A Proposal for a Major Overhaul of

High-Stakes Assessment of School Mathematics

Steven Leinwand, Mathematics Education Change Agent

Jay Meadows, CEO, Exemplars

In a sentence, 21st-century skills cannot be measured, nor developed, with primarily multiple-choice test questions. Diverse workplaces that today value communication, collaboration, reasoning, justification, critical thinking, creativity, and flexibility – all with the considerable support of technology – are neither supported by, nor aligned with, today's high-stakes mathematics assessments. A clear-eyed look at the assessment landscape reveals that current math assessments seriously undermine the teaching of mathematics, given that assessment content and format, and the instruction they drive, keep the primary focus on recall and regurgitation. Moreover, a society that is increasingly complex, nuanced, ambiguous, and technological is underserved by assessments that fail to ask questions such as "Why?", "Explain your reasoning.", "Justify your thinking", "How might you convince us?" "Can you solve this another way?"– all of which are under-emphasized in typical instruction because they distract from skills practice and test prep, and until recently have been difficult to economically assess on a large scale.

To a large degree, whether intended or unintended, what is tested determines what is taught and too often, today's assessments narrow instruction towards procedural fluency and short word problems, which often leads to bored and disengaged students. Conversely, and unfortunately, what isn't tested, despite its importance, is at best underemphasized. These realities must be addressed, and we believe that modernizing high-stakes state and national assessments is the perfect place to start. This is particularly important and timely given the range of new and powerful tools that can support far stronger assessments – assessments that are much better aligned with modern needs and how mathematics is actually performed in the real world.

While some argue that the high-stake tests merely reflect the curriculum standards that are in place, we argue that the increasingly important process standards – including communication, making viable arguments, justifying one's thinking, perseverance in solving problems, using multiple representations, constructing alternative solutions, and confronting unfamiliar situations – are seriously undervalued on most extant assessments, in large measure because they require constructing, administering and scoring constructed response items. How odd that the highly-honored capstone AP assessments all include open-ended, constructed-response tasks for which few students in non-AP courses have been adequately prepared, in part because of their notable absence on most large-scale assessments.

Moreover, most existing constructed response questions require students to perform on a digital platform that is often disconnected from their everyday, classroom experiences with mathematics and is also largely incongruent with how mathematicians engage in performing mathematics in most settings outside of the classroom.

Assessments worth doing should look and feel the same as how math is being done in the classroom. Truly valuable assessments should also look and feel like how math is used in the real world. What real-world setting asks a person to perform mathematics using drag-and-drop resources? When in daily life or the workplace is competence measured by multiple-choice items? How many students spend time typing up their thinking to mathematical problems, except in preparation for some state assessments? Our high-stakes assessments are the only place where a person is expected to engage mathematics in this way. Yet, these expectations drive huge percentages of our precious math classroom minutes to be spent learning how to navigate these tools.

How is this helping students become effective users of mathematics?

Accordingly, the case we make is straightforward:

1. Current tests and testing practices are out of sync with, and an impediment to, enabling schools to much better serve student, workplace, and societal needs

2. There is broad agreement about the deficiencies with current practice and the need for substantive improvement.

3. It is time for a new generation of summative assessments that drive instruction productively and better reflect essential future-oriented skills and competencies.

4. Newly available and emerging technology, especially aspects of Artificial Intelligence, are already in place and working well in a growing number of states across the country. These initiatives offer powerful improvements and efficiencies.

5. These changes are essential to support teachers and focus on what is needed by students throughout their lifetimes.

6. It is time for assessment specialists, psychometricians and mathematics educators to collaborate on the development and implementation of a new generation of high-stakes assessments.

We believe in assessment

We are not anti-testing. We firmly believe in the importance of assessment that supports meaningful curriculum and high-quality instruction, and reliably informs all relevant stakeholders. More particularly, we believe strongly that assessment is the third and co-equal leg of the plan-implement-evaluate stool that honors the unique and interdependent roles of curriculum, instruction and assessment.

We also believe that one important feature of mathematically high-performing countries, absent to a large degree in the United States, is a high level of alignment among curricular goals, instructional practices, and assessment content and formats. We do not believe that "teaching to the test" is cheating; rather, we advocate for tests worth teaching to.

