Are We Assessing What Really Counts?

In math, educators have long been forced to choose between assessment depth and time constraints—but AI is changing that.

By **JAY MCTIGHE** *and* **JAY MEADOWS**

Consider these familiar refrains: "What we measure signals what we value" and "Assessment drives instruction." If these adages are true, they raise two essential questions: Are we truly assessing everything we value in mathematics? In what ways might we employ artificial intelligence (AI) to advance student assessments?

To explore these questions, let's begin by examining two sets of widely utilized math standards, which outline the desired learning goals of mathematics education. The standards published by the National Council of Teachers of Mathematics (NCTM) define two essential and interconnected categories: *content* standards and *process* standards.

Content standards specify mathematical skills, algorithms, procedures, formulas, and concepts students should master, progressively structured by grade level. They build systematically on previous knowledge, creating a solid foundation of mathematical competence.

Process standards outline the essential behaviors and cognitive practices necessary for doing mathematics well. These standards—problem solving, reasoning and proof, communication, connections, and representations—are consistent across all

grade levels and adaptable to various mathematical situations.

The Common Core State Standards (CCSS) echo this emphasis on connecting mathematical practices with math content. While the CCSS are less prominent at the national level than they were at their launch in 2010, they have clearly influenced the development of many state math standards, especially with their emphasis on this confluence:

> Designers of curricula, assessments, and professional development should all attend to the need to connect the mathematical practices to mathematical content in mathematics instruction. (NGA Center for Best Practices & CCSSO, 2010)

The message within the CCSS and NCTM standards is clear: Mathematics instruction should not only impart key mathematical facts and concepts, but also develop students' capacities to reason mathematically, employ effective strategies, create models, explain and justify solutions, communicate precisely using mathematical language, and even persevere when tackling "messy" real-world problems.

This raises the question: How well are we assessing these learning goals and how can AI help us bridge the gap between what we value and what we actually measure? Let's examine this in the context of both large-scale testing and classroom assessments.

The potential of AI to analyze any form of student work, evaluate it using high-quality rubrics, and deliver specific, actionable, and immediate feedback is limitless.

Assessing the Assessments

Large-scale assessments in the U.S., including state accountability tests and the National Assessment of Educational Progress (NAEP), typically employ a selected-response (S-R) item format. Their use of "multiple choice" is understandable given that more than 3 million public school students per grade level are routinely tested across the United States every year. This format allows for inexpensive machine scoring and quick turnaround times for reporting the results.

Selected-response items are well-suited to gauging computational accuracy. Moreover, well-constructed S-R test items that present word problems can involve multistep solutions and mathematical reasoning. However, for such items, a student's reasoning must be inferred based on the chosen answer, rather than tangibly observed.

Despite their virtues, selected-response tests have notable limitations. This format does not explicitly assess students' use of problem-solving strategies, the efficiency of their solution methods, or their ability to communicate their reasoning using mathematical vocabulary with any precision. Clearly, current large-scale assessments assess some, but not all, of what the standards call for.

What about math assessments at the classroom level? We have observed that these assessments often mimic the format of large-scale tests, in that they primarily consist of multiple-choice and short-answer items. Of course, some teachers ask students to "show their work," especially on word problems involving multiple steps, which gives the teacher some insight into a student's reasoning and use of strategies. Selected-response assessments are most widely used for two reasons:

1. *Teachers feel pressured to use formats that mirror high-stakes accountability tests.* Logic suggests that if students are being tested with selected-response items, and teachers and schools are being held accountable for the test results, students should have lots of practice in the tested format.

2. *Open-ended performance assessments require considerable time to score.* This is especially challenging for secondary teachers who may have daily

loads of 100–160 students, as well as for elementary teachers who teach and assess multiple subjects.

In sum, it is clear that our current assessment methods—both large-scale and classroom-based—are not assessing everything that the mathematics standards articulate.

