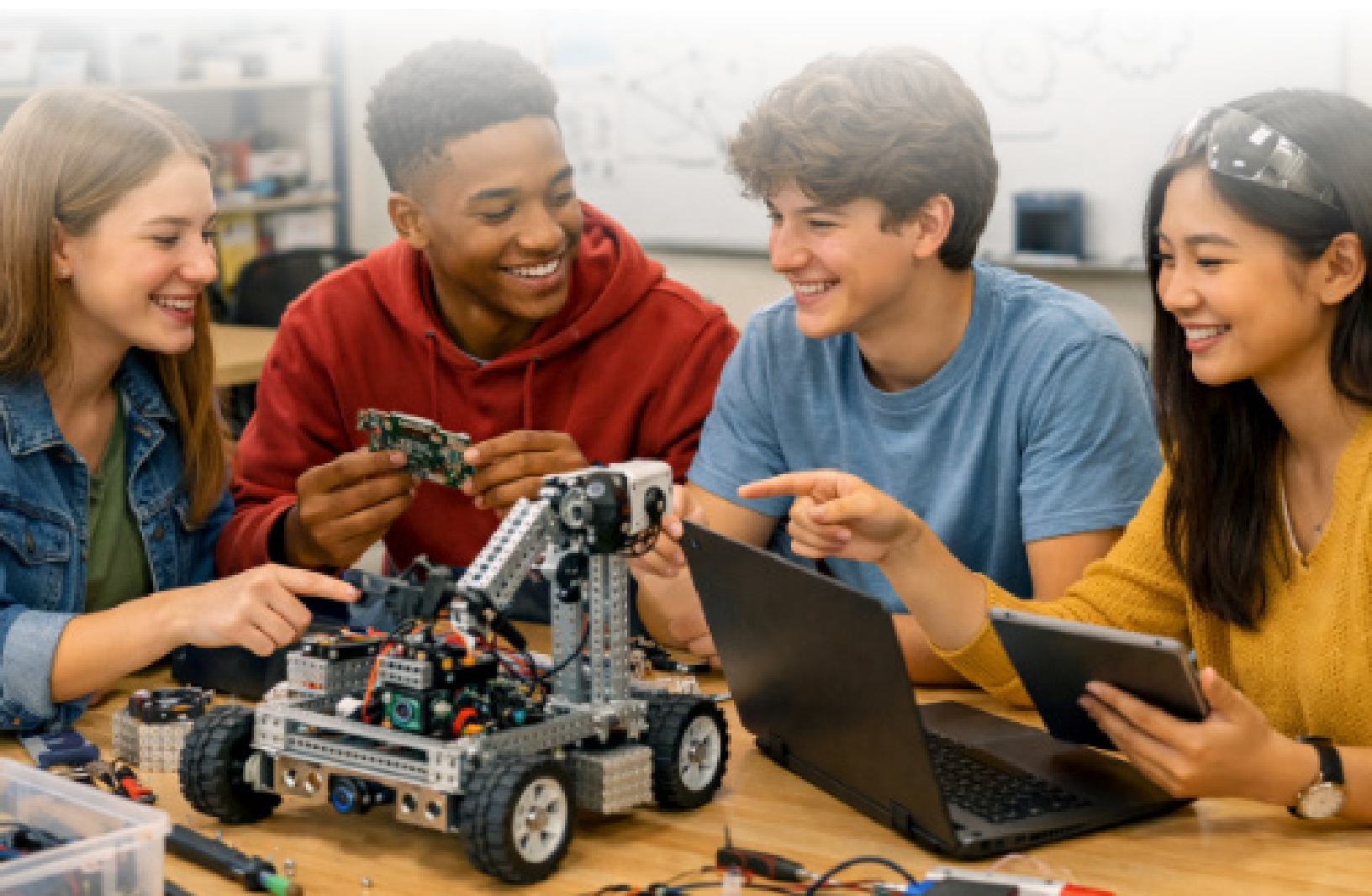


Beyond the Black Box: Why AI Needs to Show Its Work

Eric M. Tucker



Beyond the Black Box: Why AI Needs to Show Its Work

Eric M. Tucker

Artificial intelligence promises to transform how we measure human potential, but it too often demands our trust without showing its work.

