

Critical Issues in Mathematics Education
Volume 9 • Workshop 10 • April 2013

Assessing Math



to Know Math

By Mark Hoover



Assessing Math to Know Math

Critical Issues in Mathematics Education
Workshop 10 • April 2013
Mathematical Sciences Research Institute



CIME Organizing Committee

Mark Hoover, <i>Co-Chair</i>	University of Michigan
Kristin Umland, <i>Co-Chair</i>	University of Arizona
Noah Heller	Math for America
Alan Schoenfeld	University of California, Berkeley

MSRI Educational Advisory Committee

Deborah Loewenberg Ball, <i>Chair</i>	University of Michigan	Maria Klawe	Harvey Mudd College
Michèle Artigue	Université de Paris VII Université de Paris VI	Tom Leighton	Massachusetts Institute of Technology
Hélène Barcelo	MSRI	Jim Lewis	University of Nebraska Lincoln
Hyman Bass	University of Michigan	Robert Megginson	University of Michigan
Sybilla Beckmann	University of Georgia	Robert Moses	The Algebra Project Inc.
Herb Clemens	Ohio State University	Alan Schoenfeld	University of California, Berkeley
Ricardo Cortez	Tulane University	Katherine Socha	Math for America
Ted Courant	Bentley School	Hung-Hsi Wu	University of California, Berkeley
David Eisenbud	MSRI		
Roger Howe	Yale University		



MSRI received major funding for the Critical Issues in Mathematics Education Workshop series from the National Science Foundation (NSF-0932078) and Math for America.

The workshop speakers were chosen for their ability to articulate widely held perspectives on mathematics education, but this choice is not meant as an endorsement of those perspectives.

Opinions, conclusions, and recommendations expressed in this booklet are those of the author and do not necessarily reflect the views of the Mathematical Sciences Research Institute, the sponsors nor the organizing committee of the workshop.

Front cover and booklet photos: courtesy of the University of Michigan. Back cover: MSRI staff.

Foreword

In 2004, the Mathematical Sciences Research Institute (MSRI) launched a workshop series *Critical Issues in Mathematics Education* (CIME) to provide opportunities for mathematicians to cooperate with experts from other communities on the improvement of mathematics teaching and learning. In designing and hosting these conferences, MSRI seeks to legitimize such cooperation and to lend support for national progress on critical issues in mathematics education.

The tenth workshop in the series, *Assessment of Mathematical Proficiencies in the Age of the Common Core*, was held at MSRI in Berkeley, CA, April 3-5, 2013. The focus on assessment revisits the topic of the first CIME workshop. It was timely given the wide adoption of the Common Core State Standards (Common Core) and efforts to develop new assessments aligned with them. The Common Core both increases the demand and broadens the conception of what it means to be mathematically skillful. It creates new opportunities and challenges for appraising what students understand and can do. The workshop explored fundamental problems of trying to assess students' mathematical proficiency with a comprehensive perspective on what it means to learn, know, and use mathematics.

The workshop addressed three organizing questions.

1. What are fundamental problems of assessing students' mathematical proficiency aligned with a comprehensive perspective on what it means to learn, know, and use mathematics?
2. What norms and structures need to be developed to work productively across traditionally distinct professional communities?
3. What is involved in vetting assessment items in ways that contribute to developing shared professional knowledge?

In keeping with CIME goals, mathematicians, K-12 teachers, and mathematics education researchers were invited to participate in roughly equal numbers. The extensive involvement of *Math for America* teachers, a co-sponsor of this year's workshop, proved particularly successful.

An innovation at this CIME workshop was investment in working groups that developed, reviewed, and revised items of hard-to-assess mathematical proficiencies. The design of these groups took advantage of a growing understanding of what it takes to work productively across professional communities. Sessions and handouts were designed to introduce participants to key assessment issues and to provide resources for developing robust items. For two and a half days, participants worked together on the development of both formative and summative assessments of critical aspects of mathematical proficiency.

The goal of the present document, commissioned by MSRI, is to draw more mathematicians' attention to the fundamental challenges for developing robust mathematics assessment. It is my hope that by providing a coherent synthesis of the many ideas assembled at the workshop, this document will support mathematicians, mathematics educators, and others in their efforts to engage productively in the development of new assessments, whatever expertise they bring and role they take.

In drafting this booklet, I have made liberal use of ideas developed by the workshop organizing committee and those who presented at the workshop. I gratefully acknowledge these contributors, as well as those who provided feedback on a draft, including the workshop organizing committee, associates of MSRI, and numerous workshop presenters and attendees. I take full responsibility, though, for errors and inadequacies, as well as any dubious claims and arguments.

– Mark Hoover

Assessing Math to Know Math

CONTENTS

Foreword	3
The ubiquity of assessment	5
Challenges of assessing mathematical proficiencies	6
What are mathematical proficiencies?	7
What can be learned from probing student thinking?	8
Interviewing Angela	9
Challenges of aligning assessment with instruction	11
Formative assessment examples	14
Key challenges for assessment design	19
Norms and structures for productive work across professional communities	21
Working together to develop quality items	22
Developing summative assessment item prototypes	23
Views of assessment in and from the classroom	28
Fair assessments in a diverse society	34
Equitable teaching — equitable assessment	35
Reducing test bias	36
Understanding mathematical practices	37
Mathematical proficiencies and future progress	38
Readings and references	39

The ubiquity of assessment

In the wake of the release of the Common Core State Standards (Common Core) in 2010, two multistate consortia were awarded funding from the U. S. Department of Education to develop an assessment system aligned to the Common Core by the 2014-15 school year: The Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium. In addition to these two consortia, smaller projects are working to inform and reflect on the development of new assessments. The undertaking is immense and can be framed in different ways, as is evident in choices made by these different initiatives.

Efforts to develop quality assessments of mathematical proficiencies are fraught with fundamental challenges. Assessment deserves concerted professional attention precisely because its challenges are so central to, and interact with, all aspects of mathematics teaching and learning. Assessment, although commonly thought of as a culminating activity, is in fact as much a starting point as an ending point. This is evident in at least two ways.

First, assessment should not be isolated to a final phase of instruction. Diagnosis is as important as grading, and moment-to-moment teaching requires sizing up student understanding. When a student asks a question, the teacher needs to appraise what the question implies about what the student understands, will be able to hear, and could productively be thinking about. Educational assessment is the process of gathering information about what people know, broadly construed, for the purpose of improving educational efforts. It is essential to the education enterprise that assessment is an ongoing part of instruction, vital at every moment in the process; good assessment is instructional (often directly) and good instruction assesses.

