CASE STUDY

A Hint is Not a Diploma: Consequential Educational Decision Making in the Multimodal AI Era

Eric M. Tucker

A Hint is Not a Diploma: Consequential Educational Decision Making in the Multimodal AI Era

Eric M. Tucker

What happens to the learner if an AI-powered decision is wrong?

Education is trapped in a false AI binary: heavily restrict it to prevent systemic harm, or blindly embrace it to avoid leaving learners behind. This obscures a deeper risk: the unprecedented rush to integrate AI without an evidentiary framework. As Ercikan and Solano-Flores (2025) warn, "Rapid adoption of innovative sources of data in assessment without proper support of empirical evidence on validity and fairness may compromise score meaning interpretation."

Both reactions err by treating "educational AI" as a monolith. The impending fall release of a review draft of the Standards for Educational and Psychological Testing offers a generational opportunity to address this. An adaptive spellchecker, a dropout dashboard, and an automated graduation screener are not ethically, legally, or pedagogically interchangeable simply because they share underlying code.

To govern education AI responsibly, we must stop fixating on the code's novelty and focus on the consequences of its decisions. As Kane (2016) observed, "To make consequential decisions without bothering to consider the consequences of these decisions would seem to be irresponsible."

As a parent of two school-age children who learn differently, this isn't theoretical. I am fiercely protective when unfounded assumptions or low expectations shape decisions about my son or daughter, and feel palpable relief when educators demonstrate they genuinely know them. I trust decisions rooted in human connection.

I welcome AI managing brief math practice. An algorithm noticing my child struggling with fractions and reinforcing a precursor skill feels reasonable; instructionally sound and reversible. But I would fiercely object if an opaque algorithm quietly decided whether my child was retained, tracked out of Algebra, denied special education, or blocked from graduation. A hint is not a diploma.

Let the Algorithm Hold the Flashlight, Not the Gavel

To operationalize this distinction, consider the "Netflix vs. Mortgage" test. Society happily accepts Spotify playlists or Google Maps routing because algorithmic error is reversible. Conversely, we reject "black-box" algorithms denying mortgages, screening resumes, or triaging patients. We instinctively recognize these decisions alter life chances and defy easy recourse: we accept automated navigation, but reject algorithmic redlining.

Across lending, healthcare, and hiring, convenience justifies automation, but consequence justifies safeguards. Reversible nudges require basic monitoring; irreversible judgments require due process.

Education sits on this same continuum. Because assessment and credentialing systems inform society's architecture for distributing future opportunity, the entire consequence spectrum—from harmless nudges to life-altering gatekeeping—coexists within single school buildings. This demands a consequence-calibrated framework.

Tier 1: Low Consequence (Assistive & Reversible)

At the continuum's base are everyday instructional supports: AI providing practice items, adaptive hints, and formative feedback. Because these are advisory and easily overridden, the cost of algorithmic error is negligible. Governance here is simple: let responsible innovation thrive. Low-stakes personalization is a suggestion engine; high-stakes classification is a gatekeeper.

Tier 2: Moderate Consequence (Decision-Support & Advisory)

The middle tier encompasses tools informing resource allocation: placement recommendations, dropout dashboards, and intervention referrals. Though lacking legal finality, they remain highly consequential; silently shaping teacher expectations, directing scarce resources, and creating self-fulfilling pathways. These systems require documented uses, subgroup bias monitoring, clear educator notices, and crucially, human-in-the-loop review. Teachers should use algorithmic alerts as formative hypotheses, never as unreviewable classifications.

Tier 3: High Consequence (Gatekeeping & Allocation)

At the top of the continuum are AI systems that act as unyielding gatekeepers, materially influencing life chances, including: grade promotion, graduation eligibility, special education identification, college admissions, scholarship allocation, and automated test proctoring or misconduct accusations. Errors here are opaque, devastating, and alter trajectories without immediate recourse.