Our Assessments Send Messages

One consequence of the predominant use of selected-response formats is the message this can unwittingly convey to students and teachers. For students, it may suggest that problem solving can be boiled down to choosing the "right" answer from four or five provided "options." For teachers, it may imply the need to use lots of test prep materials featuring canned problems requiring the formulaic "plug-in" of numbers into memorized algorithms (formulas).

These are *not* the messages that mathematics standards intend to convey.

Needed: Expanded Use of Authentic Performance Assessments

We believe that to more appropriately provide valid and reliable evidence of all mathematical outcomes, including reasoning, use of problem-solving strategies, modeling, and communication, math assessments must expand their format to include performance tasks that call for constructed responses.

Consider this analogy from *Assessing Student Learning by Design* (McTighe & Ferrara, 2021):

> Think of effective assessment as a photo album, with any single picture providing a "moment in time" display of what a student knows, understands, and can do. Since a photo album contains a number of pictures, taken over time in different contexts (close-up, wide angle, etc.), it provides a more accurate and revealing "portrait" of an individual than any single snapshot can provide. While multiple-choice items are a perfectly acceptable component of a balanced assessment album (i.e., one type of picture), other assessment forms are needed to provide additional evidence.

Authentic performance tasks provide the needed complement to selected-response test items. These types of tasks have specific qualities that make them ideal for measuring such outcomes (McTighe, Doubet, & Carbaugh, 2020). Authentic performance tasks:

- Provide evidence of learning the key math practices and math content contained in the mathematics standards
- Establish a realistic context, including a genuine problem or goal, a target audience, and genuine constraints (e.g., budget, schedule) that reflect authentic use of math concepts and procedures
- Are open-ended, involving multiple steps, addressing multiple content standards, and allowing for a variety of strategies to be employed
- Require reasoning and explanation, not simply computation or selection of an answer from given options
- Include criteria/rubric(s) for judging performance based on the targeted standard(s).

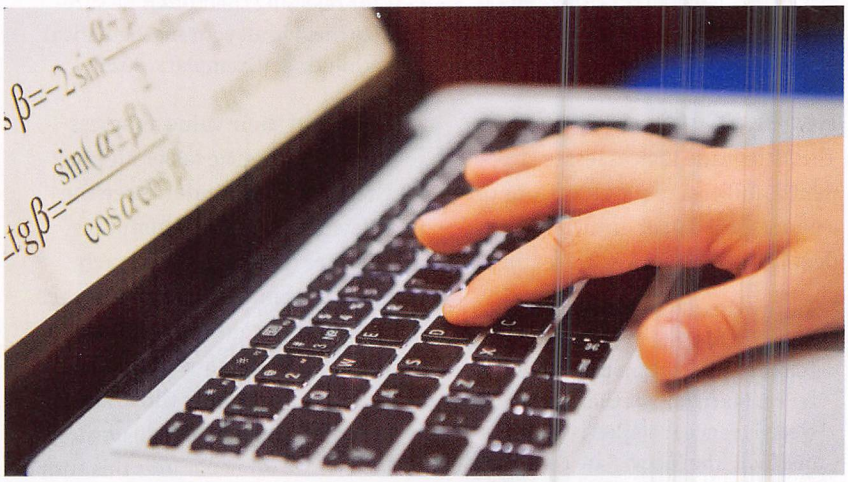
They may also:

- Naturally integrate one or more other subject areas with mathematics
- Purposefully incorporate the use of technology.

Following are two examples of such tasks.

Going to a Movie

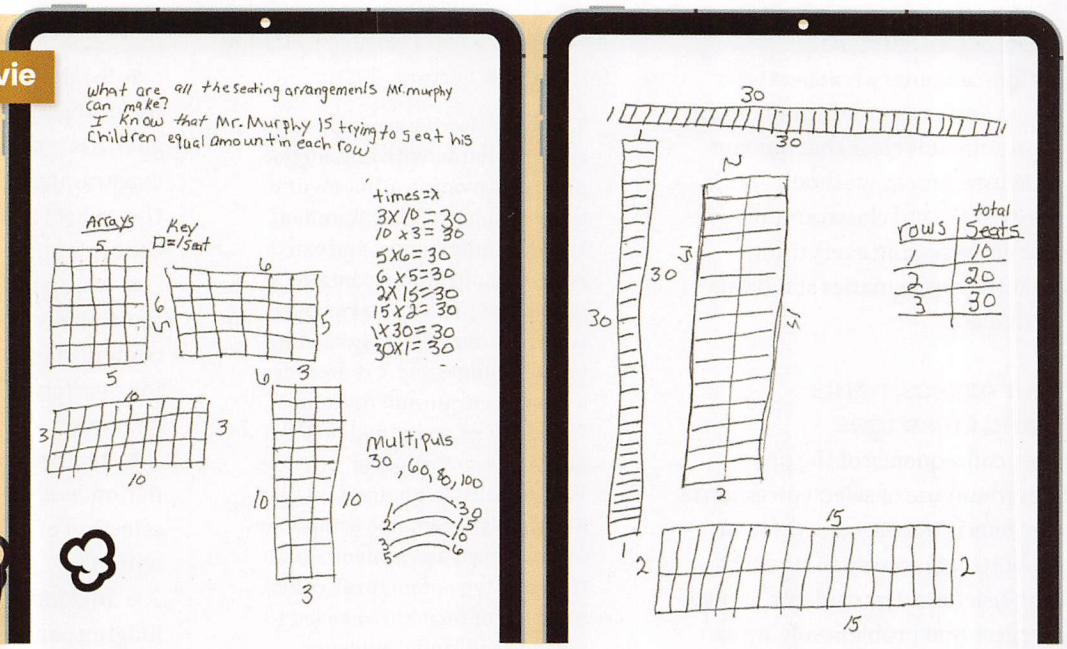
This 3rd grade task asks students to



NATALIA KIRSANOVA / ISTOCK

Going to a Movie

A 3rd grader demonstrates multiplication and division concepts through various movie theater seating arrangements.



Source: Exemplars

demonstrate their understanding of multiplication and division using both equations and drawings:

> Mr. Murphy is taking some students to see a movie. Mr. Murphy needs a seating arrangement for 30 students. Mr. Murphy wants an equal amount of students in each row. What are all the possible seating arrangements Mr. Murphy can make? Show all your mathematical thinking. (Exemplars, 2025)

Herding Cats

This 6th grade task assesses students' abilities to plot locations on a coordinated grid, multiply with decimals, and work with scaling:

> Maru owns a free-roam cat rescue. Maru is working to enclose her roaming area with a cat-safe electric fence. Maru knows the cost of an electric fence installed is $0.89 per foot. The Meow Safe Fencing Company has provided an estimated price of $3,190 for fencing in the total property.
>
> The boundaries for the property can be described using coordinates on a scaled grid overlaid on an aerial photo. Each unit is 30 feet. The boundary coordinates are (1, 0), (25, 0), (25, 22), (21, 22), (21, 16), (5, 16), (5, 22), and (1, 22).
>
> Write a letter to the Meow Safe Fencing Company either accepting or rejecting their offer. Be sure to include all your mathematical thinking. (Exemplars, 2025)

Beyond their value as assessments, authentic tasks such as these can help students see the relevance of the mathematics they are learning. Since such tasks present realistic challenges, they offer a genuine response to the often-heard questions from learners, especially those at the secondary level: *Why do we need to learn this? Whoever uses this stuff?* Well-constructed authentic performance tasks naturally reveal ways in which mathematics is employed beyond the classroom.

Performance Assessments Provide More Evidence About Valued Outcomes

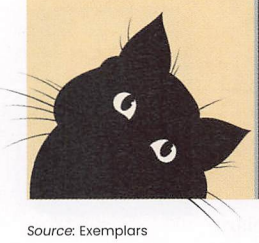
In contrast to selected-response items that simply ask students to choose an answer from several alternatives, performance tasks involve both application and explanation.

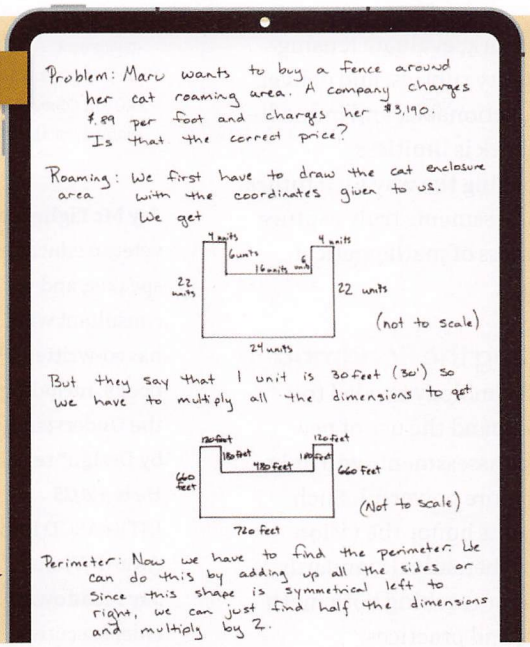
Look at the examples of student responses to the previously described performance tasks.

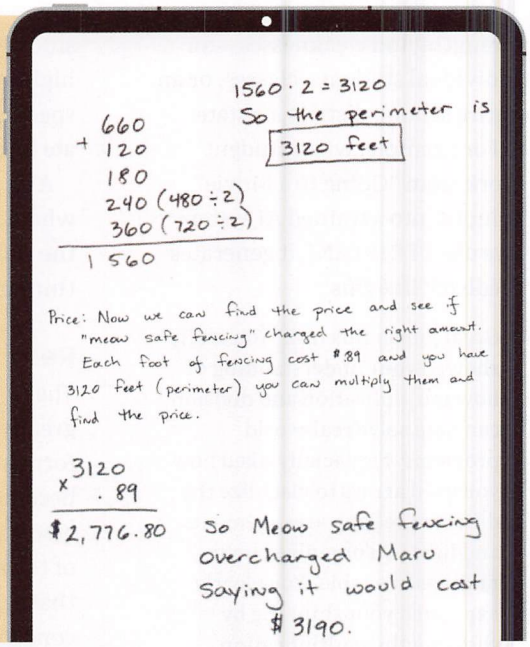
Notice in both examples, students are displaying their calculations, generating representations, and communicating their reasoning. More generally, in a mathematical performance task, a student is expected to use a model or representation aligned with the

Herding Cats

A 6th grader solves a real-world cat fencing problem through representations, calculations, and reasoning.







Source: Exemplars

task's mathematical concepts to organize information, demonstrate their thinking, and draw new conclusions. Students need to communicate their strategies using grade-appropriate vocabulary and present clear, organized calculations.

While we encourage the expanded use of performance tasks for assessment purposes, we also encourage the wider use of authentic tasks as a critical part of rich mathematical learning. When students tackle a multistep problem and choose the tools and strategies they need to solve it, they're empowered to think creatively and flexibly—and they grow their problem-solving skills.

Enter AI: A Game Changer

Recent advances in AI may finally resolve the long-standing tension between depth and practicality. Despite the value of authentic performance tasks, the amount of time these tasks take to evaluate can be prohibitive. For a teacher with 150 students, spending 10 minutes assessing the solutions of each student using a qualitative rubric and providing useful feedback would require 25 hours of work. For a 3rd grade teacher with 25 students, this would require over 4 hours of work. How can these tensions be addressed?

Artificial intelligence is already capable of various tasks like assessing essays, transcribing handwriting, and processing images. What if we could use AI to analyze student responses on open-ended math performance tasks while gaining the scalability of digital scoring?

Imagine a future where we no longer have to choose between speed and depth—where students can solve real problems, show their thinking, justify their reasoning, and receive timely, meaningful feedback at scale.

Well, that reality is already here. Rapidly expanding AI capabilities are currently being harnessed to:

- Examine student-constructed mathematical representations for accuracy and completeness
- Evaluate student reasoning and use of mathematical vocabulary while justifying their strategies and conclusions
- Analyze student calculations for correctness
- Efficiently and reliably score student-constructed responses to open-ended tasks
- Offer tailored instructional recommendations, including suggestions for differentiation, to address the specific performance areas
- Provide detailed analysis of

IGOR, DANIEL RODRIGUEZ / ADOBE STOCK

strengths and weaknesses—for individual students, classes, or an entire school, district, or state.

For example, when student work from "Going to a Movie" is input into a trained AI system (we used TRISTAN), it generates feedback like this:

> Great job on this task! You really showed your understanding of how multiplication and division can help solve real-world problems. I especially liked how you used arrays to visualize the different seating arrangements and how you organized your findings in a table. You clearly explained your thinking by showing the multiplication equations that match each arrangement. Here are a few ideas to help you grow as a mathematician:
>
> - Try to explain the connection between the multiplication facts and the seating arrangements. For example, you could say, "Since 3 x 10 = 30, we can have 3 rows of 10 seats."
> - Level up your math language by using words like "factors" and "product" when you talk about the numbers in the problem.
> - Think about how you could use what you know about factors to break down the number 30 in different ways. Can you find any prime numbers that are factors of 30?

With appropriate training, AI can also provide feedback on individual process skills such as reasoning and proof or communication, identifying specific areas for improvement or re-engagement. The potential of AI to analyze any form of student work, evaluate it using high-quality rubrics, and deliver specific, actionable, and immediate feedback is limitless.

AI is paving the way for a future where assessments truly capture the richness of mathematical thinking.

Rerouting the Roadmap

The AI technology needed to greatly expand the use of performance assessments will only become more powerful. Such assessments honor the vision of the mathematical standards that call for teaching both math concepts and practices.

We can now authentically assess learners' abilities to tackle rich problem-solving tasks and explain and justify their mathematical reasoning. We owe it to our students—and to ourselves—to expand the use of authentic math tasks that can engage students in meaningful learning. And with the power of AI at our fingertips, that future is closer than we think.

References

Exemplars. (2025). *Going to a movie.* Exemplars Library of Performance Tasks.

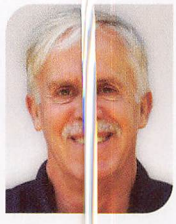
Exemplars. (2025). *Herding cats.* Exemplars Library of Performance Tasks.

McTighe, J., & Ferrara, S. (2021). *Assessing student learning by design.* Teachers College Press.

McTighe, J., & Wiggins, G. (2013, Spring). *From Common Core Standards to curriculum: Five big ideas.* Wisconsin ASCD Highlighter.

McTighe, J., Doubet, K., & Carbaugh, E. (2020). *Designing authentic tasks and projects: Tools for meaningful learning and assessment.* ASCD.

NGA Center for Best Practices & CCSSO. (2010). *Common core state standards (mathematical practice).*



Jay McTighe is a veteran educator, speaker, and consultant who has co-written 18 books, including the Understanding by Design® series. He is a 2025 ISTE+ASCD Impact Award Winner.

Jay Meadows is chief executive officer of Exemplars and a former middle school math and science teacher.

reflect + discuss

What percentage of your math assessments ask students to explain their thinking rather than just provide the right answer?

What messages might your assessment practices be sending to students about what "doing mathematics" really means?

If AI could handle the time-intensive scoring of open-ended math tasks, which authentic performance assessments would you most want to implement in your classroom or school?