We do not believe that we can have truly effective mathematics programs without high-quality, well-aligned assessments. Such assessments help us understand what worked, for whom, and how well. They provide essential feedback that enables adjustments in teaching, and informs curricular shifts. It is hard to imagine a sensible system of accountability that doesn't provide teachers, parents, schools, districts, states, and the nation with reliable, actionable, and

understandable information about the success, or lack thereof, of learning. But this information only has value when it measures what is truly important and in authentic ways that reflect changing societal and educational needs, as captured by underemphasized process standards that complement content standards.

The over-reliance on multiple-choice items

We understand that multiple-choice items have traditionally resulted in higher levels of psychometric validity and reliability. Unfortunately, these two measures come at the expense of authenticity and often measure knowledge in stilted ways where the item difficulty is as likely to emerge from the item distractors as from the actual cognitive complexity of the mathematics measured by the item.

We also understand that many mathematical skills, such as estimation, explanation, and justification, the depth of conceptual understanding, and the processes of problem-solving, are not adequately measured with multiple-choice items and are, therefore, underrepresented on most assessments. For example, all humans estimate quantities and measures, but since estimation is very difficult to assess with multiple-choice items, the development and authentic application of this critical skill often fall through the cracks in our schools. Similarly, it is easy to argue that patterning can be assessed by simply asking which one of four figures must be the next figure in this pattern. But such an approach narrows instruction and ignores the essential understandings of describing the pattern, identifying how it changes, actually drawing the next figure, and then generalizing the pattern - all scored with full or partial credit for full or partial work.

But there is another option. We believe that the rapidly advancing technology available today can assess and reliably score such tasks at far less cost than human scorers. Everyone who encountered Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced open-ended and performance tasks recognized their quality and relevance. But ten years ago, we just didn't have the tools we have today to make such assessment tasks that reflect authentic application of mathematics financially viable.

Let's be honest about what most students face on most state assessments today.

The list of problems we face as educators, and as a society, with current mathematics assessment practices is clear:

- There is a preponderance of items that assess procedures and symbol manipulation at the expense of understanding, explanation, and demonstration of more than mere answer-getting.
- There is an over-reliance on multiple-choice items.
- There is an over-emphasis on Depth of Knowledge Level 1 "recall" items at the expense of Depth of Knowledge Levels 2 and 3 "apply, solve, understand and justify" items best assessed in open-ended, constructed-response formats.
- There are large gaps in alignment between current assessments with the increasingly important "process standards" involving reasoning, communication, and authentic problem-solving.
- Despite the ubiquitous availability of Google, Photo Math, Alexa and now AI in the workplace and out of school, beyond selected use of calculators and Desmos, most current assessments ignore these technological realities and skew assessment away from the sensible and appropriate integration of technology into teaching and learning.
- In a world where nearly every survey of workplace needs and job requirements points to analytical thinking, communication, flexibility in solving problems and facility with technology, it is very hard to find even lip service paid to these domains on high-stakes assessments.
- While nearly every state and school district claims to address college- and career-readiness as an overarching goal, such a goal is a moving target as the world changes. But little has changed in the basic structure and content of large-scale, summative assessment over the entire span of post-No Child Left Behind assessments.

However, technology has brought great advances

In fact, great progress has been made in components of assessment, and this progress supports our belief that modernizing and making important additions to our assessment system is entirely possible.

- Online assessments. The majority of state and high-stakes national assessments are now administered online. This advancement enables a range of assessment item formats, reduces the expensive cost of printing and scanning, and significantly reduces the time needed to return assessment data to states, schools, districts, and teachers.
- Adaptive assessments. Highly sophisticated algorithms now use a small number of common items across a range of difficulty to narrow in on a given student's level of understanding, and then use a set of increasingly targeted items to narrow down on an accurate measure of a student's mathematical abilities. In addition to reducing the number of items needed for a reliable measure of achievement, multi-part open-ended items can easily be substituted for multiple-choice items on which guessing introduces serious measurement error.
- More teacher and parent-friendly reports. States and NAEP have wrestled for years with the usefulness and quality of information reported annually to schools, teachers and parents. However, there is a very broad range of report forms and included information that robs many parents and teachers of important information about their children and students that is available in other states. Combining AI tools and constructed-response items opens promising floodgates to far more informative reports.
- More accessible user interface. One of the challenges of constructed response questions is the requirement that students construct their answers using digital tools that are often unfamiliar to them. Student performance often depends as much on a student's ability to navigate the assessment platform and embedded tools as it does on the student's mathematical abilities. By incorporating assessment tools that look more like how mathematics is normally performed, with pencil and paper and hand-drawn representations, the more authentic the assessments can be.
- AI for scoring. Although the newest innovation available is not yet part of the contracts for many state assessments, AI is the future of strengthening item writing, adaptive assessments, assessment reports, and, perhaps most importantly, reliably and efficiently scoring student work on open-ended and constructed-response tasks.

Proposed changes in school mathematics assessment

Building on, and institutionalizing these important and consequential advancements, and in light of the many deficiencies with current practice, we propose a set of initiatives for overhauling high-stakes assessment of school mathematics.

We believe that schools and districts are drowning in assessments that take an extraordinarily large bite out of instruction time, with only minimal evidence, given test security, that assessment really improves teaching and learning. Accordingly, we strongly advocate for a range of changes to address these conditions:

- Consider moving all state assessments to the fall – late September or early October – to capture what students really know and can do after a summer off, as opposed to what is crammed in as part of wasteful test-prep that accompanies spring testing in April or May.
- Consider recognizing that fall assessment in Grades 4, 6, and 8 provides more than enough information and data to make wise decisions. Moving the first encounter with high-stakes stress-inducing assessment to grade 4 avoids prematurely branding 3rd graders as high flyers and underachievers. In addition, moving the first encounter with these assessments to grade 4 postpones the concerns about whether 3rd graders are developmentally ready for online and open-ended assessments. And once again, instead of spending millions of dollars on annual high-stakes assessments, states and districts would be able to redirect assessment savings to instructional support that responds to the test results.
- Consider revamping test designs to ensure far more open-ended, constructed response items with a focus on justification, use of graphical representation and explanation, scored with AI, in lieu of an overwhelming proportion of multiple-choice items. This can be accomplished by shifting the balance of Depth of Knowledge 1 and Depth of Knowledge 2 and 3 items in favor of the latter.
- Consider revamping test designs to ensure a better balance among skills, concepts and applications, and explore new and creative ways to assess understanding of mathematics and measure habits of mind, dispositions, and the often-ignored processes that society and the workplace increasingly value.

- Consider limiting all adaptive assessments to grade-level content, while allowing a dynamic range of item complexity and difficulty. There is no good reason to use fifth and sixth-grade content on a fourth-grade test just because we can place all of the items on a common scale, when the result is severely mixed and confusing messages to teachers and curriculum designers about exactly what are appropriate, assessed grade level standards.
- Consider increasing the proportion of assessment items in grades 6 and 8 for which an embedded calculator is available in mathematics.
- Consider beginning immediately to research and explore the many ways to wisely use Artificial Intelligence tools to strengthen item writing, adaptive assessments algorithms, assessment reports and, perhaps most importantly, scoring student work on open-ended and constructed-response tasks.
- Consider adopting, like the scoring of writing samples and essays, a common four-point rubric for all complex tasks as a way of communicating, more formally, Full and Complete, Acceptable, In Need of Work, and Unacceptable.

Next steps to move these changes forward

To accomplish this overdue and critically needed agenda, we need a multi-pronged strategy involving position papers, research summaries, legislation, and state and national leadership.

At its core:

- State legislatures and state departments of education must shift policies, update practices and raise what it expected from assessment contractors.
- Congress must amend the Every Student Succeeds Act to provide states with greater latitude over student assessment.
- Schools and school districts, professional associations, teachers, administrators and parents must pressure policy-makers to enact these changes.

Among early steps might be:

- A Council of Chief State School Officers (CCSSO) appointed task force of state assessment and mathematics leaders to provide direction, recommendations, and leadership in this initiative. We envision a coordinated program of initiatives to broadly

till the soil for change, including position papers, journal articles, editorials and commentaries, exemplar items, and state and national legislative lobbying.

- Turn to non-profit and private funders to support a joint national convening of state-level mathematics leaders and state assessment leaders to begin discussions on modernizing state assessments, coordinated by ASSM and CCSSO.
- Empower NAGB leadership to update NAEP as an existence proof for potential changes. We envision NAEP, like PISA, as an exemplar of high-quality assessment.
- College Board and ACT leadership, where these influential giants provide existence proofs for many of these changes, as the SAT, the ACT, and AP assessments support, and provide leadership for curricular and assessment updating.
- Work with and through the newly reconstituted Mathematical Sciences Education Board (MSEB) at the National Academies of Science, Engineering, and Medicine for feasibility studies and examples of modernized assessment items and practices.
- Lobby for shifting federal testing mandates with the Every Student Succeeds Act (ESSA).
- Grants to change state-wide assessments

