

We've Been Judging Symphony Rehearsals by the Last Note: Why Multimodal AI Can Finally Measure Mathematical Practice

Eric M. Tucker



We've Been Judging Symphony Rehearsals by the Last Note: Why Multimodal AI Can Finally Measure Mathematical Practice

Eric M. Tucker

Picture a sixth-grader solving a ratio problem. He sketches a table, erases it, and switches to a number line because it helps him 'see the quantities against each other.' His reasoning is sound, but a minor arithmetic slip ruins his final response.

Traditional tests mark this mathematical practice a zero, seeing only the final response. But an educator watching captures the student's active sense-making, strategic representation, and productive struggle.

For decades, assessments demanded static answers. But mathematical practices are about the 'moves.' They aren't decorative additions, they are the math. Multimodal AI has the potential to make this hidden reasoning visible. The reasoning is the main event.

Why Legacy Tests Fail the Math

Legacy tests were constrained by their era's technology and psychometrics. Measurement operated under an 'item paradigm' because a multiple choice response was a scalable way to evaluate students. One unintended consequence: effectively atomizing mathematics into isolated microstandards.

Treating math as a checklist of procedural items inhibits coherence; a tree is more than a [pile of twigs](#). By building assessment systems optimized for superficial legibility [rather than learning](#), we have essentially been judging symphony rehearsals by the last note. A static answer on a page is an incomplete witness to the full life of a mathematical practice. We must pivot away from reliance on endpoint-only testing and shift our focus toward [assessment in the service of learning](#), capturing the unfolding process of student cognition, rather than just auditing the procedural residue left behind.

Shrinking the Distance Between Action and Insight

Decades of learning and measurement science demonstrate that [process data](#) (fine-grained interactions like pauses, retries, and strategy shifts) [correlate](#) with actual numeracy outcomes.

What has fundamentally changed today is the interface and the time-to-evidence bearing ratio. Early digital assessments required long, constrained play sessions to yield a limited number of interpretable score points. Today, the advent of multimodal AI has the potential to shrink the [inferential distance](#) between a student's action and an educator's insight. Modern systems capture hand-drawn diagrams, spoken arguments, and strategic shifts in real-time, without interrupting the flow of learning.

Standards for Mathematical Practice

These [Standards for Mathematical Practice](#) represent the foundational habits of mind required to understand and perform mathematics.

- Make sense of problems and persevere in solving them.
- Reason abstractly and quantitatively.
- Construct viable arguments and critique the reasoning of others.
- Model with mathematics.
- Use appropriate tools strategically.
- Attend to precision.
- Look for and make use of structure.
- Look for and express regularity in repeated reasoning.

Making the Practices Visible

Multimodal AI operationalizes these practices by translating abstract mathematical verbs into concrete, visible evidence.

Take SMP 1 (Productive Struggle). Traditional tests only reveal failure, but multimodal AI tracks attempt sequences to distinguish productive struggle from aimless, frustrated 'wheel-spinning'.

Next, consider SMP 3 (Argument and Critique). Mathematical reasoning is discursive; it is often spoken, gestured, and negotiated before it is written down. Speech recognition captures oral explanations in noisy classrooms, allowing AI-powered systems to evaluate a live, spoken defense. This offers a more authentic measure of argumentation and oral reasoning that prevents us from mistaking a notation gap for a reasoning gap.

Finally, SMP 8 (Repeated Reasoning) involves abstracting shortcuts from patterns. Modern algorithms can detect the precise 'a-ha' moment when a student transitions from repetitive addition to a multiplicative formula. The vital event isn't the final number, but this leap to repeated reasoning.

Guardrails and Responsible AI

We must move away from opaque 'black-box' scoring by anchoring inferential systems in Evidence-Centered Design, while guarding against algorithmic biases that penalize human variation such as regional dialects. Frameworks like the [Duolingo English Test Responsible AI](#) framework offer an excellent approach for validating fairness, privacy, and transparency.

Crucially, multimodal AI must function as a teacher's 'noticing engine,' not a classroom judge. Rather than reducing students to a sterile '2.7 out of 5 on SMP 1,' dashboards should surface actionable alerts: who is ready to advance from addition to multiplication, or who excels orally but lacks precise written notation.

Towards Measurement of Mathematical Practices

As [Kristen Huff](#) observes, legacy assessments struggle to meet modern educational needs: our learning goals are increasingly complex, educators require versatile tools backed by built-in validation, and intense public scrutiny demands design rigor and clarity. If we expect students to master high-level reasoning, the infrastructure of educational measurement must match the sophistication of the minds it evaluates.

District leaders must demand tools prioritizing mathematical practices over digitized worksheets, while developers must reject cheap gamification for feedback loops and human governance. The future of assessment lies in richer evidence captured at the exact moment of thinking. It is time to transform the test score from a cold verdict into a meaningful explanation.

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 17). *We've been judging symphony rehearsals by the last note: Why multimodal AI can finally measure mathematical practice*. Getting Smart. <https://www.gettingsmart.com/2026/07/17/weve-been-judging-symphony-rehearsals-by-the-last-note-why-multimodal-ai-can-finally-measure-mathematical-practice/>

Licensing

This case study is available under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND) license.

Copyright

© 2026 Getting Smart. "We've Been Judging Symphony Rehearsals by the Last Note: Why Multimodal AI Can Finally Measure Mathematical Practice" by Eric Tucker, originally published July 17, 2026, at [gettingsmart.com](https://www.gettingsmart.com). All rights reserved.