Asking educators and families to rely on AI-powered inferences without transparency is like asking passengers to board a plane with a windowless, instrument-free cockpit. Yet, this 'black box' approach risks becoming the default, driven by proprietary models and a collective failure to prioritize [SAFE](#) guide rails specific to education.

Fortunately, this trend contradicts the trajectory of the measurement sciences, which are actively sharing their scientific source code by making resources such as [Educational Measurement \(5th edition\)](#) and the [Standards for Educational and Psychological Testing](#) open access.

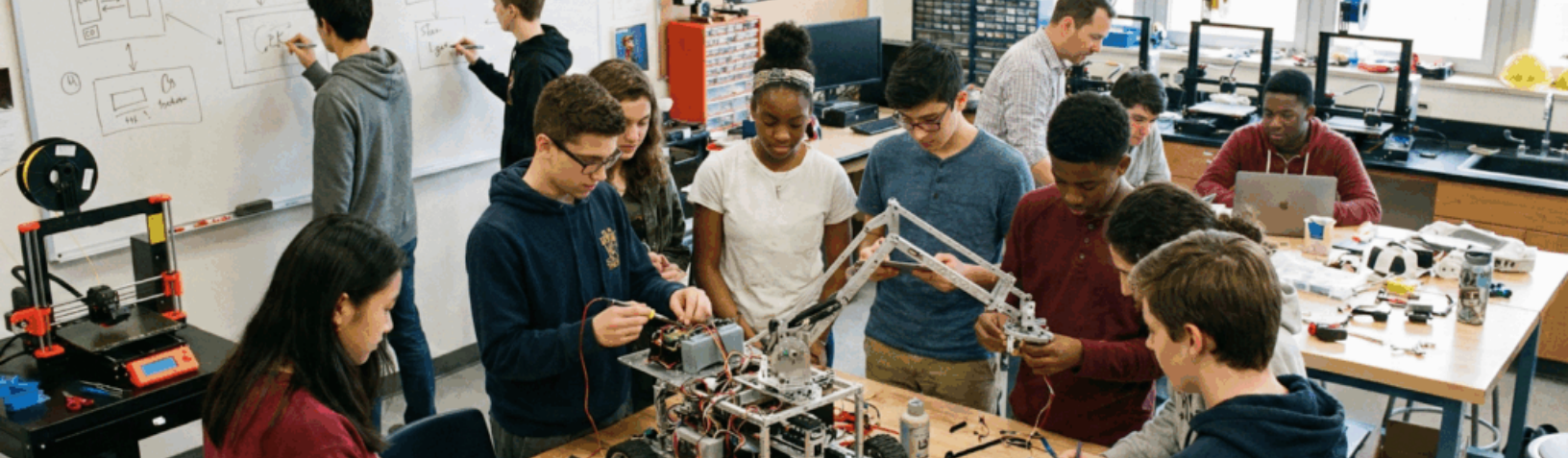
We cannot allow the promise of technology to obscure what the sciences of educational measurement are positioned to contribute. We face a choice: do we accept 'walled gardens' where AI increasingly acts as an undisclosed gatekeeper to opportunity, or do we build a 'digital public square' where design constraints are visible and explainable? Building credibility requires us to upgrade how we define quality for educational measurement in the AI era.

Scientific Soundness by Design

One allure of AI in testing is personalization: a sports fan might demonstrate math proficiency by analyzing player stats, while an astronomy enthusiast calculates differences across planets. This customization increases engagement but introduces a paradox: if every learner takes a different test, how do we know their scores mean the same thing?

This tension underscores a broader imperative: As we innovate what and how we measure, we must modernize how we understand scientific soundness—including validity, reliability, and fairness.