Conclusion

The world has changed dramatically since the importance and influence of high-stakes assessment became a key piece of the education landscape. Unfortunately, over the past 20 years, only relatively minor cosmetic, inexpensive, and easy-to-make changes have been made. These changes have not kept up with the need for schools, and particularly mathematics, to reflect current needs for greater depth of understanding, more attention to authentic modes of assessment, and a shift from regurgitation of memorized skills to explanation and justification of solutions to realistic and mathematical tasks. We have PISA as a model and can learn much from the PARCC and Smarter Balanced initiatives introduced over ten years ago and abandoned or watered down in the ensuing years. But when readily available tools such as Google, Alexa and Photo Math can easily score Proficient on nearly every current state assessment, it is time for significant improvement in what and how we assess school mathematics.

J

Steve Leinwand is a recently retired principal research analyst at AIR, the American Institutes for Research, in Arlington, VA. He has a 50-year career of leadership positions in mathematics education including teacher, presenter, consultant, mentor, coach, evaluator, researcher, podcast contributor, critical friend and change agent. Since 2002, he has served as mathematics expert on a wide range of AIR projects that focus on high quality mathematics instruction, turning around underperforming schools, improving adult education, evaluating programs, developing assessments and providing technical assistance for school improvement. Leinwand's current work focuses on systemic classroom and school-based initiatives to shift curriculum, instruction and assessment in ways that result in lasting improvement and higher levels of student achievement. He work is captured, in part, on www.steveleinwand.com.

Leinwand co-authored "What the United States Can Learn from Singapore's World-Class Mathematics System (and what Singapore can learn from the United States." He has spoken and written about effectively implementing the Common Core State Standards in Mathematics, differentiated learning and "What Every School Leader Needs to Know about Making Math Work for All Students". As part of AIR's assessment program, Leinwand has overseen the development and quality review of multiple-choice and constructed response items for AIR's contracts with Ohio, Hawaii, Delaware, Minnesota, South Carolina and the Smarter Balanced Assessment Consortium.

Before joining AIR, Leinwand spent 22 years as Mathematics Consultant with the Connecticut Department of Education where he was responsible for the development and oversight of a broad statewide program of activities in K-12 mathematics education including the provision of technical assistance and professional development, the evaluation of Title 1 and K-12 mathematics programs, the assessment of student achievement and teacher competency, and the coordination of statewide mathematics programs and activities. Steve has also served on the NCTM Board of Directors and has been President of the National Council of Supervisors of Mathematics.

Steve is an author of several mathematics textbooks and has written numerous articles. His books, Sensible Mathematics: A Guide for School Leaders in the Era of Common Core State Standards and Accessible Mathematics: 10 Instructional Shifts That Raise Student Achievement were published by Heinemann in 2012 and 2009 respectively. Invigorating High School Mathematics: Practical Guidance for Long-Overdue Transformation, co-written with Eric Milou, was published by Heinemann in 2021. Most recently, Leinwand and his 17-year-old granddaughter, Caroline Welty, co-wrote The Math Tutor's Handbook: Strategies and Tips for Suiccess that was published by Corwin in 2024. In addition, Leinwand was the awardee of the 2015 National Council of Supervisors of Mathematics Glenn Gilbert/Ross Taylor National Leadership Award for outstanding contributions to mathematics education and has been awarded the 2021 NCTM Lifetime Achievement Award.

Jay Meadows is the Chief Executive Officer of Exemplars, a leading force in redefining how students learn to think mathematically. With over 30 years in education, Jay brings a rare blend of classroom insight and research-driven innovation. A former middle school math and science

teacher, he holds advanced degrees in both teaching and mathematical pedagogy—credentials that reflect his deep commitment to meaningful, effective math instruction.

At Exemplars, Jay is not only a primary author of performance tasks but also a national thought partner for educators. He leads dynamic professional development experiences across the United States, helping teachers transform their classrooms with problem-solving, authentic assessment, and the practices that research shows make the biggest difference. His mission is bold: to cultivate a generation of learners who can tackle any challenge with confidence and grit.
Jay's impact extends far beyond U.S. borders. Before his teaching career, he facilitated diplomatic exchange programs for the U.S. State Department, partnered with international nonprofits, and served in the Peace Corps in Kazakhstan. This global lens continues to shape his belief in education as a force for empowerment and equity