Governance at this tier requires maximum scrutiny. High-consequence AI demands robust validity evidence tied directly to the specific decision, subgroup fairness audits, documented alternatives, and meaningful human appeal. Indeed, Bennett, LaMar, and Mazzeo (2025) contend, “our efforts to build a strong validity argument … should ramp up as the consequences associated with test results increase."

This consequence-tiered approach is no radical rupture; educational measurement possesses this scientific DNA. The Standards explicitly stipulate that evidence-based inferences are use-relative. As Samuel Messick(1989) and Michael Kane (2016) established, evidence validating a low-stakes hint doesn't automatically validate a high-stakes graduation decision. To vendors claiming abstract validity, we must bind regulatory norms to real-world accountability, recognizing that “validity refers to score decisions and test uses including their consequences” (Lane & Marion, 2025). Fairness is not an optional supplement; it is the core condition of test justification.

We can build a practical operational blueprint by bridging the domain safeguards of modern AI law—like the NIST AI Risk Management Framework and the EU AI Act—with the evidentiary discipline of psychometrics. “High-Risk Context” translates to assessment stakes. “Explainable AI” maps directly to clear score reporting. “Bias Audits” align seamlessly with subgroup-bias analyses.

Crucially, we must guard against 'ceremonial' human oversight. The European Data Protection Supervisor defines meaningful oversight as active involvement that "improves the quality of the decisions taken by the system, rather than as a merely procedural formality, or symbolic gesture." A rushed educator rubber-stamping a machine's output isn't oversight; it's automation bias in disguise. As Lane and Marion (2025) explain, mismatches between intended and enacted uses threaten decision validity, causing "unintended social and personal consequences." Meaningful review requires contextual knowledge, dedicated time, and ultimate authority to override the machine.

Taking the Recommendations of Machines on Trust?

As EU AI Act co-rapporteur Dragoş Tudorache observes, "We can't take the recommendations of a machine on trust. They have to be verified by a human supervisor."

Ultimately, a consequence-calibrated model is the only way to protect both learners and innovation. By concentrating heavy regulatory scrutiny only where potential harm is severe, we explicitly free educators and developers to experiment rapidly in low-stakes spaces.

AI possesses incredible potential to make learning personalized and engaging. Yet we will only reap those benefits, and maintain public trust, by explicitly distinguishing learning support from life-chance allocation. Proportionality protects innovation while safeguarding human potential. We must build an architecture of opportunity that lets the algorithm hold the flashlight, keeping the gavel firmly in human hands.

References

Bennett, R. E., LaMar, M., & Mazzeo, J. (2025). Technology-based assessment. In L. L. Cook & M. J. Pitoniak (Eds.), Educational measurement (5th ed.). Oxford University Press.

Ercikan, K., & Solano-Flores, G. (2025). The sociocultural context of educational assessment. In L. L. Cook & M. J. Pitoniak (Eds.), Educational measurement (5th ed.). Oxford University Press. https://doi.org/10.1093/oso/9780197654965.003.0003

European Data Protection Supervisor. (2025). Human oversight of automated decision-making (TechDispatch #2/2025). European Union. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2025-09-23-techdispatch-22025-human-oversight-automated-making_en

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Kane, M. T. (2016). Explicating validity. Assessment in Education: Principles, Policy & Practice, 23(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Lane, S., & Marion, S. F. (2025). Validity and validation. In L. L. Cook & M. J. Pitoniak (Eds.), Educational measurement (5th ed., pp. 191–275). Oxford University Press. https://doi.org/10.1093/oso/9780197654965.003.0004

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Macmillan.

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

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
August 2026

Citation
Tucker, E. (2026, July 8). *A Hint is Not a Diploma: Consequential Educational Decision Making in the Multimodal AI Era.* Getting Smart. https://www.gettingsmart.com/2026/07/08/a-hint-is-not-a-diploma-consequential-educational-decision-making-in-the-multimodal-ai-era/