Second, the development of assessment needs to be a practical, ongoing process that informs and is informed by other parts of the education system; it is not simply a final step in the implementation of an improvement plan for education. It is often said that what gets tested gets taught. If only it were that simple. More accurately, assessments interact with teaching, students and parents, school organization, and political and historical circumstance in a complex education system. The ubiquity of assessment in the education process has implications for public policy, the development of quality assessments, and their effective use.

The title of this document, *Assessing Math to Know Math*, is meant to convey this ongoing, double-edged nature of assessment. It suggests the notion that, with well-designed assessments, sensibly used, students may come to know

mathematics more clearly and assuredly. This is in keeping with research that has shown that the act of assessing and being assessed, appropriate to the circumstance, can be a powerful tool for student learning (for a review of evidence, see Bransford, Brown, & Cocking, 1999).

More immediately, though, the title also suggests that the development and use of assessments can be a powerful tool to help those who care about mathematics education gain sorely needed insight, not only into the extent of success and failure in mathematics teaching and learning, but into what mathematical proficiencies are and what standards should be. In other words, efforts to create good assessment tools can lead to clearer articulations of, for example, what is involved in competent use of representations in the context of mathematical explanation or what the curricular implications are for defining similarity in terms of geometric transformations, as done in the Common Core.

...the development and use of assessments can be a powerful tool to help those who care about mathematics education gain sorely needed insight...

Three specific issues may contribute to mathematicians' interest in being involved in efforts to develop better assessments.

- The intellectual challenge of finding out what a student understands.
- The intellectual challenge of designing items that provide the information needed.
- The importance of having the mathematician's voice in discussions and debates about policy issues.

Each of these is elaborated and warranted in the sections below. After sketching the challenges of assessing mathematical proficiencies and of designing assessments that work in concert with teaching and learning, this document describes the requirements for working across professional communities in ways that are productive and avoid wasted effort and gratuitous conflict. To provide concrete grounding for the discussion, numerous examples of assessment tasks are provided and a mathematics interview of student and a teacher discussion of classroom assessment are included. The document concludes with a discussion of the design of fair assessments in a diverse society.

Challenges of assessing mathematical proficiencies

Part of the challenge for developing good assessments relates to an incomplete understanding of mathematical proficiencies. Three specific issues stand out.

1. The bi-directional relationship between assessment and standards.
2. Technical measurement issues.
3. Unintended influences on instruction.

The reality is that many students, at all levels, complete school classes and college courses with good grades yet have thin understanding of the content, limited facility with talking, reasoning, and using mathematics in out-of-school contexts, and distorted ideas about its practice and its contribution to the world. When asked to say a few words about what a derivative is, many successful undergraduates struggle, making statements that reveal severe limitations, such as, “the derivative of x squared is $2x$,” as though this constitutes a definition of the derivative. Some of the deficiencies are readily apparent, but many remain unnoticed until they accumulate into a quagmire of mathematics miseducation. Students need to be able to add, subtract, multiply, and divide, but they should also know that mathematical claims are not established by voting on them, and they need good instincts about what to do mathematically when faced with a mathematics problem that has no immediate solution.

The 1989 National Council of Teachers of Mathematics (NCTM) curriculum standards focused attention on five *process standards*: problem solving, reasoning and proof, communication, representation, and connections. The 2010 Common Core State Standards for Mathematics identified eight *mathematical practices*. Unfortunately, it is unclear what these practices are (both in that they are insufficiently elaborated to be commonly understood and that their defense as strategic choices individually and as a set is underdeveloped) and how they combine with content knowledge to constitute real mathematical proficiency.

Disappointing student achievement should come as no surprise given the lack of common, technical language for talking explicitly about mathematical proficiency. The above comments imply that a basic and iterative aspect of developing an assessment is establishing an adequately clear and usable picture of desired student mathematical proficiencies. The 2010 *Common Core State Standards for Mathematics* provides an improved specification of mathematical proficiency, but much of what matters most about mathematical practices and about their interplay with specific mathematical content remains underspecified. Even research mathematicians, by engaging in the work

of developing high-quality assessments, can come to know mathematics more fully, or at least more explicitly. In doing so, they would build on past efforts of mathematicians engaged in unpacking mathematical content and practices for the purpose of teaching and learning, such as those of Klein (1908/1939), Polya (1957), Thurston (1995), Mancosu, Jørgensen, and Pedersen (2005), and others. This may seem provocative, but the reality is that there is a real need for the mathematics community to sharpen its ability to express to a wider public audience what mathematics is and what mathematicians do. Mathematicians are, in the end, the only people who can provide this insight—who can ensure that school mathematics reflects the practices of mathematics. Explicit elaboration is not required for doing mathematics, where much of professional mathematical practice can remain tacit, but effective teaching requires it and the improvement of mathematics education depends on it.

In addition to the double-edged nature of assessment, another important challenge for assessing mathematical proficiencies is that the design of assessment tasks is not straightforward, even when goals are clear. Students routinely answer questions right for the wrong reasons, and answer questions wrong when they may understand a great deal about the content, but misstep, again for a host of reasons. Assessment tasks need to elicit student thinking, cover terrain efficiently, and be practically producible, usable, and score-able. They need to be fair in light of all-too-often patterns of systematic test bias. And, evidence provided by assessment tasks needs to be reliably interpretable. Tasks for large-scale, high-stakes assessments need to satisfy important psychometric constraints. Tasks for formative assessment need to provide a clear window on student thinking. Competing design challenges for assessment are non-trivial.

Last, but not least, a big-picture challenge for the assessment of mathematical proficiencies is that the design and production of assessment needs to function well in the overall system. A good assessment not only has to produce reliable, informative, and actionable education feedback, it must simultaneously keep from derailing productive teaching and learning of worthwhile mathematical proficiencies. Furthermore, it needs to be something teachers and students can be expected to use to improve their efforts. It needs to fit strategically with other moving parts of the mathematics education enterprise, such as teacher education, professional development, curriculum materials, school improvement, and policy initiatives. Otherwise, it either fails and is forgotten or, worse yet, erodes other productive efforts.

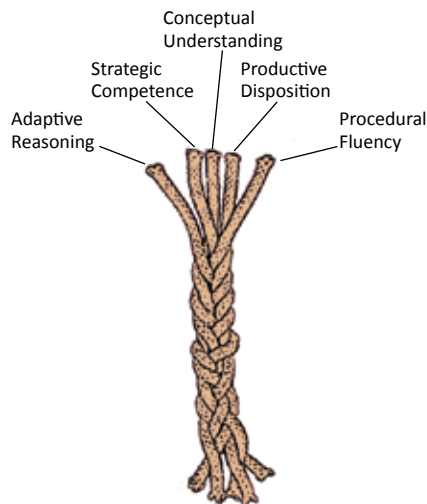
What are mathematical proficiencies?