The Organisation for Economic Co-operation and Development's [Innovating Assessments to Measure and Support Complex Skills](#) argues that validity cannot be inspected at the end of the assembly line; it must be baked in from the start. Validity is no longer a mere static property of a test. It is a dynamic argument about a learner in context. A test is invalid if its results are misinterpreted or misused. If an AI-powered reading test generates a score that a teacher uses to pigeonhole a student, the integrity of the assessment process has failed. Validity is not just about accuracy; it is about the appropriateness of inferences we make and act upon.



Explainability is a Prerequisite for Trust

If scientific soundness is the engine (the machinery), transparency is the dashboard. We need to move from black box determinations to explainability. Learners have a right to understand how decisions about them are made.

Too often, a student receives a score—say, a 78 on an essay—without explanation. This runs counter to the learning science: a student does not [benefit from feedback](#) she or he does not understand.

Imagine if every AI-powered assessment came with a nutrition label. Just as we expect accurate information about the ingredients in our food, we must expect explicit evidence and reasoning regarding test design argumentation.

The [International Test Commission's Guidelines for Technology-Based Assessment](#) suggest that test sponsors explain the function of AI in plain language. Learners and families deserve to know what is measured, how tasks are scored, and the limitations of claims.

Fairness is Foundational

Every learner deserves a fair opportunity to demonstrate their ability, free from the disadvantages of 'construct irrelevant variance' such as linguistic background or digital literacy. Because AI systems inherit biases from training data, fairness matters from the start.

While technology offers critical accommodations, it can also introduce new barriers; for example, a speech-scoring AI must provide viable alternatives for deaf students. AI-powered assessments must first 'do no harm.'

As emphasized in the [Handbook for Assessment in the Service of Learning](#), we must build a validity argument proving that the test is not only accurate, but also safe, effective, and just.

Towards The Digital Public Square

We face a defining choice: a future governed by proprietary 'black boxes' with outsized impact on learners and workers, or a 'digital public square' where assessment design is transparent and vigorously debated.

Innovation without explainability is malpractice. As we integrate multi-modal AI into educational measurement, we must build upon bedrock principles of quality. And, we must demand that vendors show their work.

An assessment's value is not just in its accuracy, but in how useful the story is to learners and educators. It is time to ensure that the story of AI in education is one of openness, scientific rigor, and earned trust.

References

- Cook, L. L., & Pitoniak, M. J. (Eds.). (2025). Educational measurement (5th ed.). Oxford University Press.
- Foster, N., & Piacentini, M. (Eds.). (2023). Innovating assessments to measure and support complex skills. OECD Publishing. <https://doi.org/10.1787/e5f3e341-en>
- International Test Commission & Association of Test Publishers. (2025). Guidelines for technology-based assessment. <https://www.intestcom.org/page/28>
- Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024). Reimagining balanced assessment systems. National Academy of Education. <https://doi.org/10.31094/2024/1>
- Nasir, N. S., Lee, C. D., Pea, R., & McKinney de Royston, M. (Eds.). (2020). Handbook of the cultural foundations of learning. Routledge. <https://doi.org/10.4324/9780203774977>
- Sireci, S. G., Tucker, E. M., & Gordon, E. W. (Eds.). (2025). Handbook for assessment in the service of learning: Vol. 2. Reconceptualizing assessment to improve learning. University of Massachusetts Amherst Libraries. <https://doi.org/10.7275/ejm6-se46>
- Tucker, E. M., Armour-Thomas, E., & Gordon, E. W. (Eds.). (2025). Handbook for assessment in the service of learning: Vol. 1. Foundations for assessment in the service of learning. University of Massachusetts Amherst Libraries. <https://doi.org/10.7275/2h95-jf35>
- Tucker, E. M., Everson, H. T., Baker, E. L., & Gordon, E. W. (Eds.). (2025). Handbook for assessment in the service of learning: Vol. 3. Examples of assessment in the service of learning. University of Massachusetts Amherst Libraries. <https://doi.org/10.7275/s78z-y897>

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

Beyond the Black Box: Why AI Needs to Show Its Work © 2026 Eric Tucker