CAST | Until learning has no limits®

Interweaving • AI • Assessment and • Accountability in Responsive Education

Interweaving Assessment in the Age of AI:

Policy Implications for Advancing Innovation that Centers Learner Variability

Cover photograph © GatesArchive/Joy Asico-Smith

Acknowledgements

This white paper was developed in partnership by CAST and Menlo Education Research (MenloEDU) as part of MenloEDU's Interweaving AI, Assessment, and Accountability in Responsive Education Initiative. Special thanks to Lillian Pace from KnowledgeWorks, Nathan Kriha and Nicholas Munyan-Penney from EdTrust, the Interweaving Assessment Project Steering Committee, and the participants from across the three Interweaving Assessment convenings.

Table of Contents

Executive Summary **1**

Introduction **2**

Summary of the Three Convenings **2**

Current Context **3**

Current Landscape of Summative Assessment in the U.S. 3

Designing for Learner Variability: UDL and Accessibility in Assessment 5

Key Learnings from the Interweaving Assessment Series **6**

R&D Priorities 7

Words of Caution & Reflections 12

Enabling Conditions & Policy Levers **14**

Increase Funding for State Assessment Innovation 14

Align Federal R&D to Assessment Approaches that Address Learner Variability 15

Coherence and Alignment 17

Build Capacity of State Leaders, Teachers, and District Leaders 18

Trust and Responsible Use 19

Conclusion **20**

Final Note: A Window into the Design of this White Paper 22

References **23**

photograph © Gates Archive/Matei Horvath

Executive Summary

Menlo Education Research facilitated the Interweaving Assessment convenings, a series of three in-person convenings designed to examine how artificial intelligence (AI)—specifically generative AI tools—and other emerging technologies are reshaping educational assessment.

This white paper synthesizes **policy-relevant insights and recommendations** from across the convenings, with a focus on the first convening's **emphasis on AI, accessibility, and designing for learner variability.**

The white paper offers a summary of the three convenings, background on the current landscape of educational assessment in the U.S., and an overview of Universal Design for Learning (UDL) and accessibility as they relate to assessment.

Building on these foundations, the white paper outlines **key research and development (R&D) priorities** identified by participants as critical opportunities to advance assessments that are more accessible, fair, and valid through the lens of learner variability.

Further, the white paper surfaces the **cautions, tensions, and unresolved questions** that participants raised with regard to AI-enabled assessment, emphasizing the need for thoughtful and responsible design and use.

Finally, the white paper articulates a set of **enabling conditions and associated policy levers** that can advance innovation at the intersections among assessment, AI, accessibility, and learner variability.

photograph © Gates Archive/Joy Asico-Smith

Introduction

Emerging technologies—specifically artificial intelligence[1] —are rapidly reshaping the landscape of educational assessment, creating both unprecedented opportunities and complex challenges. To explore these shifts, Menlo Education Research facilitated a series of three in-person convenings designed to gather together a curated mix of individuals across the assessment ecosystem. These convenings were designed to foster critical dialogue, surface shared priorities, and identify actionable pathways for assessment research and development in an evolving technological context.

The goal of this white paper is to surface policy-relevant insights and recommendations from the Interweaving Assessment series of convenings, particularly the first convening's emphasis on AI, accessibility, and designing for learner variability. This white paper examines opportunities for strengthening assessments to address learner variability with specific consideration of accessibility, fairness, and bias, as well as priority areas for research and development (R&D).

1 The convenings used the Organisation for Economic Co-operation and Development's (OECD) definition of artificial intelligence: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (OECD, 2026).

Summary of the Three Convenings

The initial convening, Interweaving Assessment: AI & Accessibility, was held in conjunction with CAST's UDL-Con in July 2025 and was designed in collaboration with Lindsay Jones of CAST and Dr. Susan Lyons of Lyons Assessment Consulting, as well as the project steering committee. This convening focused on how to advance new approaches, methods, and strategies to support innovation in assessment R&D to be responsive to learner and context variability. Responsive education recognizes that variability exists not only within learners but also within the contexts in which learning occurs. This attention to both learner variability and context variability supported participants to imagine assessments that are better able to adapt to the diversity of learners and learning environments. Participants engaged in cross-role conversations to identify psychometric topics and emerging technologies (AI as well as machine learning and other innovations) that have the potential to transform current approaches to assessment in ways that elevate accessibility as a core feature rather than a peripheral consideration. Universal Design for Learning (UDL), a framework developed by CAST to guide the design of accessible, inclusive, and challenging teaching and learning environments, proved central to the first convening.

The second convening, Interweaving Assessment: Rethinking Fairness, Bias, and Comparability in the Age of AI, was held following the National Council on Measurement in Education's AIME-Con in October 2025. This convening was designed in collaboration with Dr. Susan Lyons of Lyons Assessment Consulting, Dr. Amy Burkhardt of Cambium Assessment, Gaby López of Overflow Education Partners, and the project steering committee. The convening aimed to extend the conversations generated in the first convening by interrogating how validity (broadly conceived) must be reconceptualized within an evolving technological landscape to ensure fairness and minimize bias. The examination of unique use cases generated critical conversations, and participants prioritized opportunities, challenges, and concerns associated with addressing evolving validity approaches in response to applications of AI.

Finally, the third convening, Interweaving Assessment: AI & Accountability, was held in conjunction with the National Council on Measurement in Education (NCME) Annual Meeting and the American Educational Research Association (AERA) Annual Meeting in April 2026. Planned alongside the project steering committee, this convening built off of the learning generated in the first two convenings. This final convening focused on surfacing new priorities to address challenges and opportunities associated with integrating AI into accountability and school improvement systems. Participants discussed how AI-based innovations could address pain points in the current system across federal, state, and local policies and processes, including the need for investments in infrastructure needed for strong data governance and sovereignty policies, leveraging innovations to rebuild trust and community agency, and supporting more context-specific improvement strategies.

The first convening's **emphasis on AI, accessibility, and designing for learner and context variability was a unifying theme across the three convenings**. The conversations generated a coherent set of insights and policy implications for the field, pointing toward collaborative, high-leverage efforts to responsibly advance assessment innovation for every learner.

photograph © Gates Archive/Joy Asico-Smith

Current Context

Current Landscape of Summative Assessment in the U.S.

K-12 academic assessments, particularly federally required state summative assessments, are facing escalating political pressure as states across the country work to address persistent low and inequitable academic achievement. Although many education, business, and civil rights advocates continue to view assessments as essential tools for advancing equity, rising concerns about overly standardized formats, delayed reporting of results, and conflicting signals from federal policy on the role of summative assessments for classroom instruction threaten their effectiveness as instruments for meaningful school improvement (Mehrotra & Munyan-Penney, 2023). Advancements in assessment innovations—such as formative assessments aligned to content standards and local curricula; more culturally inclusive, relevant, and personalized items and performance tasks; and more timely and actionable reporting—could address these challenges and reaffirm the value of state summative assessments within a balanced assessment system as a tool to direct necessary resources to struggling schools to improve outcomes for students.

Meanwhile, in response to concerns, states and districts are pursuing a variety of policy strategies—sometimes multiple simultaneously—to address the most pressing pain points. First, there's been a dramatic increase in additional assessment requirements enacted by states and districts with the dual purposes of providing more instructionally useful information while helping "predict" end-of-year proficiency including literacy and math screeners and interim assessments. However, in many cases district and school policies have simply added to the number of tests students and educators experience without improved instructional utility or increased student achievement (Education First, 2026).

States are also investing in innovations in summative assessment systems, with states increasingly seeking to leverage federal funds via Competitive Grants for State Assessments (CGSA)—as federal formula grants for state assessment only cover costs for administering current assessments. Several states have developed through-year assessments, which break summative assessments into several parts administered during the course of the year, to provide more timely data for educators. Others are working to develop new assessments that measure a broader array of skills, such as critical thinking, communication, collaboration, creativity, and ethical and global citizenry.

States are also considering changes to the assessments used in their state accountability systems to better capture diverging student pathways in high school. Some are pursuing waivers from federal requirements under the Every Student Succeeds Act (ESSA), in alignment with the Trump Administration's push for states to apply for flexibility, to enable changes in how they identify low-performing high schools.

Against this backdrop, AI and other emerging technologies have the potential to help states address many of these pressures in ways that are currently not possible. However, these technologies must be leveraged thoughtfully and safely.

Designing for Learner Variability: UDL and Accessibility in Assessment

These escalating political pressures and renewed commitment to innovation create a moment of opportunity for designing more accessible and inclusive summative, interim, and formative assessments from the start. In order to accurately assess what learners know and can do, we need a balanced system of assessments that serve their unique purposes and anticipate and honor students' varying strengths and areas for growth.

Universal Design for Learning (UDL) offers a framework to proactively address and design for learner variability. Drawing from research in neuroscience, the learning sciences, cognitive psychology, disability studies, inclusive education, and asset-based pedagogies, UDL promotes a relocation of the notion of barriers—from within the individual to within the design of the learning environment (Meyer et al., 2025). The framework is grounded in three principles: design multiple means of engagement, design multiple means of representation, and design multiple means of action and expression. CAST developed the UDL Guidelines to support educators, researchers, assessment developers, and other members of the education community to apply these three principles to practice in ways that reduce barriers and increase access to meaningful and challenging learning opportunities. Momentum for UDL is growing, with schools and districts throughout the U.S. and abroad embarking on UDL initiatives, and major federal policies—including ESSA—explicitly reference UDL as a core framework for inclusive design (Rao, et al., 2023).

Accessibility lies at the foundation of the UDL framework and associated UDL Guidelines. Accessibility is shaped by what we need to do, our interactions with the environment, and our personal preferences. When designing for accessibility, the following questions are essential: 1) To whom is it accessible? 2) Under what conditions? and 3) For which tasks? (National AEM Center, n.d.). There are growing examples of accessibility in federal policy, particularly around assessment. ESSA not only explicitly references UDL but also encourages states to design assessments that reduce barriers and increase accessibility for all learners (Every Student Succeeds Act, 2015; Lazarus et al., 2022). Similarly, the Individuals with Disabilities Education Act (IDEA) and Section 504 of the Rehabilitation Act of 1973 require that students with disabilities participate in state and districtwide assessments with appropriate accommodations or alternate assessments, while Title II of the Americans with Disabilities Act (ADA) requires public educational institutions to ensure that digital services, websites, web applications, and technology-enabled learning environments are accessible to individuals with disabilities (U.S. Department of Education, 2026; U.S. Department of Justice, 2024).

The intersections of UDL and accessibility open up new possibilities for imagining innovative assessments that are responsive to every learner. The design of traditional assessments may include unintended barriers that prevent an accurate understanding of what learners know and can do. Designing assessments with UDL and accessibility from the start can preserve construct validity while removing irrelevant barriers (Johnstone, et al., 2006; Lazarus et al., 2022; Rose et al., 2018). This intentional, proactive design has the potential to embrace learner variability and, simultaneously, strengthen the validity, fairness, and usefulness of assessments, especially in technology-enabled contexts.

Guiding questions to consider as you engage with the white paper:

- What are your goals for engaging with this white paper?
- How do the themes shared in this white paper connect to your current role and context?
- Which themes resonate most with you?

Key Learnings from the Interweaving Assessment Series

Across the three convenings, participants engaged in conversations and activities that prompted them to imagine what it will take to responsibly integrate AI into educational assessment systems. The first convening's emphasis on AI, accessibility, and designing and innovating to support learner variability was a common thread of discussions across the three convenings—both by intentional design and through the ways participants' conversations naturally evolved. The key learnings presented in this white paper synthesize insights that have direct implications for policy: priority areas for assessment R&D as well as cautions and reflections that must inform future decision-making.

R&D Priorities

Participants emphasized that realizing the potential of AI-enabled assessment will require intentional and sustained investment in R&D. The priorities outlined in this section reflect areas where participants identified the greatest need—and opportunity—for advancing assessment systems that are more accessible, fair, and valid. While these priorities are not exhaustive, they represent directions for research, experimentation, and collaboration across the field.

A New Measurement Paradigm

Perhaps the most fundamental shift participants called for is a move away from the standardization paradigm that has long defined educational assessment. Traditional approaches have pursued comparability by holding the assessment and administration conditions constant across students—but many participants argue this provides only superficial comparability. A more meaningful path forward lies in tailoring assessment tasks that enable comparable inferences about student learning and skills while providing more personalized and relevant content, so that assessments reflect what a student actually knows and can do rather than how well they navigate a fixed format. This tailoring could also lead to greater instructionally relevant insights across assessment types by offering richer data on students' personal progress. This requires developing new ways of understanding student context, including how different items perform differently across different learners—variance that current methods are ill-equipped to account for.

This paradigm shift also demands reimagined approaches to gathering evidence of psychometric quality. When assessment results come from individualized activities, conventional evidence frameworks may no longer apply. The field must invest in developing new methods and frameworks suited to these emerging tools—methods that can confirm an assessment is measuring what it claims to measure and can support meaningful comparisons of students' knowledge, skills, and abilities, even when the form of that assessment varies from student to student.

Example Use Cases: A New Measurement Paradigm

- An AI-enhanced assessment environment could observe and analyze how students interact with materials and peers during performance tasks, capturing academic skills alongside evidence of durable skills like collaboration and adaptability that traditional formats cannot reliably measure.
- An AI-enhanced assessment environment could generate and administer assessment items that mimic real-world scenarios that align to students' interests, creating more authentic opportunities for students to demonstrate knowledge and skills within the context of classroom learning.
- Leverage AI-enabled personalization and connected reimagining of validity and reliability to provide a more accurate picture of skills and knowledge of students' at far ends of score distribution–where current measures are often the most imprecise–ideally leading to supports more aligned to student needs (Lord, 1980; Smarter Balanced Assessment Consortium, 2025).

Inclusive Co-design

Equally important as the new measurement paradigm is ensuring that the design and use of these tools reflects a diversity of voices. Inclusive co-design—bringing together educators, students, families, and communities across backgrounds—must be a cornerstone of how next-generation assessments are built and validated. By engaging those who will ultimately use and be affected by assessment systems, co-design can improve technical quality by strengthening validity, reducing construct-irrelevant variance, identifying potential sources of bias, improving accessibility and usability, and ensuring that assessments produce meaningful and interpretable evidence of learning across a diversity of populations. Inclusive co-design also helps ensure that assessment systems reflect the values and needs of the learners they serve. Co-design emphasizes the importance of designing with, rather than for, potential users within the community, recognizing them as active partners whose knowledge, lived experiences, and perspectives inform all stages of the design process (Costanza-Chock, 2020; Mejeh & Sarbach, 2025; Penuel, 2011).

In addition to improving the quality of data, intentionally building in the perspectives of the full range of potential users of AI-enabled assessments and the resulting data will help ensure that a diverse set of learners can access content and demonstrate their skills. Additionally, this process will build trust with students and educators—essential for sustainable implementation that tailors instructional practices to learners needs and, ultimately, improves student outcomes.

Example Use Cases: Inclusive Co-Design

- Conduct focus groups with students to more fully understand their experiences with educational assessment and to explore any barriers they encounter.
- Administer surveys to educators to uncover their perspectives on the kinds of features and options that could be embedded into assessment design to more accurately assess students' understanding.
- Conduct listening sessions with parents and caregivers to explore their perspectives on educational assessment and the barriers their children may face in terms of expressing what they know and can do.

Addressing Bias

As AI-enabled assessments increasingly incorporate behavioral and multimodal signals—such as eye contact, response time, and facial expressions—the risk of encoding existing biases into new systems becomes a pressing concern. Participants highlighted the need for rigorous research into whether new AI-developed measures–including those designed to reflect a wider range of student identities–and automated scoring are accurate, fair, and sensitive to the full range of learner variability. Assumptions embedded in design, such as equating speed of response with performance, could replicate or even amplify longstanding inequities if left unexamined. If the goal is to produce data that meaningfully reflects what students know and can do, then assessments that are skewed by bias are not simply unfair—they also produce invalid inferences about students' abilities, producing a distorted picture that cannot reliably inform instruction or support.

Participants also called for advances in AI-enabled personalization that move away from broad demographic categorizations and instead attend to students' intersectional identities. Grouping students into blunt categories obscures meaningful variation and risks producing assessments that serve some learners well while systematically failing others. Developing more nuanced, identity-aware modeling approaches will be essential to building systems that are genuinely equitable—and to ensuring that the actionable insights these systems generate are trustworthy for every student.

Example Use Cases: Addressing Bias

- AI-enabled automated item generation could quickly develop a wide range of items that are relevant and engaging for students with various interests and identities.
- Writing prompts and other constructed response items can be scored by AI trained using a wide variety of student samples and aligned rubrics, to increase both the efficiency and accuracy.

Multimodal Assessment

Where current assessments are largely constrained to students reading and writing as the primary means of expression, AI-enabled tools can process and interpret a far richer range of student output: spoken responses, drawings, movement gestures, and other forms of demonstration that better reflect the full range of human capability. New assessments also have the potential to disentangle skills and measure each on its own terms: a student who can listen to a passage, and articulate a thoughtful oral response is demonstrating literacy skills that traditional assessment might fail to capture. Often these assessments conflate content knowledge with the ability to decode text and produce written output which can impact and inform important instructional choices.

AI is uniquely positioned to make this kind of assessment feasible at scale, processing and interpreting diverse forms of student output in ways that were previously impractical. Research is needed not only to develop these tools, but to understand how to interpret and act on the richer data they produce—ensuring that educators receive clear, usable insights rather than an overwhelming volume of information without guidance.

Example Use Cases: Multimodal Assessment

- AI could leverage student responses and interactions within the assessment platform, alongside curriculum documents to suggest opportunities to support students in upcoming units or lessons, or activities to reinforce skills at home.
- Performance tasks could be observed (and also developed) by AI, watching how students interact with materials and their classmates, to demonstrate a wide range of knowledge and skills.
- AI can generate and observe assessment items that mimic authentic, “real-world” scenarios, more actively engaging students.

Asset-Based and Personalized Feedback

Participants stressed that assessment data is only as valuable as what is done with it—and that AI's potential to deliver timely, personalized feedback represents one of its most powerful opportunities to improve learning. Too often, feedback in educational settings is corrective and deficit-focused, telling students what they got wrong without building on what they got right. Feedback designed from an asset-based perspective—one that identifies students' strengths and frames growth as an extension of existing capability—has the potential to sustain engagement, nurture curiosity, and support students through productive struggle in ways that deficit-oriented approaches cannot.

As AI systems become more capable of personalizing feedback at scale, there is a meaningful opportunity to redesign these feedback loops intentionally, including research into how feedback should be timed, framed, and differentiated across learners with different needs, identities, and learning histories. This includes educators and families who are too often left with assessment reports that are dense, delayed, and difficult to translate into action. AI can change this by drawing on student responses and interactions within the assessment platform—alongside curriculum documents and other data systems to surface targeted, timely recommendations for educators and families. Getting all of this right will require close collaboration with educators, families, and students to ensure that feedback at every level feels accurate–including overcoming biases–and actionable in practice.

Example Use Cases: Asset-Based and Personalized Feedback

- Students receive personalized, AI generated feedback that identifies how students can build on existing strengths to develop new skills and understanding that increase student motivation, and can be leveraged by educators to have student-driven conversations about learning.
- Families and educators could ask AI chatbots questions about assessment results to better understand what students know and can do, and where they may need additional support.
- Pulling data from assessments, as well as additional data, such as from Multi-Tiered System of Supports (MTSS) or state accountability systems, to identify what additional resources and supports could be targeted to specific students, classrooms, or schools.

Best Practices and Standards

Across all of these priorities, **participants recognized that individual innovations will only reach their potential if the field invests in the shared infrastructure needed to develop, evaluate, and disseminate them responsibly.**

Two interconnected areas of standards work emerged as especially important.

1. The first is the **development of frameworks for fairness and transparency.** As AI-enabled assessment tools proliferate, there is an urgent need for shared definitions of what fairness looks like in practice—and for transparency standards that require developers to clearly document how their systems are built, what data they draw on, and how they have been tested across different student populations. Without such standards, the burden of evaluating these tools falls unevenly on under-resourced districts and educators, and accountability for harmful outcomes becomes difficult to assign.
2. The second is **investment in infrastructure for sharing and building knowledge across the field.** Research and innovation in AI-enabled assessment currently risk being siloed within institutions or proprietary platforms, limiting the field's ability to learn collectively. Developing shared repositories, open standards, and collaborative networks would allow researchers, educators, and developers to build on one another's work—accelerating progress, reducing duplication, and ensuring that breakthroughs in accuracy, fairness, and actionability reach the students and schools that need them most. Together, these standards and infrastructure investments form the connective tissue that will allow the rest of this R&D agenda to grow into something durable and equitable.

Words of Caution & Reflections

While the focus on AI, accessibility, and designing for learner variability generated momentum around opportunities for innovation, participants regularly emphasized the need for careful reflection and caution as AI-enabled assessment approaches move from experimentation to broader use. Across discussions, the assessment R&D ecosystem emerged as a critical intermediary—responsible for ensuring that new tools and approaches support every learner, strengthen validity and fairness, and do not introduce unintended harm. **The reflections in this section highlight key risks, tensions, and unresolved questions that must be addressed to ensure that innovation serves students rather than placing them at risk.**

AI Training Data Gaps

AI-enabled tools are only as strong as the data used to train their underlying components. When training data is unrepresentative, AI technologies may perform as intended for some students while producing inaccurate or unreliable results for others. This is a risk that is particularly acute for students who are already underrepresented and underserved in the current educational system. How the field can realistically collect sufficiently representative data, covering the full range of learner variability, remains an open and pressing question the R&D ecosystem faces.

Algorithmic Bias

AI tools are trained on, and use data, that may reflect dominant cultural norms in ways that systematically disadvantage students from traditionally marginalized communities. For example, nascent research has found evidence that large language models can meaningfully alter outputs depending on the inclusion of first and last names suggestive of different races or genders. Bias is currently an unavoidable reality of many AI technologies raising questions about what's needed to ensure AI-enabled assessment tools can consistently and accurately measure what they are built to measure for all students.

Student Data Ownership and Privacy

AI-enabled assessment tools will collect a significant amount of data on student users, raising serious questions about privacy, data ownership, and how to appropriately handle student data in the AI era. Participants expressed concern about student data being collected and exploited by organizations, and the potential for AI-collected data to be misused by other AI-enabled technologies to inappropriately profile students.

Overreliance

As AI-enabled tools become more widely adopted in education systems, there is a real risk that teachers and leaders may over-rely on AI-generated insights at the expense of their own professional judgment. When AI outputs are treated as an objective truth rather than one metric among many, consequential decisions about students' educational outcomes and experiences may be made on incomplete or inequitable grounds. How the field builds AI literacy capacity to overcome issues of overreliance is a crucial policy question to address moving forward.

Prompts to support engagement and reflection:

- Which of these R&D priorities resonate most with your context?
- Is there a theme that seems missing based on your experiences and perspectives?

photograph © Gates Archive/Joy Asico-Smith

Enabling Conditions & Policy Levers

The key learnings across the Interweaving Assessment series generated thoughtful conversation around the ways policy can support the necessary R&D infrastructure and other enabling conditions that can advance innovation at the intersections among assessment, AI, accessibility, and learner variability.

Increase Funding for State Assessment Innovation

Enabling Conditions

Meaningful state and local assessment innovation depends on sustained and sufficient funding to develop, implement, and evaluate promising ideas. Yet growing demand for more authentic, learner-centered assessment systems collides with the current reality of constrained education budgets and the ongoing costs of maintaining existing statewide assessment systems. As a result, assessment innovation tends to be episodic, under-resourced, and limited in its ability to scale.

While the federal government provides funding to help states administer required assessments, direct and sustained investment from states and philanthropy in assessment innovation has been minimal. Congress currently appropriates between $12 million and $29 million annually for CGSA, despite strong state demand for funds. Even at its 2024 peak, CGSA awards supported only 10 of the 20

states that applied to improve their assessment systems, and recent competitions capped awards at between $3 million and $4 million, making it challenging to advance more complex solutions at scale. Increased and sustained federal funding is needed to meet demand and to help states responsibly harness emerging technologies to advance more learner-centered approaches to assessment.

Policy Levers

- Increase and sustain federal funding for CGSA to meet demonstrated state demand and lift arbitrary caps on the award size to ensure grants are sufficient to support high-quality, scalable innovation.
- Prioritize state-level investment in assessment innovation to enhance accessibility and learner-centered designs across state summative, interim, and formative systems.

Align Federal R&D to Assessment Approaches that Address Learner Variability

Enabling Conditions

The federal government plays a critical role in shaping the national education research agenda to advance evidence-based improvements to education systems. The Institute of Education Sciences (IES) leads these investments, investing in research, evaluation, and dissemination to strengthen educational practice and outcomes. Recent efforts to reconsider and rebuild IES have sparked a national conversation about how federal education research can be more responsive, timely, and impactful. A focus on assessment innovation has begun to emerge in these conversations, with advocates emphasizing the need to align federal R&D with a focus on learner variability.

To meet this charge, IES will require enhanced capacity and a structure better suited to responding to urgent and emerging education challenges. For years, federal policymakers have explored creation of a rapid-response R&D entity within IES modeled after the Defense Advanced Research Projects Agency (DARPA). A comparable education approach in education could accelerate innovation in a field undergoing rapid change, including shifting workforce expectations and the emergence of new technologies such as AI. Such a structure would be especially helpful to the advancement of learner-centered assessment approaches.

Additionally, the federal government can provide states with a viable policy pathway to test and refine new ideas for strengthening assessments. Historic efforts to create federal assessment pilots have fallen short, with minimal state participation due to concerns over heavily bureaucratic processes, constraining guardrails, and lack of aligned funding. Going forward, federal policymakers must work in partnership with states to co-design a more viable pathway that preserves technical quality while enabling responsible innovation.

Policy Levers

- **Prioritize field-identified R&D priorities**, including the R&D concepts identified in these convenings, into the scope of future IES investments. These priorities reflect critical research needs identified by assessment designers, researchers, and practitioners. Align these investments with related federal initiatives such as Innovative Assessment Demonstration Authority (IADA), CGSA, and ESSA Titles I and II-A.
- **Create and fund a new center within IES dedicated to rapid research and development of promising education innovations**, including those with the potential to transform student assessment. Bipartisan proposals that advance these goals include:
    - *The New Essential Education Discoveries (NEED) Act*, which would establish a National Center for Advanced Development in Education at IES to develop and disseminate cutting-edge, evidence-based practices.
    - *The Developing and Advancing Innovative Learning Models Act*, which would provide $500 million annually to support R&D for innovative learning models with a strong emphasis on learner-centered education.
- **Develop a viable pathway that provides states with optimal conditions to test, refine, and scale new approaches to assessment** that maximizes advances in AI and emphasizes learner-centered designs. These conditions should provide states with an on-ramp to plan new ideas, grant permission to iterate mid-stream based on key learnings, establish realistic timelines to scale innovations, and provide sufficient funding to build something meaningful. Align federal funding for assessment innovation, including CGSA and relevant R&D initiatives, to this opportunity to provide states with coherent policy flexibility and funding.

Coherence and Alignment

Enabling Conditions

A foundational enabling condition for AI-enabled assessment innovation is a coherent, shared vision for student success—one that is grounded in what learners will actually need to thrive in postsecondary education and the modern workforce. Without clarity on what the system is preparing students to do, it is difficult to evaluate whether assessments are measuring what matters or to make principled decisions about where innovation is most needed. With that vision in place, states are better positioned to make strategic investment decisions—directing resources toward the specific pain points in their assessment systems that are most pressing, rather than pursuing innovation for its own sake that could introduce additional pain points into assessment systems and data use.

Coherence is needed not only between assessment systems and broader state education goals and aligned initiatives—such as high-quality instructional materials adoption or science of reading reforms—but also across the different types of assessments themselves. Screeners, diagnostics, interims, formative assessments, and summative assessments each serve distinct purposes, and a coherent system requires that these tools work in concert rather than at cross-purposes.

Policy Levers

- **Establish diverse expert workgroups to guide assessment development and procurement.** States should convene workgroups that bring together expertise in assessment design and implementation, teaching and learning, AI, and accessibility to develop a shared vision for next-generation assessments and establish clear criteria to guide RFPs for new assessment vendors. A diversity of representations in these groups is essential to ensuring that new tools are designed with equity, usability, and alignment front-of-mind from the outset.
- **Review standards frameworks to reflect critical knowledge and skills.** As visions for student success evolve, states should review existing standards frameworks to ensure they are aligned to the academic knowledge and skills that matter most—potentially including durable skills that current assessments often fail to capture.

Policy Levers *(continued)*

- **Launch aligned assessment pilots and capacity-building networks.** States should invest in piloting new assessment approaches in ways that are deliberately connected to pain points in current assessments—including those beyond state summative assessments. These should be paired with networks that build educator capacity and feedback loops, allowing educators to share what is working and surface implementation challenges early.
- **Build coherent and aligned assessment systems.** As new AI-supported assessments are piloted and scaled across schools and districts, state leaders should develop tools and guidance that help districts in the implementation of fully balanced assessment systems. These tools should support local leaders in ensuring all assessments are aligned to state standards, used only for their designed purpose, and coherent with and not duplicative of other assessments.

Build Capacity of State Leaders, Teachers, and District Leaders

Enabling Conditions

Building the capacity of state and district education leaders is another key condition for advancing innovation at the intersections of assessment, AI, accessibility, and learner variability. Central to this work is creating sustained professional learning opportunities that develop leaders' deep conceptual understanding of UDL and its application to assessment design, as well as leaders' mindsets toward assessment as a flexible system in which students can demonstrate mastery through multiple pathways.

Further, support can be offered to state and district leaders making procurement decisions around AI and assessment tools. Developing an array of guidance and resources can support leaders to develop procurement structures and processes that lead to assessment ecosystems that safely and effectively integrate AI.

Policy Levers

- **Leverage National Comprehensive Centers** to provide technical assistance and guidance to state and district leaders about how to safely and effectively leverage AI to support assessment innovations, aligned with states' CGSA-funded projects to the extent possible.
- **Align federal investments in teacher preparation and development** to build educator capacity for designing and implementing innovative assessment approaches with particular emphasis on data literacy and instructional use.
- **Fund and invest in accessible AI literacy training** for state and district leaders—along with educators and school staff—that goes beyond foundational knowledge of generative AI to include responsible and legally required data privacy practices, strategies for identifying and mitigating systemic bias of AI technologies, awareness of environmental considerations, and guidance on how to transparently communicate with students about the use of AI in their assessments.

Trust and Responsible Use

Enabling Conditions

Responsible integration of AI into educational assessments depends not only on the quality and efficacy of the AI tools, but also on the trust of students, families, educators, and communities that they are designed to serve. Without meaningful investment in building that trust, even well-designed and research-backed innovations risk low adoption, resistance, or failure to be implemented as intended.

Trust must be cultivated across various groups through meaningful action. Students and families need assurance that sensitive data collected through AI-enabled tools are protected and kept internal, used only for educational purposes, and the technologies are implemented in safe and responsible ways that mitigate against negative or unintended consequences. Educators, school leaders, and district leaders need confidence that AI tools have been rigorously tested for bias and that they have been

adequately trained to identify and mitigate the limitations of said tools in practice. And communities, which include students, parents, and practitioners, need to see themselves reflected in the design of AI tools through authentic co-creation processes aligned with community values and goals. Additionally, as AI adoption in professional and educational settings grows, policymakers cannot ignore important questions about the environmental costs–including which communities are most impacted–of these technologies. Acknowledging and proactively addressing these concerns is itself a trust-building act.

Policy Levers

- **Examine existing student data privacy frameworks**—including the Family Educational Rights and Privacy Act and the Children's Online Privacy Protection Act—to identify gaps specific to AI-enabled assessment contexts and explore what stronger, more targeted protections for student data may be needed.
- **Fund and incentivize co-design processes** that meaningfully engage students, families, educators, and community members in the development, implementation, and evaluation of AI-enabled assessment tools.
- **Build awareness around the environmental implications of AI adoption** among education stakeholders and highlight real-world use cases that reflect a shared commitment to environmental values and justice.

Prompts to support engagement and reflection:

- Which of these enabling conditions feel most urgent to you based on your context?
- What action(s) might you take to support a policy lever that feels most relevant and meaningful to you?

photograph © Gates Archive/Matei Horvath

Conclusion

The convergence of advances in artificial intelligence, accessibility, and our understandings of learner and context variability creates a unique opportunity to rethink educational assessment. Emerging technologies are uncovering possibilities that would have been difficult to imagine only a decade ago—from assessments that adapt to learner needs and contexts, to multimodal demonstrations of knowledge, to personalized feedback that is timely, actionable, and grounded in students' strengths. At the same time, these technologies raise important questions about validity, fairness, privacy, transparency, and trust. **The challenge facing the field is not whether AI will influence assessment, but how to ensure that its influence advances educational opportunity rather than reinforces existing inequities.**

Across the three Interweaving Assessment convenings, participants consistently emphasized that **learner and context variability must be at the center of this work**. Rather than framing variability as a challenge, future assessment systems can be designed to anticipate and respond to the diversity of strengths, needs, identities, and contexts that learners bring to educational environments. Leveraging frameworks like Universal Design for Learning (UDL) and designing with accessibility and inclusivity from the start can support a move toward assessment systems that reduce barriers and increase access to more meaningful, actionable feedback for students, educators, and families.

Realizing this vision will require more than technological innovation alone. It will require a coordinated commitment to research and development, new approaches to validity and reliability, inclusive co-design processes, rigorous attention to bias and accessibility, and investments in the people and systems responsible for implementation. It will also require policymakers, researchers, educators, developers, students, and communities to work together to establish the conditions under which innovation can be pursued responsibly. The policy levers outlined in this paper—including **increased funding for assessment innovation, alignment of federal R&D priorities, coherence, capacity building, and trust-centered governance**—represent important steps toward creating those conditions.

Convening participants repeatedly returned to a foundational insight: **a core challenge with current assessment systems is not a lack of data, but barriers within the design.** Assessments built to achieve standardization across students have often produced only the appearance of comparability while failing to capture what many learners actually know and can do. This insight underscores the need for a new measurement paradigm—one that leverages advances in AI, accessibility, and our understanding of learner and context variability to generate more meaningful, accurate, and actionable evidence of learning, especially for learners whose strengths are often overlooked by traditional approaches.

The conversations generated across these convenings make clear that this moment presents a window of opportunity. As states, districts, educators, researchers, and developers consider the future of educational assessment, there is an opportunity to move beyond incremental improvement and toward a more fundamental reimagining of how learning is understood, measured, and supported. **By centering learner and context variability, accessibility, and responsible innovation, the field can design assessment systems that are not only more technologically advanced, but also more inclusive, valid, and responsive to every learner.**

Final Note: A Window into the Design of this White Paper

This white paper was intentionally designed using the UDL framework. The framework is built on three principles: Design Multiple Means of Engagement, Design Multiple Means of Representation, and Design Multiple Means of Action and Expression. Below we share **how we leveraged each of these three principles to reduce barriers and increase access to the ideas presented in this white paper**. We offer this window as a model for the design of other more accessible and inclusive white papers and reports.

- Design Multiple Means of Engagement
  - Background information on the project highlights the relevance of our message and invites the audience to engage with the ideas presented
  - Guiding questions at the beginning of the white paper—and reflection questions throughout—invite the audience to connect to their own context
- Design Multiple Means of Representation
  - Fully accessible PDF format can be accessed using assistive technology such as a screen reader
  - Images and image descriptions offer multiple representations
  - Call-out boxes, icons, and heading structure highlight critical features
- Design Multiple Means of Action and Expression
  - Example use cases and policy levers support putting the ideas we share into action
  - Executive summary and table of contents support organizing and managing the information presented

Prompts to support engagement and reflection

- Which of these UDL features best support you as a learner?
- What other features would support you to more fully engage with the ideas presented in this white paper?

References

CAST. (2024). Universal Design for Learning guidelines version 3.0. https://udlguidelines.cast.org

Costanza-Chock, S. (2020). Design justice: *Community-led practices to build the worlds we need.* MIT Press.

Education First. (2026). Rethinking the test pile: A national study of K–8 academic assessments. https://www.education-first.com/wp-content/uploads/2026/04/Final-Local-Assessments-Landscape-Report_February-2026.pdf

Every Student Succeeds Act, 20 U.S.C. § 6301 et seq. (2015).

Johnstone, C., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. National Center on Educational Outcomes, University of Minnesota.

Lazarus, S. S., Johnstone, C. J., Liu, K. K., Thurlow, M. L., Hinkle, A. R., & Burden, K. (2022). *An updated state guide to universally designed assessments* (NCEO Report 431). National Center on Educational Outcomes. https://nceo.info/Resources/publications/OnlinePubs/Report431/default.html

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Mehrotra, S. & Munyan-Penney, N. (2023). *Future of assessments: Centering equity and the lived experiences of students, families, and educators*. https://edtrust.org/rti/future-of-assessments-centering-equity-and-the-lived-experiences-of-students-families-and-educators/

Mejeh, M., & Sarbach, L. (2025). Co-design: From understanding to prototyping an adaptive learning technology to enhance self-regulated learning. *Technology, Knowledge and Learning, 30*(2), 1021-1046.

Meyer, A., Rose, D. H., & Gordon, D. (2025). *Universal design for learning: Principles, framework, and practice* (3rd ed.). CAST Professional Publishing.

National AEM Center. (n.d.). What is accessibility? https://aem.cast.org/get-started/defining-accessibility

OECD. (2026). OECD AI Principles overview. https://oecd.ai/en/ai-principles

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher, 40*(7), 331-337.

Rao, K., Gravel, J. W., Rose, D. H., & Tucker-Smith, T. N. (2023). Universal Design for Learning in its 3rd decade: A focus on equity, inclusion, and design. *International encyclopedia of education*, 6, 712-720.

Rose, D. H., Robinson, K. H., Hall, T. E., Coyne, P., Jackson, R. M., Stahl, W. M., & Wilcauskas, S. L. (2018). Accurate and informative for all: Universal design for learning (UDL) and the future of assessment. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices: Issues, innovations, and applications* (pp. 167-180). Cham: Springer International Publishing.

Smarter Balanced Assessment Consortium. (2025). 2023–24 summative assessment technical report. https://technicalreports.smarterbalanced.org/2023-24_summative-report/_book/index.html

U.S. Department of Education. (2026). Individuals with disabilities. https://www.ed.gov/laws-and-policy/individuals-disabilities

U.S. Department of Justice. (2024). Nondiscrimination on the basis of disability: Accessibility of web information and services of state and local government entities. Federal Register, 89(80), 31320–31406. https://www.federalregister.gov/documents/2024/04/24/2024-07758/nondiscrimination-on-the-basis-of-disability-accessibility-of-web-information-and-services-of-state

This work was supported, in whole or in part, by the Gates Foundation. The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. Please note works submitted as a preprint have not undergone a peer review process.