In describing the historical pendulum swings in what it means to be successful in learning mathematics, the National Research Council's report, *Adding It Up* (Kilpatrick, Swafford, and Findell, 2001), argues that, in the end, everyone wants a comprehensive set of learning outcomes for students, which it goes on to describe as a set of five intertwined strands of mathematical proficiency:

- *Conceptual understanding*—comprehension of mathematical concepts, operations, and relations.
- *Procedural fluency*—skill in carrying out procedures flexibly, accurately, efficiently, and appropriately.
- *Strategic competence*—ability to formulate, represent, and solve mathematical problems.
- *Adaptive reasoning*—capacity for logical thought, reflection, explanation, and justification.
- *Productive disposition*—habitual inclination to see mathematics as sensible, useful, and worthwhile, coupled with a belief in diligence and one's own efficacy (p. 116).

The report stresses that, “the five strands are interwoven and interdependent in the development of proficiency in mathematics,” and offers the visual image of a braided rope (p. 117).

Intertwined strands of proficiency



This characterization of mathematical proficiency is important because it reminds us, in practice and in research, that improvement is not a matter of deciding whether to teach procedural skills or conceptual understanding, but both. Indeed, choosing between such false dichotomies ensures failing to educate children adequately. The report's characterization also points out the need to consider the composite nature of mathematical proficiency in its development and assessment.

Although the report provides an important overall framing, it does not offer a specification of mathematical proficiencies that could be used to organize curricula or instruction. Identifying and sufficiently elaborating a set of components usable by practitioners in improving teaching and learning remains an important open concern.

The NCTM process standards identify problem solving as important. Research on problem solving has contributed important insights into the teaching and learning of problem solving skills. Everyone wants children to learn to solve problems, yet it is not clear how central this goal should be or how it fits into a helpful articulation of mathematical proficiencies.

The Common Core standards combine NCTM's process standards with the National Research Council's proficiency strands to identify eight mathematical processes or practices.

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

These practices “describe ways in which developing student practitioners of the discipline of mathematics increasingly ought to engage with the subject matter as they grow in mathematical maturity and expertise” (p. 8). They are referred to as mathematical practices and have a more disciplinary bent than lists of the past. However, with only a one-paragraph description for each, they may point in useful directions, but they leave much underspecified, with implications for teaching and assessing unclear.

Adding to the above discussion is a characterization of mathematical topics, or content standards. As with the Common Core practice standards, the Common Core content standards blend disciplinary sensibilities with pedagogical experience to yield a set of standards widely acceptable across relevant professional communities. The Common Core content standards provide much needed focus and coherence given the kitchen-sink documents that have characterized state standards in recent decades. They “are logical and reflect, where appropriate, the sequential or hierarchical nature of the disciplinary content from which the subject matter derives” (p. 3). The difficult task of figuring out an effective way of attending to the content and practice standards simultaneously is vital to the development of quality assessments.

What can be learned from probing student thinking?¹

Teaching mathematics is about helping students travel toward more and more developed forms and expressions of mathematics, both theoretical and technical. To effectively assist in that journey, the teacher needs to know something about students' current thinking — the mathematical ideas and resources students bring to the work and how they think about and express them. This is not about how full the glass is, but about a much more complex profile. For example, as a young student who correctly reads and writes 45 begins to understand more about numbers and knowingly recognizes that the 4 represents 40, he or she may write 405 instead of 45. A teacher needs to be able to see and name this progress.

Probing student thinking is a core practice of teaching and is probably the teaching practice least familiar to many mathematicians, who tend to focus their ventures into education more on school curricula and the development of mathematical ideas. Pedagogical interviewing, to probe student thinking, is a highly skilled version of formative assessment. In pedagogical interviewing, the student's ideas and ways of thinking are primary. The interviewer is, in an important sense, the learner (about the student's thinking), and the student is the teacher (the one who knows his or her own thinking). The questions and prompts that the interviewer uses must be non-invasive yet make visible the complex boundaries of a student's mathematical world and the nature of the mathematics at hand. In conducting or even observing a student interview, it is important to track simultaneously on what is revealed about student thinking and on the prompts and interactions that bring that knowledge into view.

An example of a pedagogical interview is discussed on page 9. A video of the interview is available at: <http://www.msri.org/workshops/696/schedules/16544>. The point of such an interview is not to suggest that teachers set aside time to interview each of their students outside of the context of



Multiple types of insights are important for informing conversations about assessing students' mathematical proficiencies.

classroom instruction. Instead, it is to highlight an aspect of skilled teaching (an important form of assessment that occurs on the insides of teaching) and to make it available for public observation and professional discussion. The intent is also to raise questions about what it means for a student to exhibit desired mathematical proficiencies. As student thinking is skillfully uncovered, questions surface about the mathematical proficiencies that really matter, the design of tools for gathering information about students' developing proficiencies, and valid and practical ways of interpreting the students' productions — all central concerns for effective instruction.



Instruction must assess understanding of content and provide insight into student thinking.

Some specific questions to keep in mind when observing skillful probing of students' mathematical proficiencies are the following:

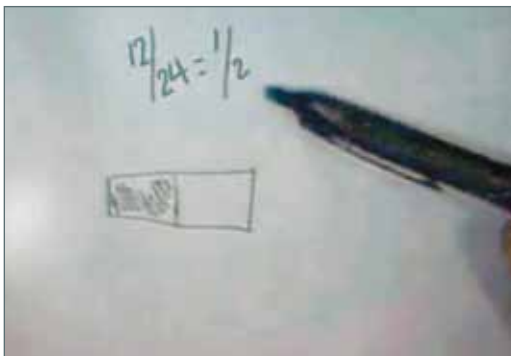
- What mathematical terrains does the interview probe?
- How would the student's knowledge, skills, and dispositions in each part of the landscape be characterized? What kinds and levels of mathematical proficiencies, or practices, does the student exhibit?
- What questions about the student's thinking still remain and how might the student's thinking be probed for answers?
- What does the student's thinking suggest about what there is to be learned and ways of characterizing mathematical proficiencies?

Teachers are experienced at listening to the work of children and may have fine-grained observations. In contrast, those less familiar with children's thinking might notice important mathematical aspects of the questions and of student responses. Both types of insights are important for informing conversations about assessing students mathematically.

¹This section is adapted from a presentation given by Hyman Bass (University of Michigan).

Interviewing Angela²

Early in the interview, the teacher asks the student to write an example of a fraction that she knows — any fraction she wants.³ The student writes that $12/24 = 1/2$ and says, “So I wrote twelve twenty-fourths, which also equals a half.” When asked to represent what “that fraction” means, she draws a “bar” and shades half of it.



