. For Brunskill, trust depends on what a system can reliably do, what it is intended to support, and where it may fall short. She used the example of a microwave to show that users do not need to understand every inner mechanism to trust a tool, but they do need to have faith there is a broader system which provides safeguards.

Demonstrating value: Kristen DiCerbo, chief learning officer at Khan Academy, connected trust with usefulness, suggesting that systems should "give educators what they don't have." Stanford's Domingue expressed hope for AI-supported systems that can connect what someone is learning with a comprehensive portfolio of work. Their comments imply that trust is built when a system provides tangible benefits.

Documenting quality: Technical documentation offers another way to support trust. ETS researchers Johnson and McEachin pointed to the technical documentation ETS provides for automated scoring as an example of how organizations can make relevant information available. Heymans of Google described adversarial testing to identify issues such as bias in generative AI models, along with efficacy studies, as additional ways to build trust in the system.

Protecting integrity: The discussion also made clear that transparency has limits. ETS's Johnson stressed that too much transparency can create vulnerabilities. Protecting some aspects of a system's mechanics may be necessary to preserve the trustworthiness of the final result. He also cautioned that trust can be undermined when AI models attempt to extrapolate data for small cultural groups they weren't trained on, leading to outputs that don't make sense for those communities.

Offering solutions: Data represent another area where trust must be earned. ETS researcher Zapata-Rivera suggested a framework in which users know exactly why their data are used and can remove their data from particular uses. McEachin and Tony Chan, former president of King Abdullah University of Science of Technology, added that explainable AI, which shows what inputs influenced a model's outputs, can support transparency by giving users some basic insights into AI-supported decision-making.[55]

4.3 Translating Insight into Action

The Responsible Assessment in the AI Era convening made clear that AI is already reshaping assessment. Forging a path forward will require coordinated action across education systems, research organizations, developers and funders. Each group has a distinct role, but the work is interconnected. The following subsections translate the insights from the convening into concrete next steps.

4.3.a For Education Systems: Pilot Innovative Models of Assessment

Pilots and experimental programs can give educational systems a manageable way to explore innovative initiatives such as competency-based learning or novel formative assessment practices before wider implementation. These efforts might begin in one course, grade level, or program. Training is built in from the start as well as clear plans for explaining results to learners and families.

Stanford Graduate School of Education Associate Professor and Stanford Accelerator for Learning faculty affiliate Jason Yeatman leads the Rapid Online Assessment of Reading (ROAR), which partners with school districts to pilot the online platform. Built on the premise that traditional assessments that assess foundational reading skills can be costly and time-consuming, ROAR offers a free, online test that provides "precise indices of reading ability."[56] ROAR shows that schools don't always have to partner with edtech companies for pilots. Research and assessment organizations are also seeking education partners to test emerging approaches in real learning environments.

Areas for action:

- **Pilot continuous, embedded assessment models**
  - Integrate assessment into daily learning activities
  - Use formative feedback loops to guide instruction
- **Adopt portfolio and competency-based approaches**
  - Capture evidence of learning across time and contexts
  - Include artifacts, projects, and collaborative work
- **Reduce reliance on high-stakes, one-time testing**
  - Rebalance summative assessments within broader systems
  - Explore multiple measures for accountability
- **Build educator capacity**
  - Train educators to interpret complex evidence
  - Provide tools for integrating assessment into instruction
- **Engage learners and families**
  - Increase transparency of assessment systems
  - Provide accessible interpretations of results

4.3.b For Researchers: Redefine What and How We Measure

Researchers have a critical role in the future of assessment. The conference poster session highlighted work underway, while also showing the magnitude of what still needs to be clarified. The Measuring Acquisition & Growth of Inquiry & Curiosity (MAGIC) Project, led by Stanford Graduate School of Education Professor and Stanford Accelerator for Learning faculty affiliate Jelena Obradović, is an example of sustained research, with philanthropic support and research practice partnerships, focused on developing performance-based assessments that measure curiosity, creative thinking, problem-solving, and scientific inquiry in young learners.[57] As Sandip Sinharay of ETS noted, research initiatives can stall when funding ends, underscoring the need for deliberate efforts to continue them, including district research fellows, alternative funding sources, and sustainable cross-sector partnerships.

Areas for action:

- **Clarify and operationalize complex constructs**
  - Develop shared definitions for constructs such as AI literacy, sensemaking, and collaboration
  - Align constructs with real-world performance and domain specificity
- **Advance ecological validity**
  - Design assessments embedded in authentic tasks and environments
  - Study performance across contexts rather than isolated conditions
- **Develop standards for AI-assisted assessment**
  - Establish validity frameworks for AI-generated scoring and feedback
  - Investigate bias, transparency, and reliability of AI systems
- **Leverage new forms of evidence responsibly**
  - Incorporate process data, interaction data, and longitudinal patterns
  - Define limits of interpretation and avoid overreach

4.3.c For Developers: Build for Learning, Not Just Efficiency

Assessment developers should design AI-supported assessments around learning, interpretability, and human judgment rather than speed or scale alone. AI can support feedback and identify patterns that affect learner achievement. Despite this feature, assessment developers should avoid creating products for fully automated high-stakes decision-making, especially when results influence access and opportunity.

Otero, from the Gates Foundation, emphasized assessment developers' responsibility to support scale while making model behavior and guardrails more transparent. Stanford Graduate School of Education Professor and Stanford Accelerator for Learning faculty affiliate Bryan Brown cautioned that developers can make broad assumptions about users without adequate context: "Cultural cues and modes of communication with AI are a direct reflection of who is designing software." Contributors, including Otero and Lee, stressed the importance of involving learners and families so assessment tools better reflect what young people want to understand about themselves.

Areas for action:

- **Design for interpretability and transparency**
  - Make AI-generated feedback explainable
  - Surface uncertainty and limitations clearly
- **Support learning processes, not just outputs**
  - Capture process data (e.g., revision, iteration, collaboration)
  - Provide feedback that supports growth
- **Enable human-in-the-loop systems**
  - Design workflows that incorporate educator judgment
  - Avoid fully automated high-stakes decision-making
- **Prioritize equity and bias mitigation**
  - Test systems across diverse populations
  - Monitor for unintended consequences
- **Align tools with educational goals**
  - Start with clearly defined learning goals
  - Match use of AI tools with standards

4.3.d For Funders: Invest in the Next Generation of Assessment

Funders can accelerate the field by investing in new constructs, stronger assessment, validation studies, and the infrastructure needed to implement promising programs. Otero pointed to questions funders are asking, including what counts as good evidence and whether assessments are capturing what they are intended to capture. Investment is especially important in the AI era because new tools may scale more quickly than the evidence base supporting them.

Contributors also emphasized that funding models can shape how innovation reaches schools and learners. Kenneth Koedinger, a professor at Carnegie Mellon University described research-practice partnerships as strategic tools for driving innovation while aiEDU's Christian Pinedo raised the possibility of human capital fellowships that fund people rather than projects. Others discussed intermediaries that help high-wealth individuals direct capital toward credible, high-impact efforts. Private-sector investment in AI learning tools and certifications suggests that market activity is already moving quickly. Funders can build upon existing momentum through joint RFPs, coordinated projects, and partnerships.

Areas for action:

- **Invest in new constructs and measurement models**
  - Fund the development and validation of assessments for complex capabilities such as AI literacy, adaptability, and relational capacity
  - Support interdisciplinary research teams (learning science, AI, psychometrics, sociology)
- **Build shared infrastructure**
  - Invest in longitudinal data systems and interoperable platforms
  - Support synthetic data environments to enable safe, scalable research
- **Fund translation, not just discovery**
  - Require clear pathways from research to implementation
  - Support partnerships between researchers, schools, and developers
- **Create coordinated funding mechanisms**
  - Launch joint RFPs across funders
  - Prioritize projects that integrate research, practice, and technology

Conclusion

This paper summarizes the discussions and contributions from leaders in research, technology, K-12 schools, and higher education institutions. The sections presented reflect the depth of insights shared during the convening. Contributors emphasized that assessment must evolve without losing sight of fairness and validity. They also called for clearer definitions of emerging constructs, strong evidence for measurement approaches, and more attention to the social and cultural contexts that can influence how learners demonstrate what they know and can do.

Several tensions introduced during the convening will need continued attention. AI can help make more forms of evidence available, but more data does not automatically lead to better assessment. Continuous assessment may reduce reliance on isolated testing events, but there are also concerns about privacy, learner awareness, and overinterpretation. High-stakes assessments are still important, yet people can still benefit from assessments that can show how learners progress over time instead of during a single point. Addressing these issues will require coordinated work across the assessment ecosystem.

The convening demonstrated that dialogue is a strong entry point. Gathering individuals from varied backgrounds and perspectives brought shared concerns, practical questions, and areas for action into focus. Education systems can pilot models in low-stakes settings to determine whether those models produce useful evidence. Researchers can help clarify emerging constructs and strengthen the validity of new forms of measurement. Developers can co-design tools that support responsible use. Funders can support the infrastructure and partnerships needed to move promising work into practice.

This paper reflects the insights and questions that emerged from one convening, therefore it offers a focused view rather than a complete account of the field. The discussion does not capture every advance in AI-supported assessment, but it highlights areas where research, tools, and partnerships are already beginning to take shape. In addition, key terms related to assessment are still being defined. These limitations point to the value of continued dialogue and research as assessment continues to evolve.

The question that opened the convening remains the right one: “How do we shape the current moment?” The answer depends on whether assessment evolves to meet the needs of learners in a changing world. Traditional assessment still has an essential role in the AI era. At the same time, assessment must capture how people now learn and work. AI can expand what is feasible, but it cannot be held accountable for the consequences of assessment decisions. People can. Ultimately, AI creates new possibilities for assessment, but humans must decide which possibilities are worth pursuing.

“We need instruction that produces adaptive learners and assessment that can tell.”

— Daniel L. Schwartz, dean of Stanford Graduate School of Education and Halper Family Faculty Director of the Stanford Accelerator for Learning.

Glossary[58]

Ambient Data: Data collected passively and unobtrusively from the surrounding environment, often through sensors and other digital devices.

Assessment: A systematic process for collecting information in order to draw conclusions about entities of interest, such as people or programs.

Capability: a multidimensional construct that incorporates both stable traits and dynamic conditions.

Construct: The underlying knowledge, skill, ability, or characteristic that an assessment is intended to measure.

Evidence-centered design: A framework for designing assessments that makes explicit the relationships among intended constructs and claims, evidence needed to support the claims, and tasks used to elicit that evidence.

Fairness: The extent to which an assessment supports valid interpretations for all groups of test takers.

Formative Assessment: An assessment process that prioritizes providing timely feedback during learning, with the goal of improving instruction and supporting students' progress toward intended learning outcomes.

General purpose AI: Artificial intelligence systems designed to perform a wide range of tasks, often involving multiple types of input and output data (e.g., audio, images, text), rather than a specific set of pre-set functions.

Generative AI: A subset of artificial intelligence systems whose primary task is to generate new content, such as digital audio, code, images, and text, by learning patterns from large datasets.

High-stakes test: A test whose results can lead to significant consequences for individuals, programs, or institutions, such as certification, admission, or resource allocation.

Inference: The process of reasoning that connects existing knowledge with new observations to support an explanation, conclusion, or prediction.

Longitudinal data: Data collected from the same subjects repeatedly over time.

Low-stakes test: A test whose results carry limited consequences, typically for instructional feedback, practice, monitoring progress, or research purposes.

Measurement: A systematic process of assigning numbers to characteristics of people, objects, or events according to defined rules.

Operationalize: To define an abstract concept in concrete and observable terms.

Process data: Data that record how individuals work through tasks, typically including the actions and steps they take as well as their associated time and duration information.

Reliability: The degree to which assessment results are consistent across repeated measurement conditions.

Standardization: The process of ensuring uniform procedures, materials, and conditions in administering and scoring an assessment such that results are comparable across test takers and across administrations.

Summative assessment: An assessment, typically conducted at the completion of a learning period, to evaluate what students have learned and to summarize their achievement relative to intended outcomes.

Test: A systematic process in which a structured sample of a person's behavior or performance is collected according to a set of prescribed policies.

Validity: The degree to which evidence and theory support the interpretation and use of assessment results.

Endnotes

1. Johnson, M. S. (2025). Responsible AI for Measurement and Learning: Principles and Practices. (Research Report No. RR-25-03). ETS Research Institute. https://www.ets.org/Media/Research/pdf/RR-25-03.pdf

2. Kyllonen, P. and Sevak, A. (2024). Charting the Future of Assessments. ETS Research Institute. https://www.ets.org/Rebrand/pdf/FoA_Full_Report.pdf

3. Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K., & Thissen, D. Augmented Scores - "Borrowing Strength" to Compute Scores Based on Small Numbers of Items. *Test Scoring*, D. Thissen & H. Wainer (Eds.), pp. 355–400. Routledge, 2001

4. The Zone of Proximal Development, developed by Lev Vygotsky, refers to the space between what a learner can do independently and what they can do with guidance. It emphasizes that learning often develops through supported interaction before becoming independent performance. See Vygosky, L.S. (1978). *Mind in society: The Development of Higher Psychological Processes.* Harvard University Press.

5. Organisation for Economic Co-operation and Development. (2026). *Exploring Effective Uses of Generative AI in Education*.(OECD Digital Education Outlook 2026) OECD. https://doi.org/10.1787/062a7394-en

6. See page 53, Organisation for Economic Co-operation and Development. (2026).*Exploring Effective Uses of Generative AI in Education.* (OECD Digital Education Outlook 2026), OECD. https://doi.org/10.1787/062a7394-en

7. See also Ahmanjavaheri, Z., & Zeraatpishe, M. (2020). The Impact of Construct-Irrelevant Factors on the Validity of Reading Comprehension Tests. *International Journal of Language Testing*, *10* (1), 1–10. https://files.eric.ed.gov/fulltext/EJ1291152.pdf

8. Solano-Flores, G. (2025, April 3). Uncovering Bias in Testing. ASCD. https://www.ascd.org/blogs/uncovering-bias-in-testing

9. Kim, K.H., & Zabelina, D. (2015). Cultural Bias in Assessment: Can Creativity Assessment Help?*International Journal of Critical Pedagogy*, *6* (2), 129–148. https://janeway.uncpress.org/ijcp/article/724/galley/1317/view/

10. Chen, Y.H., Li, I.Y., & Chason, W. (2018). Construct underrepresentation. B.B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications. https://methods.sagepub.com/ency/edvol/sage-encyclopedia-of-educational-research-measurement-evaluation/chpt/construct-underrepresentation#_

11. Kaldaras, L., Akaeze, H.O., & Reckase, M.D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education*, 9, Article 1399377 https://doi.org/10.3389/feduc.2024.1399377

12. Kuchemann, S., Avila, K.E., Dinc, Y., Hortmann, C., Revenga, N., Ruf, V., Stausberg, N., Steinert, S., Fischer, F., Fischer, M., , Kasneci, E., Kasneci, G., Kuhr, T., Kutyniok, G., Malone, S., Sailer, M., Schmidt, A., Stadler, M.,Weller, J & Kuhn, J. (2025, February 25). On opportunities and challenges of large multimodal foundation models in education. *npj Science of Learning*, *10*, Article 11. (https://doi.org/10.1038/s41539-025-00301-w

13. Karg, R. (,2025, July 16). Inter-Rater Reliability and Calibration: Ensuring Consistency, Validity, and Scientific Rigor in Mental Health Research. SCID Institute.https://scidinstitute.com/inter-rater-reliability-and-calibration-ensuring-consistency-validity-and-scientific-rigor-in-mental-health-research/; See also McCaffrey, D.F., Casabianca, J.M., Ricker-Pedley, K.L., Lawless, R., & Wendler, C. (2022). Best Practices for Constructed-Response Scoring. Research Report No. RR-22-17. ETS Research Institute. https://www.tr.ets.org/pdfs/about/cr_best_practices.pdf

14. For a comprehensive discussion about assessing actual interactions or applications and not just decontextualized knowledge, see Sparks, J.R., Ober, T.M., Tenison, C., Arslan, B., Roll, I., Deane, P., Zapata-Rivera, D., Gooch, R., & O'Reilly, T. (2024). Opportunities and challenges for assessing digital and AI literacies. ETS Research Institute. https://www.ets.org/pdfs/rd/ets-digital-literacy-ai-full-report.pdf

15. Kim, A-.S., Jones, K., Lang, G., Raider, H.J., Sheikh, A., & Simione, K. (2025). Beyond Technical Skills: Uncovering Durable Competencies Through Multi-Level Stakeholder Analysis. *Information Systems Education Journal*. *23(*6), 4–18, Nov. 2025. https://doi.org/10.62273/IKXH4521

16. Stobierski, T. (2026, February 6). "4 Important Human Skills AI Can't Replace." *Harvard Business School Online*. https://online.hbs.edu/blog/post/human-skills-ai-cant-replace

17. Stanford Accelerator for Learning. (2025). *Future-Ready Voices. (*Accelerate EdTech Impact Summit 2025), Stanford University. https://acceleratelearning.stanford.edu/app/uploads/2025/12/Future-Ready-Voices-Brochure_SAL.pdf

18. Hazan, E., Madgavkar, A., Chui, M., Smit, S., Maor, D., Singh Dandona, G., & Huyghues-Despointes, R. (2024, May 21). A new future of work: The race to deploy AI and raise skills in Europe and beyond. *McKinsey Global Institute*. https://www.mckinsey.com/mgi/our-research/a-new-future-of-work-the-race-to-deploy-ai-and-raise-skills-in-europe-and-beyond#/

19. World Economic Forum. (2025, January 7). *3. Skills outlook* (The Future of Jobs Report 2025) https://www.weforum.org/publications/the-future-of-jobs-report-2025/in-full/3-skills-outlook/

20. Organisation for Economic Co-operation and Development. (n.d.). *International Vocational Education and Training Assessment* (PISA-VET). OECD. https://www.oecd.org/en/about/projects/international-vocational-education-and-training-assessment-pisa-vet.html

21. World Economic Forum. (2025). (The Future of Jobs Report 2025) https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

22. Organisation for Economic Co-operation and Development. (2025). *Empowering the Workforce in the Context of a Skills-First Approach*. OECD. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/empowering-the-workforce-in-the-context-of-a-skills-first-approach_0e3be363/345b6528-en.pdf

23. Schwartz, D.L. (2025). Achieving an adaptive learner. *Educational Psychologist*, *60* (1), 7–22. https://doi.org/10.1080/00461520.2024.2397389

24. Ewell, P.T. (2022). An emerging scholarship: A brief history of assessment. *Building a scholarship of assessment*. Jossey-Bass.

25. Levy-Feldman, I. (2025). The Role of Assessment in Improving Education and Promoting Educational Equity. *Education Sciences*, *15*(2), 224. https://doi.org/10.3390/educsci15020224

26. Shaughnessy, M., Boerst, T., and Pynes, D. (2026). Leveraging simulated student interactions to learn about prospective teachers' formative assessment practice. *The Journal of Mathematical Behavior*, *81*, Article 101298. https://doi.org/10.1016/j.jmathb.2025.101298

27. Kim, A-.S., Jones, K., Lang, G., Raider, H.J., Sheikh, A., & Simione, K. (2025). Beyond Technical Skills: Uncovering Durable Competencies Through Multi-Level Stakeholder Analysis. *Information Systems Education Journal*. *23 (*6), 4–18. https://doi.org/10.62273/IKXH4521

28. OECD / European Union (2026, June 18). *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education.* OECD Publishing, Paris. https://doi.org/10.1787/65cd27d4-en

29. To compare definitions of AI literacy, see U.S. Department of Labor, Employment and Training Administration. (2026, February 13). The U.S. Department of Labor's artificial intelligence literacy framework (Training and Employment Notice No. 07-25) https://www.dol.gov/sites/dolgov/files/ETA/advisories/TEN/2025/TEN%2007-25/TEN%2007-25%20%28complete%20document%29.pdf; OECD / European Union (2026, June 18). *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education.* OECD Publishing, Paris. https://doi.org/10.1787/65cd27d4-en; Miao, F., Shiohira, K., & Lao, N. (2024). *AI competency framework for students*. UNESCO. https://doi.org/10.54675/JKJB9835; Digital Promise (n.d.). *AI literacy*. https://digitalpromise.org/initiative/artificial-intelligence-in-education/ai-literacy/

30. Kizelcec, R.F., Huber, E., Papanastasiou, E.C., Cram, A., Makridis, C.A., Smolansky, A., Zeivots, S., & Raduescu, C. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Computers and Education: Artificial Intelligence*, *7*, Article 100269. https://doi.org/10.1016/j.caeai.2024.100269

31. See page 79,Foster, N., and M. Piacentini (Eds.). *Innovating Assessments to Measure and Support Complex Skills*, OECD Publishing, Paris, 2023. https://doi.org/10.1787/e5f3e341-en.

32. In assessment, cognitive interviews are traditionally used to examine how people interpret and respond to questions as part of item development and validation. For an overview of cognitive interviews and a comparison to think aloud interviews, see Ketterlin-Geller, L.R., Haider, M.Q., & McMurrer, J. (2024). Using Evidence from Task-Based Interviews for Development and Validation of Classroom Assessments: Approaches and Applications in Early Grades Mathematics. *Educational Assessment*, *29* (4), 229–250. https://doi.org/10.1080/10627197.2024.2398421

33. For another application of AI-powered role play, see Chen,A., Jia,J., Li,Y. & Fu.L. (2025). Investigating the Effect of Role-Play Activity with GenAI Agent on EFL Students' Speaking Performance. *Journal of Educational Computing Research*, *63*(1), 99–125. https://doi.org/10.1177/07356331241299058

34. See Holdsworth,S., Sandri,O., Hayes,J. & Maslen,S. (2025). Scenario-based capability assessment in professional learning contexts. *Journal of Adult and Continuing Education*, *31*(3), 600–627. https://doi.org/10.1177/14779714241297069 for an example of scenario-based assessment in an adult learning context.

35. See Zapata-Rivera, D., Forsyth, C., Graf, E.A., & Jiang, Y. (2024). Designing and evaluating evidence-centered design-based conversations for assessment with LLMs. *Leveraging large language models for next generation education technologies*. (CUER Workshop Proceedings). https://par.nsf.gov/servlets/purl/10591735; See also Jackson, G.T., & Zapata-Rivera, D. (2015, October). *Conversation-based assessment* (R&D Connections No. 25). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RD_Connections_25.pdf

36. Sparks, J.R., Lehman, B., Gladstone, J.R., Zhang, S., Schroeder, N.L., & Israel, M. (2025). Measuring persistence and academic resilience of K-12 students: Systematic review and operational definitions. *Frontiers in Education*, *10*, Article 1673500. https://doi.org/10.3389/feduc.2025.1673500

37. Stanford University. (n.d.). Guess What? https://guesswhat.stanford.edu/

38. Kyllonen, P. and Sevak, A. (2024, April 10). Charting the Future of Assessments. ETS Research Institute.. https://www.ets.org/Rebrand/pdf/FoA_Full_Report.pdf

39. Fu, Q., Nicholson, G. L., & Easton, J. M. (2024). Understanding data quality in a data-driven industry context: Insights from the fundamentals.*Journal of Industrial Information Integration*, *42*. https://doi.org/10.1016/j.jii.2024.100729

40. Casabianca, J. M., McCaffrey, D. F., Johnson, M. S., Alper, N., & Zubenko, V. (2026). From Feature-Based Models to Generative AI: Validity Evidence for Constructed Response Scoring.https://doi.org/10.48550/arXiv.2603.19280

41. Ibid.

42. Garcia, J.S. 2022. Relational Learning: Creating a "working alliance" in the classroom. Westfall-Rudd, D., Vengrin, C., & Elliott-Engel, J. (Eds.), *Teaching in the University: Learning from Graduate Students and Early-Career Faculty.* Blacksburg: Virginia Tech College of Agriculture and Life Sciences, Chapter 1. https://doi.org/10.21061/universityteaching; not to be confused with meanings in machine learning and psychology.

43. Liu, O. L., Liu, L., Sherer, D., & LeMahieu, P. (2025). A skills-based vision for assessment, insight, and educational improvement. E. M. Tucker, E. L. Baker, H. T. Everson, & E. W. Gordon (Eds.). *Handbook for Assessment in the Service of Learning, Volume III*. Examples of assessment in the service of learning." University of Massachusetts Amherst Libraries.

44. Liu, O. L., Kell, H. J., Liu, L., Ling, G., Wang, Y., Wylie, C., Sevak, A., Sherer, D., LeMahieu, P., & Knowles, T. (2023). A New Vision for Skills-Based Assessment. Princeton, NJ: Educational Testing Service. https://www.ets.org/pdfs/rd/new-vision-skills-based-assessment.pdf

45. Yin, S., Liu, Z., Goh, D.L., Quek, C., & Chen, N. (2025). Scaling up collaborative dialogue analysis: An AI-driven approach to understanding dialogue patterns in computational thinking education. In *Proceedings to the 15th International Learning Analytics and Knowledge Conference*, 47–57. Association for Computing Machinery. https://doi.org/10.1145/3706468.3706474

46. Poskitt, S., Waylen, K.A., & Ainslie, A. (2021). Applying pedagogical theories to understand learning in participatory scenario planning. *Futures*, *128*, Article 102710. https://www.sciencedirect.com/science/article/pii/S0016328721000197

47. Black, P., & Wiliam, D. (1998). Inside the Black Box: Raising Standards through Classroom Assessment. *The Phi Delta Kappan*, *80* (2), 139–148.

48. Sperber, L., MacArthur, M., Minnillo, S., Stillman, N., & Whithaus, C. (2025). Peer and AI Review + Reflection (PAIRR): A human-centered approach to formative assessment. *Computers and Composition*, *76*. https://doi.org/10.1016/j.compcom.2025.102921

49. Brown, B.A., (n.d.). Introduction to MOSAIC. [Video]. YouTube. https://youtu.be/b-VXVK27Okk

50. "Webb, M., Galamba, A. (2026, February 12) Transforming pedagogy with GenAI-supported formative assessment: Challenges for teacher education. *British Journal of Educational Technology.* https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.70051

51. Johnson, M.S. (2025). Responsible AI for Measurement and Learning: Principles and Practices.Research Report No. RR-25-03. ETS. https://rr.ets.org/index.php/etsrr/article/view/9/5

52. Sparks, J.R., Ober, T.M., Tenison, C., Arslan, B., Roll, I., Deane, P., Zapata-Rivera, D., Gooch, R., and O'Reilly, T. (2024). Opportunities and Challenges for Assessing Digital and AI Literacies.ETS Research Institute. https://www.ets.org/pdfs/rd/ets-digital-literacy-ai-full-report.pdf

53. Chand, S.P., & Pillay, K. (2024). Understanding the fundamental differences between formative and summative assessment. *Global Scientific and Academic Research Journal of Education and Literature*, 2, 6–9. https://gsarpublishers.com/wp-content/uploads/2024/02/GSARJEL052024-Gelary-script.pdf

54. Hadzhikoleva, S., Hadzhikolev, E., Gaftandzhieva, S., & Pashev, G. (2025). A conceptual framework for multi-component summative assessment in an e-learning management system. *Frontiers in Education*, 1–12. https://doi.org/10.3389/feduc.2025.1656092

55. For more information on explainable AI, see Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Morai, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Diaz-Rodriguez, N., and Herrera, F.. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, Article 101805. https://doi.org/10.1016/j.inffus.2023.101805

56. Stanford University (n.d.). *ROAR: The Rapid Online Assessment of Reading*. https://roar.stanford.edu/

57. Stanford Project on Adaptation and Resilience in Kids (n.d.). *Studies*. Stanford University. https://sparklab.stanford.edu/studies

58. For more details on the terms listed in this glossary, see National Council on Measurement in Education (n.d.). *NCME assessment glossary*. https://ncme.org/resources/professional-learning/glossary

Photography: Ryan Zhang

Stanford | Accelerator for Learning