

The Architecture of Inference: Appreciating Robert Mislevy's Evidence-Centered Design in the Age of AI

Eric M. Tucker



The Architecture of Inference: Appreciating Robert Mislevy's Evidence-Centered Design in the Age of AI

Eric M. Tucker

Robert “Bob” Mislevy often shared thought experiments. Imagine, he’d say, that an expert English-speaking chemist who is learning German and a native German undergraduate both take a chemistry test written in German. He’d ask: “If a test taker struggles writing an essay, is it their skill at chemistry or German?” That same low score might tell completely different stories: one of a language barrier, the other of an actual knowledge gap. With that parable, he’d raise his point: any assessment score must be a *defensible story* about a learner, inseparable from their context.

Mislevy’s most enduring contribution was not a static product but a dynamic process, a blueprint for how professionals might solve problems too complex for any single person. This process was formalized in his framework, [Evidence-Centered Design](#) (ECD), a formal ‘grammar’ that enables experts from disparate fields (e.g., psychometrics, design, AI) to reason together about what constitutes valid evidence of learning. In the age of AI, where blackbox systems are making consequential decisions about learners and workers, [evidentiary reasoning](#) approaches such as ECD are an essential architecture of inference.

Evidence-Centered Design

Mislevy was a storyteller who pointed to real-world teams—such as F-15 mechanics, EA game designers, dental clinicians—to reveal this architecture. He celebrated systems like [Hydrive](#), which trained F-15 mechanics, or the [Cisco Networking Academy](#) lab. These simulators didn’t rely on multiple-choice questions; they created dynamic, live portraits of skill. By logging decisions, correction, and sequence of choices, the systems inferred a trainee’s troubleshooting strategy. The inference was tethered directly to authentic work: the pathway the student took to repair the fault, not simply the final correct answer.

Lessons for the Age of AI

The rise of AI in education, from generative models that evaluate essays to adaptive platforms that scaffold learning, has created a crisis. Automated systems can now track many data points, but without an architecture of inference, those points are meaningless, potentially yielding scores that are difficult to explain or trust. ECD provides a framework for assessment designers. This framework yields three lessons for building assessment systems that are both powerful and trustworthy in the age of automation:

1. Measure Skills in Context

Mislevy insists that a skill is inseparable from its context. A low-stakes quiz on grammar is not the same as a high-stakes clinical diagnosis. ECD demands that we design tasks built from the authentic demands of the work itself, a *humane argument*.

This insight is alive in language tests: the Occupational English Test, which measures English braided with the messy genres of clinical practice (reading patient charts, parsing prescription notes), and the [Duolingo English Test](#) (DET). The DET is a digital-first adaptive measure of language proficiency used for higher education admissions.

It leverages AI extensively to measure integrated skills (like speaking and listening within a conversation) and reflects Mislevy's conviction that to know if someone can navigate a system, you let them navigate that system. The validity lies in the resonance of the task to the real world.

2. Inferential Pathways: Telemetry as Evidence

The most common limitation of traditional testing is measuring *if* a student found a solution, not *how* they found it. When applied to digital learning environments, ECD shifts the focus in this context: the complete sequence of actions, the pathway, becomes meaningful evidence.

This focus on telemetry powered the design of many early educational games and simulations. In Game-Based Assessment, such as the design of [SimCityEDU: Pollution Challenge!](#), a student could try an approach and watch the system—the economy or air quality—push back in real time. Mislevy called this a “live argument.” In these decision traces, the system's telemetry revealed the student's approach. ECD has inspired games by groups like GlassLab and newer initiatives across [Roblox](#), [Project Lead the Way](#), [PBS Kids](#), and [Save Patch](#), proving that context is the construct, so assessment belongs inside authentic activity.

This idea now drives modern digital platforms with formative insights. Platforms from [Khan Academy](#) and [Age of Learning](#) to [Carnegie Learning](#) and [Curriculum Associates](#) collect and analyze interaction data to provide real-time, skill-level insights that inform instruction and course correction.

3. Artifacts Make Assumptions Transparent

For learning that involves complex creation (such as art, writing, or scientific inquiry) the evidence is the artifact itself. The challenge is translating a personal creation into a fair, shared evidentiary claim.

Mislevy celebrated [AP Art and Design](#). Here, the challenge was translating hundreds of personal studio hours (textured by charcoal, clay, and light) into a common standard. The “miracle” was the rubric that artists, educators, technologists, and psychometricians built together. It became the bridge turning a student's creativity into a shared claim, allowing raters to make inferences without sanding off the very edges that made the art unique.

Similarly, [Learning Maps](#)—visual illustrations of the relationships among knowledge and skills—become the shared artifact. Each node represents a specific concept, probabilistically linked to precursor skills. Crucially, they provide a common, shared language and pathway for assessing progress.

The Breadth of ECD's Legacy

Mislevy's influence extends far beyond the examples described above. ECD is a foundation for:

- International assessments like the [OECD's PISA](#) exam.
- U.S. Department of Defense [military training simulations](#) (CRESST).
- Professional certification and licensing exams such as [Cisco Networking Academy](#) Certification Exams to [Skills for the Future](#).
- Formative assessments aligned with the [Next Generation Science Standards](#).

In a career dedicated to crafting defensible stories about learning, Mislevy gifted our field intricate statistical models built upon simple parables. He demonstrated that the most powerful models we have are those that connect evidence to claims with rigor and precision. Mislevy's legacy is not a monument to be admired, but a practice to be inhabited: a clear, collaborative challenge to solve the next tough problem together, centered on the reliable architecture of inference.

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).

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

June 2026

Licensing

This case study is available under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International ([CC BY-NC-ND](https://creativecommons.org/licenses/by-nc-nd/4.0/)) license.

Copyright

The Architecture of Inference: Appreciating Robert Mislevy's Evidence-Centered Design in the Age of AI © 2026 Eric Tucker