When asked which fraction she is representing, she says “both,” then qualifies that she’s really representing the half because she, “divided it into two and shaded only one of them.” The teacher notices the shift and probes the student’s thinking further.

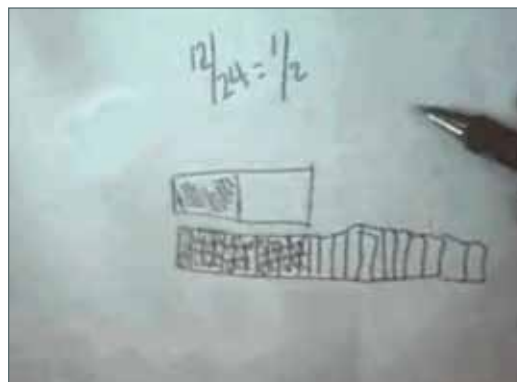
Teacher: You started to say that you’d drawn both of them a moment ago. Is that—did you draw both of them? Both fractions or did you only draw one half?

Student: Well I—since I know that twelve is half of twenty-four and I know that twelve twenty-fourths is also equal to a half. So it’s kind of—not both because if you were to draw both Wait. Never mind.

Teacher: Well how—how would you draw twelve twenty-fourths? Maybe you don’t want to—yeah, I don’t know if you want to do it or not, but could you describe what you would do?

As the student describes what she would do, she initially says she would draw a longer bar, then says maybe it would be a shorter bar, “because twenty-fourths are smaller.” Interested in the student’s thinking, the teacher subtly maintains the question, “So you would need So about how long would that bar have to be?” The student wavers, “Maybe as long as this one,” and then, “Maybe a little longer.” When the teacher asks if she wants to try drawing it, the student draws a rectangle about the same size as the first and marks off small rectangles to represent twenty-

fourths. When she only gets 14 pieces in the rectangle, she appends enough more pieces to the end to make 24 in all.



When asked to explain what she drew, she explains having shaded in twelve of the twenty-four pieces to show twelve twenty-fourths. When asked to use her drawings to explain her initial equation, she explains that “the thing that they have in common is that they’re both half of the bar” — that they are the same because “they both fill in half of the bar.”

The teacher acknowledges (reassures her) that what she has said makes sense, but poses that someone might say that it looks as if twelve twenty-fourths is greater. The student explains that it is not bigger and why it happens to look that way.

Student: No, it’s actually—it’s not twelve—twelve-twenty-fourths isn’t greater, its just the bar might be longer because of the tiny pieces and—but the one half is—isn’t—is originally—is not bigger its just that the twenty-four pieces are so tiny that they just can’t fit in a bar like this. So there’d need a longer bar.

Teacher: So the problem is not—is that it’s hard to fit—it’s impossible or it’s hard to fit the twenty-four in there? Is it impossible to do it or just difficult to do it?

Student: It’s kind of difficult, but it’s not impossible.

This short, initial segment of this interview opens up an important set of issues for mathematics teaching and for assessment. In isolation, her representation for comparing these two fractions might be concerning, as might her wavering on key issues. At the same time, her facility with this complex set of ideas is impressive. She exhibits a great deal

²This interview was developed and conducted by Deborah Ball (University of Michigan).

³This text refers to a “teacher” and a “student” generically rather than by name because the point is not about this particular teacher and student but about the content and dynamics of student thinking and their educational implications.

of confidence and stability in her thinking. She does not seem to be guessing about what is wanted by the teacher, but is laying out relatively recent developments in her own emergent thinking about fractions. At the end, she ably critiques a line of reasoning not her own.

In addition to observations about this student, this brief segment of the interview provides insights into the mathematical content to be taught. What does it mean for two fractions to be equal? Are $\frac{1}{2}$ and $\frac{2}{4}$ only one fraction or two? Fractions are being represented with reference to different units or wholes all of the time. When is it okay to use a different whole and when not? The definition of a fraction in the Common Core is as follows.

Understand a fraction $\frac{1}{b}$ as the quantity formed by 1 part when a whole is partitioned into b equal parts; understand a fraction $\frac{a}{b}$ as the quantity formed by a parts of size $\frac{1}{b}$. (3.NF.1)

This standard is followed by standards about fraction equivalence and comparison.

3. Explain equivalence of fractions in special cases, and compare fractions by reasoning about their size.
 - a. Understand two fractions as equivalent (equal) if they are the same size, or the same point on a number line.
 - b. Recognize and generate simple equivalent fractions, (e.g., $\frac{1}{2} = \frac{2}{4}$, $\frac{4}{6} = \frac{2}{3}$). Explain why the fractions are equivalent, e.g., by using a visual fraction model.

What do these standards suggest about this student's reasoning? What does this student's reasoning suggest about ways of thinking about the mathematics to be learned? The teaching of mathematics requires a kind of unpacking of mathematical content in highly nuanced ways so that the dynamic between students' thinking and disciplinary knowledge can be thoughtfully considered in the design of instruction and in assessment that appraises and informs it.

Efforts to develop assessments of students' mathematical proficiencies must inevitably contend with questions about priorities regarding mathematical proficiencies and about the nature of students and their thinking. Even when standards are given, as one considers evidence for proficiencies, questions arise about the meaning of the standards. Furthermore, even when standards are clear, in wanting to inform teaching and learning, questions arise

about the nature of students and their thinking in relation to the standards.

This last comment deserves an additional remark. Some mathematicians and mathematics teachers cringe at things their students say and do (or fail to say and do) because they deem it to be, if not wrong, at least out of keeping with mathematical understanding and sensibility. Some even say that they do not want to know what their students think because it can be distressing to hear and grates on their mathematical sensibilities. Certainly students are responsible for the efforts they make and for their mathematical learning, but an aversion for student thinking is an unfortunate disposition for one responsible for teaching. Teaching is about connecting students' current thinking to developed thoughts of a subject. Ignoring half of the equation reduces teaching to a hit-or-miss proposition, one likely to be more successful with students like oneself or even limited to those who could just as well learn on their own. It is also professionally irresponsible, and with children, morally irresponsible.

Moreover, attending to student thinking opens up its own mathematical exploration. As a search for isometries of the plane or solutions to Diophantine equations provides intellectual challenge and reward, so too can figuring out where a student is on solid ground and what mathematical route might bring the student from where the student is into what is to be learned. In discovering this work, mathematical ideas that are often so implicit that they are invisible suddenly become visible. These occasions can also informatively point out aspects of mathematical proficiencies that educators are leaving implicit and failing to teach, and they can refine our language for explicitly expressing mathematical proficiencies in ways that support the entire educational enterprise.

The interview with Angela continues, exploring the placement of fractions on a number line, the issue of unequal partitions in an area model, and comparison of fractions given only symbolically. The last half of the interview explores three combinatorial problems, including questions about structural similarities and differences among the problems. The examination of mathematical structure reveals a relatively large swath of mathematical proficiencies that is probably important yet largely unaddressed in current instruction and in the design of assessment. (Mathematical structure as it is explored here may well be an example of an aspect of mathematical proficiency for which we lack adequate language to support teaching and learning and where this interview provides a resource for exploring this important issue.)

Challenges of aligning assessment with instruction⁴

People often use the phrases summative assessment and formative assessment to distinguish assessment that evaluates student progress at a point in time (with scored feedback for external accountability) from diagnostic assessment used by teachers during instruction to inform teaching and learning (with qualitative feedback used to modify classroom activities). However, the division is not so clear. For instance, Newton (2007) argues that the phrase summative assessment only applies to a kind of result while formative assessment only applies to a use of results. Distinguishing between types of information and uses of information is important for thinking and communicating about assessment. In particular, although there are in general many different purposes for any particular assessment, the purpose of informing teaching and learning is critical to all assessments.

Hugh Burkhardt coined the acronym WYTIWYG (what you test is what you get) to point to the inadvertent influence of summative assessment on what happens in classrooms. Even though a high-stakes test may be designed to evaluate student progress affordably and reliably, and not meant to inform teaching directly, high-stakes tests, including the format of items⁵ and their administration, inevitably drive instruction as much or more than standards do. This places a heavy burden on high-stakes tests. They must be designed with attention to impact on instruction as well as being affordable, accurate, and reliable at scale. Such tests can be a positive or negative force, vis-à-vis standards and classroom practices. Good tests, ones well aligned with instruction, able to inform it, and useful as a model for it, can push things in the right direction. Poor tests can undermine the very activities of teaching and learning they are meant to inform.

Effective teaching requires simultaneous attention to both the content to be taught and to the meanings and engagement of learners. Thus, tests aligned with instruction must both assess understanding of content and provide insight into student thinking. Assessment items, then, must get at the substance of what students think and can do mathematically.

In brief then, formative assessment is about understanding student thinking and using that understanding to help students to learn more effectively and more deeply. That is, formative assessment is about building instruction around what is learned about student thinking. In contrast, summative assessment must rigorously assess understanding of content — yet it too must consider instruction. It must

provide images of proficiency and information about what students can and cannot do that is well enough aligned with instruction to not derail it. Thus, summative assessment needs to adhere to many of the design principles of good formative assessment, while still tackling the need for large-scale, external accountability.

To illustrate the WYTIWYG concern, consider a few sample test items. The item below is a released item from the California Standards test and represents the upper limit of what eighth grade students in California are being asked to do. (<http://www.cde.ca.gov/ta/tg/sr/css05rtq.asp>)

What is the y -intercept of the graph of
 $4x + 2y = 12$?

A -4
 B -2
 C 6
 D 12

One might think about this as a two-step problem, one in which you substitute in $x=0$ and a second in which you solve for y , or you could put it in standard slope-intercept form and read off the y -intercept. Although everyone is likely to agree that students who have studied linear equations should be able to do this problem, such a problem can readily be taught as a mindless routine, and students with little understanding of linear equations might still get this item correct. Indeed, focusing instruction on being able to get this problem reliably correct might well lead students to develop an understanding of mathematics similar to that of a student described by Erlwanger (1975, p. 25) as “an inflexible rule-oriented attitude toward mathematics, in which rules that conflict with intuition are considered ‘magical’ and the quest for answers ‘a wild goose chase.’” The point here is that, while students should be learning the mathematics implicated by this problem, we need to be attentive to the ways in which such problems can perniciously turn back on instruction in ways that undermine the very goal they are meant to set. Answering such an item correctly does not necessarily imply understanding of the mathematics associated with this problem or that instruction that takes the answering of this item as the goal is appropriately focused. Minor rephrasing might avoid more immediate problems. For instance, a question could ask for the point on the line in the first quadrant farthest from the x -intercept, or halfway between the y -intercept and the x -intercept. Even then, though, as such problems are deemed

⁴ This section is adapted from a presentation given by Alan Schoenfeld (University of California, Berkeley).

⁵ The term *item* is common in the context of assessment and refers to a discrete *problem* or *task* used in an assessment (such as a question on a test). As the interaction between assessment and instruction is considered, the terms become nearly interchangeable. This document extensively uses *item* as a signal that the mathematics problems, or tasks, are being thought of specific to the purpose of assessment.

important, attention needs to be given to ways in which they may lead to instruction that misses the point.

Here is another example, in which you can see that the line in option A has a positive slope with a y -intercept below the x -axis.

Which best represents the graph of $y = 2x - 2$?

A **C**

B **D**

Although it might seem as though this item tests conceptual understanding of slope and intercepts as represented in the equation of a line, if students are taught heuristics for answering test questions of this form, the item may measure nothing more than basic pattern recognition. We want to know that students can answer simple mechanical questions about the properties of graphs. We also want to know that they can answer simple well-constructed conceptual questions about the properties of graphs. Unfortunately, an over-abundance of problems such as the two above can encourage instructional approaches that undercut the healthy development of students' mathematical proficiency and engender regrettable attitudes about and views of mathematics.

In a different context, with different purposes, are two assessment items developed by educators at the Shell Centre.

Looking at these graphs, many students will initially notice that C is furthest to the right, so must have won. Some will pause and realize that the horizontal axis is time, so C took the longest, so C lost. Students might ask what happened in the horizontal section of the graph for runner C. Interpreting this part of the graph requires a

Hurdles Race item

The rough sketch graph below describes what happens when 3 athletes A, B and C enter a 400 metres hurdles race.

Distance (meters)

Time (seconds)

— A
- - B
... C

Imagine that you are the race commentator. Describe what is happening as carefully as you can. You do not need to measure anything accurately.

non-trivial inference, even though many with mathematical training will immediately interpret it correctly. Notice what is required. First, the horizontal segment means the runner is not making progress. Second, the context is a high-hurdles race, so the runner may have fallen over a hurdle and then limped along to a late finish. Students might also think about what happened at the point of intersection for runners A and B. Throughout most of the race, runner A was ahead, having run a greater distance in less time, but near the end of the race runner B caught up with runner A because they are in the same place at the same time. Then, runner B kicked and won, while runner A, who had been leading, ran out of energy.

This item assesses, among other things:

- Interpreting distance-time graphs in a real-world context.
- Realizing “to the left” is faster.
- Understanding points of intersection in that context (they are tied at the moment).
- Interpreting the horizontal line segment.
- Putting all this together in an explanation.

It also involves several mathematical practices:

- 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. . . .

We want students to be able to reason in such ways, as a result of learning mathematics, in addition to answering mechanical questions and well-constructed conceptual questions about the properties of graphs.

Mathematical practices, along the lines of those identified in the Common Core standards, are where the content “lives.” If you look at the mathematical content of either the *Hurdles Race* or the *Sale 25%* items, it is far richer content assessment than what is often tested, such as in the first two items, but additionally, to deal with the latter questions, students have to think mathematically and carry out mathematical work. Assessments need to capture this “union” of content and practices.

Sale 25% item

Part 1: In a sale, all the prices are reduced by 25%. Julie sees a jacket that cost \$32 before the sale. How much does it cost in the sale?

Part 2: In the second week of the sale, the prices are reduced by 25% of the previous week’s price. In the third week of the sale, the prices are again reduced by 25% of the previous week’s price. In the fourth week of the sale, the prices are again reduced by 25% of the previous week’s price. Alan says that after 4 weeks of these 25% discounts everything will be free. Is he right? Explain your answer.

If the exams currently being developed to assess the Common Core standards stay true to the integrated notions of content and practices represented in the standards, then there will be a revolution in testing, and these exams will compel (or at least demand) changes in classroom practices. However, teachers and schools will need support to make such change.

In other words, one challenge for the development of rich assessments that measure desired outcomes associated with real mathematical proficiencies is that unless students and teachers get support for working productively toward those challenges, the chances of success are slim. The premise behind a focus on assessments is that they will drive improvement, but this premise breaks down if students and teachers are not positioned to use the assessment information to inform their efforts. An important role for the professional



Formative assessment is about getting information about what students think for the purpose of informing teacher decisions about how best to engage and support students.

community concerned with the improvement of assessment is to provide tools for improving ways of preparing students to do well on rich assessment tasks.

Formative assessment is about getting information about what students think for the purpose of informing teacher decisions about how best to engage and support students. It is not summative assessment given weekly. The purpose of formative assessments is not simply to show what students “know and can do” after instruction, but to reveal their current understandings in order to help them improve.

Resources on the web:

- Mathematics Assessment Project
- Silicon Valley Math Initiative
- Illustrative Mathematics
- Inside Mathematics
- Math Forum
- National Council of Teachers of Mathematics

As Black and Wiliam (1998) point out, if you return student papers with only scores on them, they look at the scores and crumple them up and throw them away; if you return papers with only comments, they read them; and if you return papers with scores and comments, they look at the scores and crumple them up and throw them away. Such dynamics need consideration.

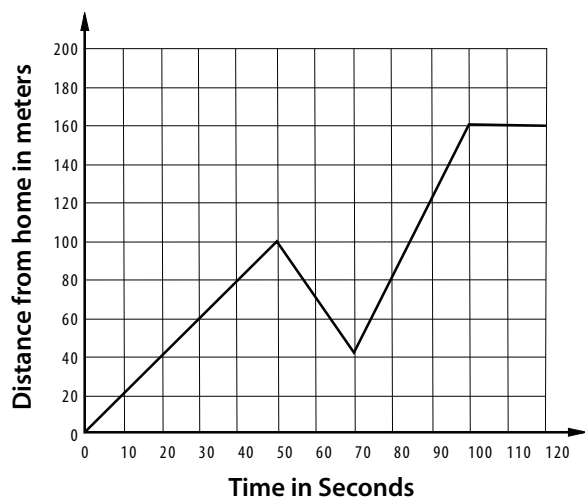
Formative assessment examples

Developing formative assessment is hard work. Tools, examples, and support are essential. One such resource is the formative assessment lessons developed by the Mathematics Assessment Project, which offer the following:

- Rich “diagnostic” situations
- Descriptions of what the mathematical issues are for students (what is assessed)
- Things to do as a teacher on seeing the results of the diagnosis

For instance, for a formative assessment lesson related to the hurdles race item discussed above, Shell Centre materials provide information about common student graphing misconceptions, such as ways in which, for distance-time graphs, students often confuse a picture of a story with a graph of the story or interpret distance as speed. As support for anticipating and planning for possible misconceptions, the materials provide a task to give before the lesson, as homework or in class. The task given for the hurdles race item is about walking from home to a bus stop. It is simpler and provides more complete information from which to reason.

Every morning Tom walks along a straight road from his home to a bus stop, a distance of 160 meters. The graph below shows his journey on one particular day.



Describe what may have happened. Is the graph realistic? Explain.

Some students are likely to talk about Tom going up a hill and down a hill or other misconceptions, so the materials provide a description of common issues and suggest questions and prompts for helping students confront their understandings.

Graph interpreted as a picture

E.g. The student assumes that as the graph goes up and down, that Tom’s path is going up and down.

E.g. The student assumes that a straight line on a graph means that the motion is along a straight path.

E.g. The student thinks the negative gradient means Tom has taken a detour.

Suggested questions and prompts

- If a person walked in a circle around their home, what would the graph look like?
- If a person walked at a steady speed up and down a hill, directly away from home, what would the graph look like?
- In each section of his journey, is Tom’s speed steady or is it changing? How do you know?
- How can you work out Tom’s speed in each section of the journey?

Graph interpreted as speed v time

The student has interpreted a positive gradient as speeding up and a negative gradient as slowing down

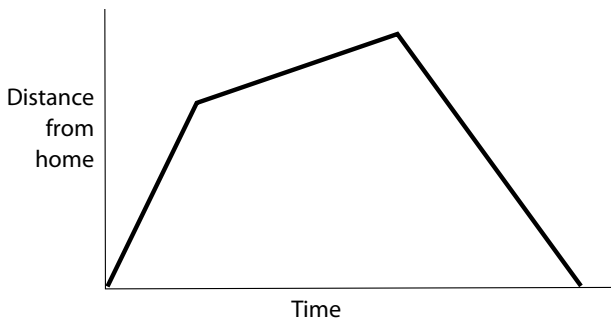
Suggested questions and prompts

- If a person walked for a mile at a steady speed, away from home, then turned round and walked back home at the same steady speed, what would the graph look like?
- How does the distance change during the second section of Tom’s journey? What does this mean?
- How does the distance change during the last section of Tom’s journey? What does this mean?
- How can you tell if Tom is travelling away from or towards home?

For example, if a student thinks a straight line on a graph means the person is moving in a straight line, a teacher might consider asking about graphing distance traveled over time for a situation in which a person walked in a circle around a house.

Supports, such as this pre-lesson task, give teachers initial insights into what students might think and initial ideas for pressing on student thinking. The full lesson for the *Hurdles Race* item begins with a task that asks students to match a story with a graph.

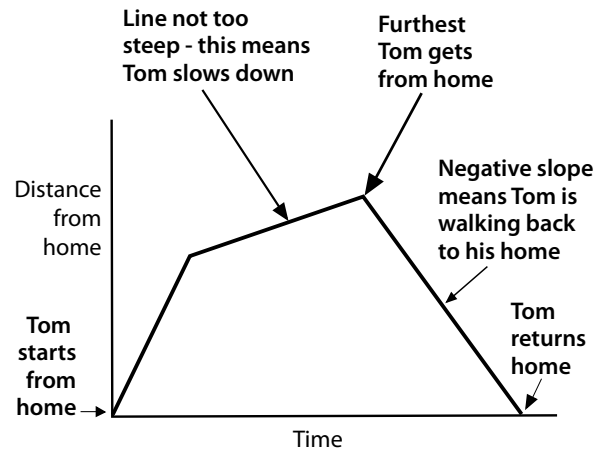
Matching a graph to a story



- A. Tom took his dog for a walk to the park. He set off slowly and then increased his pace. At the park Tom turned around and walked slowly back home.
- B. Tom rode his bike east from his home up a steep hill. After a while the slope eased off. At the top he raced down the other side.
- C. Tom went for a jog. At the end of his road he bumped into a friend and his pace slowed. When Tom left his friend he walked quickly back home.

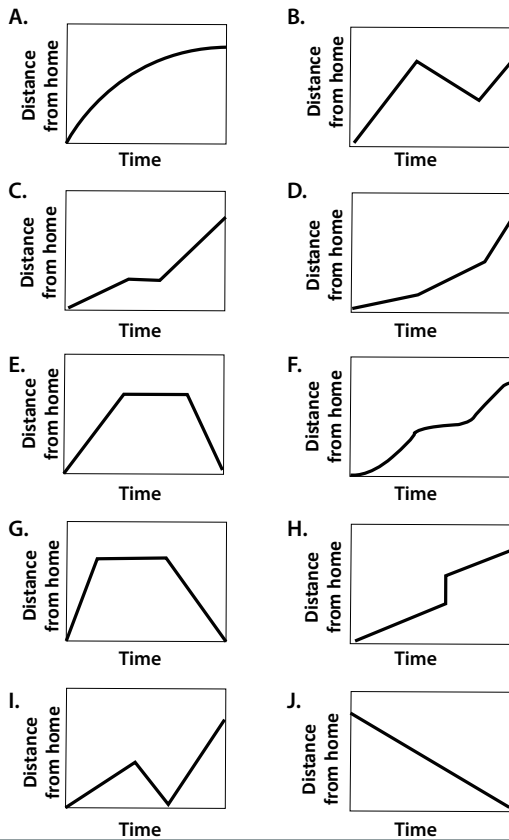
The lesson also gives students a chance to annotate and explain.

An annotated graph



The full lesson then has students match a collection of graphs and stories and convert the original graphic to a table.

Card set A: Distance-time graphs

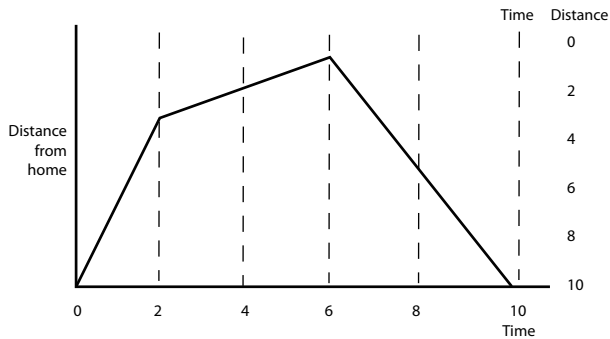


Card set B: Interpretations

1. Tom ran from his home to the bus stop and waited. He realized that he had missed the bus so he walked home.
2. Opposite Tom's home is a hill. Tom climbed slowly up the hill, walked across the top, and then ran quickly down the other side.
3. Tom skateboarded from his house, gradually building up speed. He slowed down to avoid some rough ground, but then speeded up again.
4. Tom walked slowly along the road, stopped to look at his watch, realized he was late, and then started running.
5. Tom left his home for a run, but he was unfit and gradually came to a stop!
6. Tom walked to the store at the end of his street, bought a newspaper, and then ran all the way back.
7. Tom went out for a walk with some friends. He suddenly realized he had left his wallet behind. He ran home to get it and then had to run to catch up with the others.
8. This graph is just plain wrong. How can Tom be in two places at once?
9. After the party, Tom walked slowly all the way home.
10. Make up your own story!

Whole class discussion: Interpreting tables (15 minutes)

Bring the class together and give each student a mini-whiteboard, a pen, and an eraser. Display Slide 5 of the project resource:

Making up data for a graph

On your whiteboard, create a table that shows possible times and distances for Tom's journey.

It then has students convert all of the graphs to tables and use all three sources (graphs, stories, and tables) to reason about the fit among them.

The Mathematics Assessment Project's formative assessment lessons are built around tasks that provide information about student thinking as a basis for guiding instruction. They are developed into full instructional plans, but begin with an assessment task designed to provide students with an introduction to the content and teachers with baseline information about what their students know and can do and about how they think about the content of the lesson.

Here is the shell of another formative assessment lesson, including a description of the goals and three statements that students need to decide are always, sometimes, or never true.⁶

Lesson goals

This lesson unit is intended to help you assess how well students can:

- Understand the concepts of length and area.
- Use the concept of area in proving why two areas are or are not equal.
- Construct their own examples and counterexamples to help justify or refute conjectures.

Common Core State Standards

This lesson involves *mathematical content* in the standards from across the grades, with emphasis on:

G-CO Prove geometric theorems

This lesson involves a range of *mathematical practices*, with emphasis on:

2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.

1. James says:

If you draw two shapes, the shape with the greater area will also have the longer perimeter.

Is James' statement Always, Sometimes or Never true?

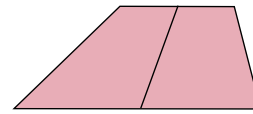
Fully explain and illustrate your answer.

2. Clara says:

If you join the midpoints of the opposite sides of a trapezoid, you split the trapezoid into two equal areas.

Is Clara's statement Always, Sometimes or Never true?

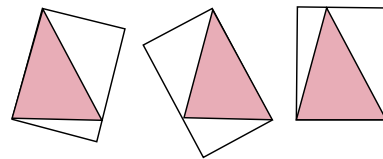
Fully explain and illustrate your answer.

Evaluating statements about length and area**3. Alex says:**

There are three different ways of drawing a rectangle around a triangle, so that it passes through all three vertices and shares an edge. The areas of the rectangles are equal.

Is Alex's statement Always, Sometimes or Never true?

Fully explain and illustrate your answer.



A nice feature of this item is that it calls attention to the fact that “the” area formula for a triangle is in fact three different formulas—one for each base. However, the construction that Alex proposes is possible if and only if the triangle is scalene. If there is an obtuse angle, then the two shortest sides will only be *contained in*, *not equal to* a side of the corresponding rectangle. Moreover, the rectangles have equal areas if and only if the triangle is scalene.

These formative assessment lessons have students explain their thinking and critique the thinking of others, as in the following.

⁶ These materials are available at <http://map.mathshell.org/materials> or by searching for “mathematics assessment project” online. Developers plan to upload 100 lessons that are downloadable free for non-commercial use.

Diagonals of a quadrilateral

If you draw in the two diagonals of a quadrilateral, you divide the quadrilateral into four equal areas.

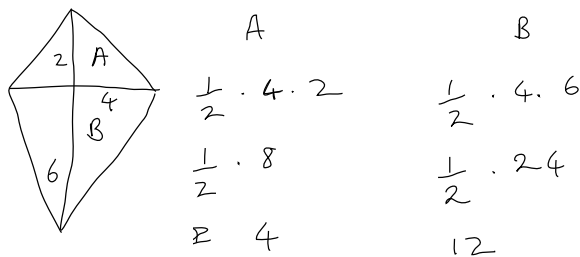
Is this statement always, sometimes or never true?

If you think the statement is always true or never true, then how would you convince someone else?

If you think the statement is sometimes true, would you be able to identify all the cases of a quadrilateral where it is true/not true?

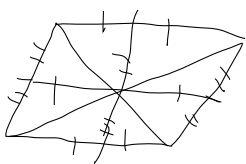
The materials have students discuss the task and provide other (hypothetical) student work and help students critique it. These are central skills called for in the Common Core.

Student work 1



Kite = not true

Student work 2



All triangles are congruent
(SSS)

Parallelogram = always true

As Ball and Cohen (1999) argue, teachers will not improve simply by being told to do so, in particular when they do not know how to do so.

Formative assessment materials such as these are an important resource, but materials alone provide little promise of improvement. Assessment is often seen as an engine for driving change in teaching and learning. However, improvement depends on capacity. As Ball and Cohen (1999) argue, teachers will not improve simply by being told to do so, in particular when they do not know how to do so. Likewise, they will not improve simply by being handed new materials or new assessments without adequate will and support for learning how to use them effectively. Existing mathematics teaching practice and professional education in the United States are broken systems. Establishing goals and yardsticks is a helpful step, but the success of standards, curricula, and assessments depends on the creation of opportunities for teachers to reconsider their current practices, examine others, and learn more about the subjects and students they teach as they experiment with professionally vetted alternatives.

These observations lead to questions.

- How do we provide professional development and material support for the teaching community? At the national level? At the district/building level? What issues do we face in trying to provide professional development in a resource-limited environment?
- What are the implementation challenges at the district and school level (e.g., alignment, teacher time for PD, building capacity, and coping with change)? How do we identify those challenges; and how do we work to help district and school leaders support teachers in attaining envisioned goals?

Questions such as these lie beyond the context of creating assessments, but are important to keep in mind because features of assessments significantly shape opportunities that can be provided to teachers.

Key challenges for assessment design⁷

In recent years the field of educational measurement has developed a relatively stable set of professional standards. (See, e.g., *Educational Measurement* [Brennan, 2006], jointly sponsored by the American Council on Education and the National Council on Measurement in Education, and the *Standards for Educational and Psychological Testing* [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999].) However, along with this development has come a growing recognition of the profound challenges associated with test validity (a measure of whether inferences and actions based on the test are appropriate). Increasingly, validity arguments being integrated into the process of test development, with a focus on coherence of an overall interpretive argument about the nature and role of the test content and about the use of scores. (See, e.g., the chapters by Mislevy, 2006, and by Kane, 2006, in *Educational Measurement*.) These represent important progress, but the WYTIWYG phenomenon — a kind of Heisenberg uncertainty principle on steroids — poses serious threats to validity.

As touched on earlier, the challenges for formative assessment are threefold: (i) developing tasks that draw out and open up student thinking in relation to the content being taught; (ii) fitting in with a coherent approach to instruction; and (iii) developing support for teachers in interpreting and responding to student thinking. Formative assessment, at its core, is simply good instruction.

The assessment design challenges for summative assessment are related but different. In addition to the challenges that result from inadequacies in understanding and expressing a robust notion of mathematical proficiencies, in particular regarding mathematical practices, challenges from several other tensions are worth noting.

- *Splitting ideas* into components to be measured distinctly, *while integrating content with practices and other content to maintain integrity* of the mathematical proficiency to be assessed.
- Creating interesting compelling *tasks* that measure important aspects of proficiency, while needing to have prototypes *that can be developed, multiplied, modified, and otherwise scaled*.
- Reliably *ranking students* (norm-referenced-like testing), while keeping an eye on whether students are *learning the simple basics of their grade* (criterion-referenced-like testing).
- Generating a test that is *fair across local curricular variation* (general background test), while designing a test that can *detect change* due to teaching and learning (sensitive to local action).

To split and not to split

A tension exists between isolating discrete knowledge and skills to be measured and maintaining a more complete picture of mathematical performance. It is really a set of related tensions, for instance between low-level routine and higher-order thinking, between procedural skill and conceptual understanding, between clear specification of particulars and compelling categories with general force, and between content and practices. These are compounded by a tension between what is readily measured and what is hard to measure, between assessment expertise on the one hand and content expertise on the other. Too often the interests of psychometricians and content experts in the development of assessments are like a cold front and a warm front colliding and dancing around each other. Fair weather requires learning to play well together.

The field has come to recognize that mathematical proficiency involves both procedural skill and conceptual understanding and that both need to be tested. Less clear at this point in time is the issue of assessing both content standards and practice standards. In all likelihood, content and practice need to be assessed in tandem, but this rubs up against the need to assess distinctly—to identify what is known and not known. When a student gets such an item wrong, is it because the student does not understand the procedures and concepts of the content or because the student does not know how to engage in mathematical modeling or mathematical explanation? Perhaps this is an important question, or maybe it is the wrong question, both for assessment and for informing teaching and learning. The field needs a better understanding of the relationship between content and practices and its implications for designing instruction and assessment. One might argue that the content lives in the practice standards. