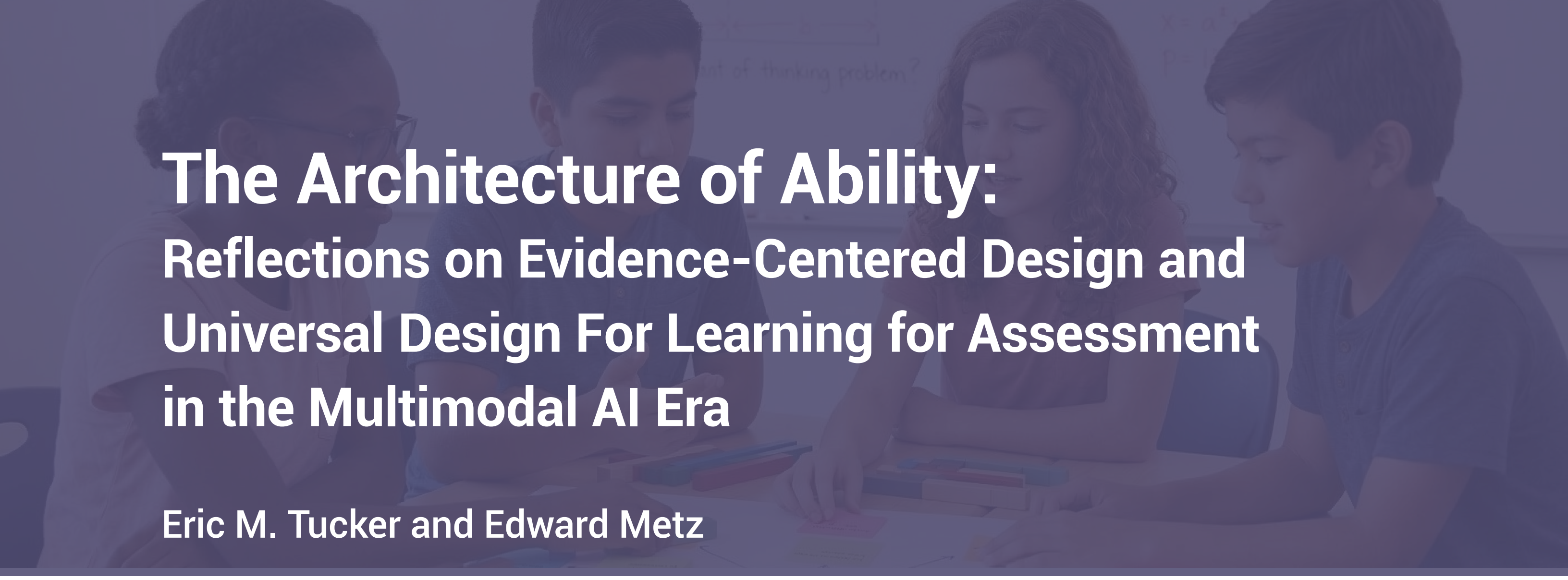
CASE STUDY

The Architecture of Ability:

Reflections on Evidence-Centered Design and Universal Design For Learning for Assessment in the Multimodal AI Era

Eric M. Tucker and Edward Metz



The Architecture of Ability:
Reflections on Evidence-Centered Design and Universal Design For Learning for Assessment in the Multimodal AI Era

Eric M. Tucker and Edward Metz

PART I: The Illusion of Scale and the Threat of Automated Noise

As state education agencies and school systems navigate the future of assessment, the field of education stands at a precarious technological inflection point. For decades, educational measurement has relied on an industrial-era model of mass production: rigid, one-size-fits-all summative tests that deliver verdicts months after the critical learning moment has passed. Driven by advances in multimodal AI, test designers and developers can break this mold. Liberated from the constraints of paper and uniform pacing, modern platforms can deliver continuous, adaptive assessments that dynamically respond to learners in real time.

Yet, as the sector pursues this AI-driven frontier, we risk harnessing extraordinary computing power merely to pave over old, multiple-choice cow paths. Superimposing flashy generative AI interfaces atop flimsy psychometric architectures will simply automate and scale bad measurement and historical biases. As the Principled Assessment Designs for Inquiry (PADI) project presciently warned two decades ago: "Technology is as seductive as it is powerful. It is easy to spend all one's time and money designing realistic scenarios and gathering complex data, and only then to ask 'How do we score it?'" (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003, p. 21). The evidentiary argument must precede the technology. AI is a powerful engine, but principled assessment design—including Evidence-Centered Design—is the required steering mechanism.

To safely and effectively navigate this modern dilemma, we cannot rely on engineering alone; we must return to the legacy of the late Dr. Robert J. Mislevy. Mislevy helped transform educational assessment from the somewhat artisanal, intuitive practice of "item writing" into a transparent, scientifically rigorous discipline of evidentiary reasoning. He taught us a foundational truth: "Educational assessment is at heart an exercise in evidentiary reasoning. From a handful of things that students say, do, or make, we want to draw inferences about what they know, can do, or have accomplished more broadly" (Mislevy & Riconscente, 2005, p. iii).

His frameworks—including Evidence-Centered Design (ECD) and its exploratory integration with Universal Design for Learning (UDL)—are indispensable blueprints required to safely architect AI-driven assessment today. As Mislevy and colleagues asserted, "Assessment design is often identified with the nuts and bolts of authoring tasks. However, it is more fruitful to view the process as first crafting an assessment argument, then embodying it in the machinery of tasks, rubrics, scores, and the like" (Mislevy, Steinberg, & Almond, 2003, p. 4).

The Courtroom of Assessment and Eliminating the "Unless" Clause

ECD treats the design of an educational assessment as analogous to a lawyer preparing for a courtroom trial. Drawing on the Toulmin model of argumentation, ECD frames the test-maker as the prosecutor making a Claim (e.g., "This 8th-grade student understands the laws of kinematics"). To support this claim, the test provides Data (e.g., "The student successfully plotted a velocity-time graph"). Connecting the Data to the Claim is the Warrant, the underlying scientific and pedagogical rationale that justifies the inference. By structuring these arguments explicitly before moving to operational tasks, ECD ensures that assessments remain anchored to their substantive evidentiary rationale (Mislevy & Riconscente, 2006).

However, in the courtroom of assessment, the defense can always offer an Alternative Explanation. Every test score makes a claim with a hidden caveat: "This student understands this concept... UNLESS."

- *Unless* they couldn't see the tiny font on the graph.
- *Unless* they have a specific learning disability and the working-memory load of the task's multi-step instructions was overwhelming.
- *Unless* they are an English Language Learner who tripped over the complex, passive syntax of the word problem.

In psychometric terms, these "unless" clauses are the literal embodiment of construct-irrelevant variance—and they are fatal to validity. When a test score fluctuates due to an 'unless' rather than the student's actual mastery of the target skill, the assessment has successfully measured the barrier, not the underlying construct. As the PADI Assessment for Students with Disabilities (ASD) research notes, "The existence of alternative explanations that are both significant and credible might indicate that validity is threatened or being compromised." (Haertel, et al., 2010, Report 1, p. 8).

A primary goal of Evidence-Centered Design is to engineer the 'unless' clauses out of a central role. Rather than waiting for a student to fail and subsequently investigating why, or retroactively applying accommodations to a rigid test, ECD maps the chain of evidentiary reasoning from its inception. By making this underlying argument explicit, the framework forces developers to proactively address validity threats, avoiding costly rework and making operational elements more amenable to examination, sharing, and refinement.

The Paradigm Shift: From Marginal to Conditional Inference

Eliminating these "unless" clauses requires a paradigm shift. For a century, the testing industry operated under the illusion of Marginal Inference: the idea that standardizing a test's surface ensured fairness. Prevailing wisdom dictated that if every student faced the exact same fonts, wording, and time limits, individual disadvantages would simply "average out" as statistical noise. Fairness meant sameness.

Mislevy and his colleagues dismantled this fallacy. They proved mathematically and pedagogically that fairness actually requires variance. As the PADI reports assert: "...making tests identical for all examinees can make a testing procedure less fair: Equivalent surface conditions may not provide equivalent evidence about examinees" (Mislevy et al., 2013, p. 1).

Consider the metaphor of the scientific instrument, championed by UDL pioneers at CAST. A rigidly standardized test operates much like an unadjustable microscope. If the focal length is permanently fused in place, a user with varying visual needs will perceive only a blur. When this occurs, the failure resides neither in the user's intellect nor in the slide itself (the focal academic content); rather, it stems entirely from the instrument's rigidity obscuring the observation. By refusing to adjust the microscope, the test introduces construct-irrelevant variance, measuring the limitations of the tool rather than the brilliance of the test taker. "Like a microscope whose eyepiece is out of focus, the instrument actually gets in the way" (Rose, Murray, & Gravel, 2012, p. 7).

This distinction becomes obvious when comparing an eye exam to an alphabet test. If an optometrist is measuring visual acuity, standardizing the twenty-foot distance to the eye chart is absolutely mandatory—the visual distance is the very construct being measured. But if an educator is evaluating a student's knowledge of the alphabet, forcing a student with low vision to stand twenty feet away is an act of measurement malpractice. The unadjusted visual demand introduces massive construct-irrelevant variance, becoming the primary alternative explanation for failure. The test stops measuring the alphabet and starts measuring the student's eyesight.

We must therefore shift our assessment paradigm from Marginal Inference to Conditional Inference—standardizing the validity of the construct while dynamically varying the task's delivery to ensure construct-irrelevant demands do not impede the learner (Mislevy et al., 2013). As Proposition 3 of the PADI framework establishes: "Surface conditions that differ in principled ways for different learners can provide equivalent evidence." (Mislevy et al., 2013, P. 5)

The Signal and the Noise: Untangling KSAs

To achieve this principled variation and execute Conditional Inference at scale, ECD asks us to dissect the demands of any task. Every educational task requires a tangled web of Knowledge, Skills, and Abilities (KSAs). The assessment designer must untangle these into Focal KSAs (the target constructs being measured) and Additional KSAs (the access skills required to interact with the task) (Seeratan & Mislevy, 2008; Zhang et al., 2010).

The Focal KSA is the pure Signal—the exact cognitive capability we intend to measure. The Additional KSAs—such as visual acuity, fine motor control, or decoding complex syntax—are the Noise.

Consider the metaphor of a traditional verbal spelling bee versus a keyboard-optional alternative. While the traditional bee intends to isolate orthographic knowledge as the Signal (the Focal KSA), its rigid format forces a spoken response, inadvertently injecting the Noise of public speaking and verbal articulation (Additional KSAs). For a student with a severe speech impairment, the noise completely drowns out the signal; the bee ceases to measure spelling and instead measures the construct-irrelevant variance of their speech impairment. Allowing that student to type the word on a keyboard acts as a noise-canceling filter. It neutralizes the construct-irrelevant hurdle without diluting the rigor of the spelling construct. Validity, therefore, is application-specific, not item-specific. "It is only by knowing the purpose of a test and the intended examinee population that one can answer how a given change will impact the evidentiary value of data for the construct meant to be assessed." (Haertel et al., 2010, Report 1, p. 1)

When an assessment inadvertently measures an additional KSA that a student lacks, it generates "construct-irrelevant variance." This is not simply an ADA compliance issue; it is a fundamental mathematical error. Drawing on Matthias von Davier's General Diagnostic Model, Mislevy explains that valid inference about the targeted construct is mathematically conditional on the necessary but construct-irrelevant KSAs not being appreciable impediments to the student. If a student encounters an insurmountable linguistic, executive functioning, or perceptual barrier, it acts as an absolute roadblock; the probability of a correct response no longer depends on their target proficiency, effectively preventing them from demonstrating their underlying capabilities (Mislevy et al., 2013). Their chance of success drops to the baseline rate of random guessing, meaning the test yields absolutely zero valid information about their target capability. As Mislevy et al. (2013) note, "In other words, getting an item wrong due to lack of some [construct-irrelevant skill] ... is misleading evidence ... while in the second case it is apposite evidence."

This is where Universal Design for Learning (UDL) acts as the operational and psychometric engine of ECD. UDL is often misunderstood as a post-hoc checklist for making things "easier." The opposite is true. Within PADI's Design Patterns, developers identify *Characteristic Features* (things that must remain constant to measure the Focal KSA) and *Variable Task Features* (elements that can be adjusted to support Additional KSAs). UDL provides the exact knobs and dials to adjust these Variable Task Features to separate *desirable difficulties* from *undesirable difficulties*.

The rigor of the math or science concept is a desirable difficulty; we want the student to wrestle with it. Tiny font, convoluted syntax, and working-memory overload are undesirable difficulties. As the PADI/UDL integration research states explicitly, "...the goal of UDL is not to make assessments that are easier (e.g., by providing options that reduce the difficulty of relevant construct) but to make them more focused and accurate, largely by reducing the "undesirable difficulties" that are sources of error" (Rose, Murray, & Gravel, 2012, p. 8). We adjust variable task features to preserve the golden rule of alteration: "Specifically, if an alteration changes the construct, then construct validity has been violated. If the alteration does not change the construct, then construct validity has not been violated"(Zhang et al., 2010, p. 20).

By integrating ECD and UDL, assessment designers can strip away construct-irrelevant noise, allowing the focal signal of a student's internal knowledge representations and underlying capability to emerge with precision (Haertel, Haydel DeBarger, Cheng, et al., 2010; Seeratan & Mislevy, 2008). The critical question is whether this theoretical architecture can translate into research-validated, commercially viable tools utilized in real classrooms at scale.

The working examples detailed in Part II answer this question affirmatively. These case studies were seed funded over two decades by a federal innovation engine specifically designed to bridge rigorous research and commercial development - the Small Business Innovation Research program.

PART II: Working Examples

The US Department of Education and Institute of Education Sciences administers the Small Business Innovation Research Program (ED/IES SBIR), a highly competitive seed-funding initiative supporting early-stage R&D of research-based education technologies. The program funds high-risk, high-social-impact innovations, requiring projects to pair rigorous evidence building with credible commercialization pathways. (Metz & Tucker, 2025)

Historically, federal R&D funding dedicated to special education has been limited or unavailable. Simultaneously, innovations for low-incidence populations serve smaller markets with longer validation cycles, making them unattractive to venture capital. This creates a dual constraint: limited public funding and scarce private investment. For many developers focused on special education and accessibility, early-stage R&D without federal support is not simply difficult; it is economically implausible.

Despite these structural barriers, over more than two decades the ED/IES SBIR program invested in a small but consequential portfolio of inclusive, assistive, UDL- and ECD-informed innovations designed from inception to reduce construct-irrelevant barriers and expand access. Seeded with modest federal awards, many of these projects have since scaled nationally, validating a central tenet of this paper: designing for the margins strengthens the center. The case studies that follow illustrate that trajectory.

Assistive & Inclusive Technologies Designed For Students With Learning Differences And Disabilities

Alchemie

> ***Kasi proves that the best accessible design isn't a workaround, it's just good design, delivering belonging and mastery, not just accommodation.***

Kasi (Finnish for "hand") is an assistive hands-on chemistry learning intervention that combines physical molecular models with a machine-learning powered digital companion. Using computer vision, the system recognizes students' tactile interactions with proprietary chemistry manipulatives in real time and provides audio hints to support conceptual understanding. The platform embodies universal design for learning principles, benefiting all students while providing full accessibility support for students with visual impairments and other learning differences. Kasi was developed through awards from ED/IES SBIR, and led to follow-on awards from NSF, NIH, and NIDILRR SBIR to extend the product line. Pilot research demonstrated that blind and visually impaired students learned as effectively with the Kasi system as with a human guide, with subsequent research showing meaningful gains in agency, engagement, and sense of belonging. The product line is being adopted at scale by secondary and postsecondary programs through partnerships with leading courseware and assessment platforms. Alchemie was showcased at the White House Demo Day in 2023, and was named as one of 5 Big Ideas at the Disability Innovation Forum in 2025.

NimbleTools

NimbleTools reframed what accessible testing means: not a separate process for students with disabilities, but a single system designed to work for everyone from the beginning.

NimbleTools was a universally designed assessment delivery system that integrated accommodation supports directly into standard test administrations to meet the needs of students with a range of disabilities. The system combined specialized hardware and software, including custom keyboard overlays for students with sensory–motor difficulties, read-aloud supports for students with reading challenges, magnification of text and images for students with visual impairments, and an onscreen avatar that presented assessment content in American Sign Language or Signed English for students who are deaf or hard of hearing. Initial versions of NimbleTools were developed through academic research supported by the Institute of Education Sciences and the commercial version was subsequently developed through ED/IES SBIR Program to enable use in operational assessment settings. Pilot studies with more than 2,600 students demonstrated the feasibility and validity of the tools for meeting documented accommodation needs. In September 2010, Nimble Assessment Systems was acquired by Measured Progress, and the platform was subsequently used for test delivery in the New England Common Assessment Program (NECAP), including Rhode Island's statewide assessments.

IQ Sonics

A research-backed, music-based speech intervention that makes early language learning engaging, multisensory, and personalized for every child.

Sing and Speak 4 Kids is a music-based language intervention designed to support early speech and language development for young children, including those with speech delays and language disorders. The program uses interactive musical games and exercises to engage preschool-aged children while complementing traditional speech-language therapy practices. It integrates multisensory audio-visual stimuli with structured, developmentally appropriate practice to strengthen early expressive language skills across a range of developmental profiles. The program was initially developed with support from ED/IES SBIR, with pilot research demonstrating improved speech language skills among children ages 2–6 with autism, intellectual disability, and dual language learner backgrounds compared to a control group that did not use the program. To date, more than 1,000 educators and parents have used Sing and Speak 4 Kids. Program development continues through a 2025 ED/IES SBIR project with a new AI-assisted feature Make Your Own Song, helping personalize the lessons for each child's interest and environment.

Attainment Company

Significant disabilities shouldn't mean significantly lower expectations for literacy.

Early Reading Skills Builder and Access: Language Arts are blended online literacy programs designed for students with intellectual disability or autism. The programs provide systematic instruction in foundational reading, comprehension, and writing using structured lessons, adapted texts with professional narration, and instructional strategies such as constant time delay, modeling, least intrusive prompting, and built-in progress monitoring. Early Reading Skills Builder focuses on phonemic awareness, decoding, sight words, and connected text, while Access: Language Arts supports secondary students with grade-aligned adapted literature and scaffolded instruction across Bloom's Taxonomy. Both products were developed through iterative research supported by the ED/IES SBIR. Single-subject studies and randomized controlled trials showed that students using Early Reading Skills Builder and Access: Language Arts demonstrated significantly greater gains in reading skills and comprehension than students in comparison conditions. Following SBIR-funded development, both programs were commercially scaled by Attainment and are now used by hundreds of school districts nationwide to support literacy instruction for students with significant disabilities.

IDRT

Breaking down the language barrier in testing, myASL Quizmaker delivers valid, ASL-first assessments so deaf and hard-of-hearing learners can demonstrate what they truly know.

myASL Quizmaker is an assessment and practice tool designed to enable equitable testing for deaf, hard-of-hearing, and other learners who rely on American Sign Language (ASL) for communication. The platform enables educators, psychologists, and other practitioners to create customized tests, exams, and quizzes with automatic ASL video and graphic translations, administer assessments, and generate scored results, statistical analyses, reports, and graded feedback for individual students. myASL Quizmaker is explicitly designed to address the limitations of assessments delivered solely in written or spoken English by supporting ASL as the primary language of access. By embracing ASL as a distinct visual-gestural language with its own grammar and syntax, the platform reduces the construct-irrelevant barriers that obscure cognitive mastery when assessments default to English literacy. Supported by ED/IES SBIR, research demonstrated that educators successfully implemented the program to deliver valid assessments to students who use ASL. myASL Quizmaker is part of the myASLTech software suite and has been used by hundreds of professionals and families in the United States and Canada for the past 15 years.

ObjectiveED

BuddyBooks doesn't lower the bar for students with learning disabilities, it removes the barriers that were keeping them from clearing it.

BuddyBooks is an enhanced literacy platform for students in grades 2–8 with learning disabilities, including dyslexia, autism, and other language-based challenges, as well as persistent reading difficulties. The platform accommodates for learning differences through a structured system of built-in scaffolds, alternative representations, and personalized pacing aligned to individual learner profiles. These design features are operationalized through adjustable text presentation, multimodal audio-visual supports, embedded comprehension checks, and adaptive feedback that reduce cognitive load, support executive functioning, and enable students to engage with grade-level content despite underlying decoding or language challenges. BuddyBooks was developed and evaluated in part through SBIR programs at ED/IES, NSF, NIH, and NIDILRR. A pilot study with 1,250 students across 20 schools over multiple semesters demonstrated pre-to-post gains in fluency, decoding accuracy, reading duration, and reading comprehension. The product is commercially deployed within RTI and MTSS frameworks in hundreds of schools and by thousands of homeschool families, receiving recognitions such as the QS Reimagine Education Bronze Award for AI in Education.

Presence

No qualified therapist nearby shouldn't mean no therapy at all. Presence brings evidence-based speech, OT, behavioral, and mental health services to students wherever they are.

Presence is a teletherapy platform that enables schools to deliver speech-language, occupational, behavioral, mental health, and psychological services to students with diverse needs through secure remote sessions. The platform was purpose-built for teletherapy and teleassessment delivery and case management, allowing licensed clinicians to provide services to students regardless of location. The platform includes interactive tools to accommodate for use by students with learning difficulties, communication disorders, and other developmental or behavioral challenges with learning differences, including shared whiteboards, visual supports, screen sharing, and structured session workflows. An early version of Presence was built through ED/IES SBIR, with pilot studies and a randomized controlled trial demonstrating that remote speech therapy delivered via the platform produced outcomes equivalent to in-person services for children with speech sound disabilities. Today, Presence is used by thousands of schools and districts across the United States to expand access to evidence-based services, improve continuity of care, and address persistent shortages of qualified providers while supporting individualized, accessible service delivery for students with complex needs.

Accommodation Features For All Students, Including Those With Learning Differences

Teachley

Teachley was built on a simple idea: that every student, including those with learning differences, deserves math instruction that makes sense, not just practice until something sticks.

Teachley is an elementary math learning platform that builds conceptual understanding through virtual manipulatives, visual models, and interactive game-based problem-solving rather than procedural practice alone. The platform supports students in building number sense and fluency by making mathematical structures explicit and manipulable. Teachley was initially developed through ED/IES SBIR special education track award to design and test intervention software for improving single-digit math fact learning. From its earliest SBIR-funded development, Teachley incorporated built-in accessibility and accommodation supports, including adjustable pacing, multiple representations, and design features intended to support students with learning differences such as dyslexia, ADHD, and processing challenges. Research from SBIR-funded studies demonstrated that students using Teachley showed stronger gains in math fact learning and conceptual understanding than comparison students. Today, Teachley Apps have been used by more than one million students and the intervention is adopted by hundreds of schools.

Filament Games

PLEx meets middle schoolers at exactly the moment many check out of science, and brings them back in through games that are as accessible as they are engaging

The PLEx Life Science (Play, Learn, Experience) suite of learning games was designed to engage middle school students in core science concepts while accommodating diverse learning needs on topics such as photosynthesis, cell biology, heredity, human health, and engineering design. Targeting a period when many students disengage from science, particularly those who struggle with reading and traditional instruction, the games coupled gameplay mechanics with learning objectives to support meaningful learning. The games were developed through a special education award from ED/IES SBIR and were explicitly informed by Universal Design for Learning principles. Built-in accommodation features included multiple means of representation and engagement, in-game glossaries, optional voice-over for all text, and scaffolded supports designed to improve accessibility for students with reading difficulties and other learning differences. SBIR-funded research and field testing with hundreds of middle school students demonstrated higher levels of engagement compared to traditional instruction alone and that gameplay was associated with learning gains. The games were used more than one million times as part of classroom instruction and supplemental science learning. The games, including Reach For the Sun and You Make Me Sick, won many national industry awards for innovation.

OKO Labs

OKO reframes math class, less about speed and right answers, more about thinking out loud, struggling productively, and discovering you're more capable than you thought.

OKO is a collaborative, game-based math platform for students in grades 3–10 that supports small-group problem solving and mathematical discourse. The platform is designed to make student thinking visible, providing AI-powered facilitation that helps groups work through challenges together. Rather than emphasizing speed or a single correct pathway, OKO focuses on the collaborative process—encouraging persistence, discussion, and sense-making as students work through increasingly complex problems. OKO's design is particularly supportive of students with learning differences and math anxiety. By incorporating insights from The Math Narrative Project, OKO's facilitation helps students develop a positive math identity and reduces anxiety by normalizing struggle and providing repeated opportunities to build confidence. For example, OKO uses voice interactions to help students express their reasoning in multiple ways, reducing the reliance on written responses alone. An educator dashboard enables teachers to monitor group and individual progress along both academic and durable skill dimensions, such as collaboration, communication, and critical thinking. OKO was developed through ED/IES SBIR, with pilot research demonstrating that students who used the program increased in math proficiency while experiencing less math anxiety. OKO is currently used in dozens of school systems across the country.

CAPTI

What if a reading assessment actually reflected the student taking it? Capti ReadBasix makes that the standard, not the exception.

Capti ReadBasix is a literacy assessment and instruction platform that screens, diagnoses, and monitors reading students in grades 3–12, including students who struggle with literacy development, across six domains. Capti ReadBasix is used in thousands of schools across 24 states by more than 500,000 students. Capti is developing scenario-based assessments (SBAs) that are designed to accommodate learners from diverse linguistic, cultural, and educational backgrounds. Using NLP and generative AI, educators can create curriculum-aligned and scenario-based assessment items that are localized to students' lived experiences, instructional contexts, and community settings—reducing cultural bias while increasing relevance and engagement. A real-time dashboard supports educators in monitoring individual and group progress, implementing Response to Intervention and MTSS workflows, and tracking benchmark growth across the school year. Pilots show Capti SBAs are feasible and engaging: teachers can assign/proctor them effectively, students stay engaged, and 71% of students say having a clear purpose through the assessment helped them do their best. Capti's products are supported by funding from ED/IES SBIR.

Education Modified

An IEP is only as good as its implementation. Education Modified bridges the gap between the plan on paper and what happens in the classroom every day.

Education Modified is a student support platform, powered by ethical AI, that equips all teachers and administrators to implement and monitor Individualized Education Programs (IEPs), or other support plans for students with diverse learning needs. The EdMod platform employs AI to synthesize complex student data—including goals, services, accommodations, and progress monitoring—empowering educators to translate individualized plans into daily, IDEA-aligned instructional practices. The platform features were initially developed through ED/IES SBIR, with pilot research demonstrating that it improved educators ability to monitor goals, as well as manage and adjust practices based on student information. Subsequent development focused on integrating multiple data sources, supporting evidence-based instructional recommendations, and improving administrative oversight. Today, Education Modified is used by over 400 schools across the US.

Part III: Scaling The Future

Scaling Conditional Inference

For decades, the theoretical insights of PADI, ECD, and UDL faced a stubborn resource bottleneck: manually creating conditional task variants for every learner profile was labor-intensive and difficult to scale. The field understood Conditional Inference but lacked scalable delivery mechanisms; despite possessing formal task templates and modular measurement blocks, assessment assembly remained a largely manual endeavor (Hamel & Schank, 2005; Riconscente et al., 2005).

The convergence of artificial intelligence with these frameworks shatters that constraint, positioning AI as the engine that makes Conditional Inference scalable. Through generative and multimodal AI, modern assessment platforms can dynamically adjust Variable Task Features, rendering the exact same psychometric construct across diverse interfaces tailored to the learner's needs. AI might simplify syntax for an English Language Learner, provide on-demand audio scaffolding for a student with dyslexia, or use computer vision to track physical manipulatives for a visually impaired student—all while keeping the *Focal KSA* intact.

However, the immense power of AI brings new threats, including the hallucination of validity. Large Language Models do not inherently understand construct validity. If an educator simply prompts a generic AI to 'make this 8th-grade science test accessible for a 4th-grade reading level,' the model is equally likely to dilute the Focal KSA (the core scientific phenomenon) as it is to properly modify the Additional KSA (simplifying the linguistic syntax). In doing so, the AI risks erasing the Signal along with the Noise. As Mislevy noted, a student who can work with Newton's laws but can't figure out the mechanics of the simulation gives 'false negative' misleading information (Baxter & Mislevy, 2005). Without an ECD architecture acting as a noise-canceling filter, the AI risks erecting an absolute roadblock.

This is why the transparency of the chain of reasoning is paramount, and why PADI Design Patterns must serve as a safeguard for the AI era. These design patterns provide the ontological rulebooks—the literal "system prompts"—that AI agents require to ensure the target construct is not diluted. As the PADI framework anticipated years before the AI boom: "A design pattern can be the basis of principled, even algorithmic, generation of tasks, so that the ideas of UDL adaption... can be incorporated into systems that generate items in real time" (Mislevy et al., 2013, p. 15). AI assessment without the affordances of ECD is fast, expensive noise.

Procurement, Policy, and the Margins

As state education agencies and school districts procure the next generation of assessments, policy requirements must align with measurement science by mandating that EdTech vendors build their platforms upon rigorous ECD and UDL blueprints.(Haertel, Haydel DeBarger, Villalba, et al., 2010). Innovation cannot just be a flashy new dashboard or a conversational chatbot; it must be grounded in principled evidentiary arguments. If an AI-generated assessment is biased or inaccessible, an integrated design framework and the ECD layered approach allow policymakers to trace exactly which layer the error occurred in, creating auditable accountability in assessment design (Baxter & Mislevy, 2005).

Prioritizing upfront design over retrofitting is an economic and moral imperative. Attempting to retrofit accommodations onto rigid, finalized AI tools is not only scientifically sloppy and legally perilous, but financially exorbitant. Consider television captioning: originally an expensive retrofit for the deaf community, it became a universally beneficial feature once designed directly into the hardware. Utilizing universal design as a proactive foundational layer builds better choices for everyone from the start (Rose, Murray, & Gravel, 2012).

In ECD terms, this reflects the Shearing Layers of Architecture. Modifying an assessment's user interface (the "Skin") shouldn't require dismantling the underlying psychometric construct (the "Structure"). Ultimately, the vanguard companies emerging from the ED/IES SBIR portfolio prove a fundamental reality of product design: designing for the margins secures the needs of the center. When we build technologies rigorous enough to isolate the Focal KSAs for students with severe physical, cognitive, or sensory disabilities, we inadvertently create frictionless, accurate measurement tools that benefit *all* learners.

Honoring the Vision

Just as the transition from DVDs to streaming revolutionized delivery without altering the fundamental art of storytelling, artificial intelligence does not change the science of evidentiary reasoning; rather, it finally provides the computational scale required to execute it (Tucker & Metz, 2025). By embedding Large Language Models within a formal delivery architecture, AI can manage presentation and evidence accumulation, automating the extraction of validity data from complex digital traces (Almond, Steinberg, & Mislevy, 2002).

As we architect the future of learning, we must return to Samuel Messick's reminder, a quote reverentially echoed throughout Bob Mislevy's life work: "Validity, reliability, comparability, and fairness are not just measurement issues, but social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made" (Messick, 1989, p. 13).

In the AI era, Evidence-Centered Design and Universal Design for Learning are the architectural safeguards that help to ensure our algorithms serve our highest social values. By explicitly untangling the signal from the noise, we can begin to guarantee that every student, regardless of how they learn or develop, is provided an unclouded, focused lens through which to show us their brilliance.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. Journal of Technology, Learning, and Assessment, 1(5).

Baxter, G. P., & Mislevy, R. J. (2005). *The case for an integrated design framework for assessing science inquiry* (PADI Technical Report 5). SRI International.

Haertel, G., Haydel DeBarger, A., Cheng, B., Blackorby, J., Javitz, H., Ructtinger, L., Snow, E., Mislevy, R. J., Zhang, T., Murray, E., Gravel, J., Rose, D., Mitman Colker, A., & Hansen, E. G. (2010). *Using evidence-centered design and universal design for learning to design science assessment tasks for students with disabilities* (Assessment for Students with Disabilities Technical Report 1). SRI International.

Haertel, G., Haydel DeBarger, A., Villalba, S., Hamel, L., & Mitman Colker, A. (2010). *Integration of evidence-centered design and universal design principles using PADI, an online assessment design system* (Assessment for Students with Disabilities Technical Report 3). SRI International.

Hamel, L., & Schank, P. (2005). *Participatory, example-based data modeling in PADI* (PADI Technical Report 4). SRI International.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

Metz, E., & Tucker, E. M. (2025). *From seed funding to scale: U.S. Department of Education and Institute of Education Sciences' Small Business Innovation Research (ED/IES SBIR) program impact analysis (2012–2022)*. The Study Group.

Mislevy, R. J., Haertel, G., Cheng, B. H., Rutstein, D., Vendlinski, T., Murray, E., Rose, D., Gravel, J., & Mitman Colker, A. (2013). *Conditional inferences related to focal and additional knowledge, skills, and abilities* (Assessment for Students with Disabilities Technical Report 5). SRI International.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). SRI International.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, structures, and terminology. In S. Downing & T. Haladyna (Eds.), Handbook of Test Development (pp. 61–90). Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G. D., & Penuel, W. R. (2003). *Leverage points for improving educational assessment* (PADI Technical Report 2). SRI International.

Riconscente, M. M., Mislevy, R. J., Hamel, L., & PADI Research Group. (2005). *An introduction to PADI task templates* (PADI Technical Report 3). SRI International.

Rose, D., Murray, E., & Gravel, J. (2012). *UDL and the PADI process: The foundation* (Assessment for Students with Disabilities Technical Report 4). SRI International.

Seeratan, K. L., & Mislevy, R. J. (2008). *Design patterns for assessing internal knowledge representations* (PADI Technical Report 22). SRI International.

Tucker, E., & Metz, E. (2025, December 3). Education's Netflix moment: A "yes, and" approach for assessment system innovation. *Getting Smart*. https://www.gettingsmart.com/2025/12/03/educations-netflix-moment-a-yes-and-approach-for-assessment-system-innovation/

Zhang, T., Mislevy, R. J., Haertel, G., Javitz, H., Murray, E., Gravel, J., & Hansen, E. G. (2010). *A design pattern for a spelling assessment for students with disabilities* (Assessment for Students with Disabilities Technical Report 2). SRI International.

About the authors

Eric M. Tucker is the President and CEO of the Study Group, which exists to advance the best of artificial intelligence, assessment, and data practice, technology, and policy. He has served as President of Equity by Design, Superintendent and Executive Director of Brooklyn Laboratory Charter Schools, CEO of Friends of Brooklyn LAB, Cofounder of Educating All Learners Alliance, Executive Director of InnovateEDU, director at the Federal Reserve Bank of New York, and Cofounder and Chief Academic Officer of the National Association for Urban Debate Leagues. As an entrepreneurial, strategic, and impact-focused leader, Eric has over 25 years of experience building catalytic partnerships in education, securing over $300 million of investments for enterprises and initiatives that have transformed outcomes for learners and educators. Eric has expertise in measurement and assessment system innovation, participatory and advanced R&D, analytics, and human infrastructures for improvement and co-edited *The Sage Handbook of Measurement*. He earned a doctorate and a masters of science in measurement sciences from the University of Oxford and bachelors degrees from Brown University. Eric served as an ETS MacArthur Foundation Fellow with the Gordon Commission on the Future of Assessment in Education. He served as a Senior Research Scientist at the University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Edward Metz is a developmental psychologist and education researcher, and he is currently a Visiting Scholar at the Digital Futures Institute at Columbia University's Teachers College. He previously served as the Program Manager for the U.S. Department of Education and the Institute of Education Sciences' Small Business Innovation Research program.

About the Study Group

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

Date of Publication
March 2026

Licensing

Copyright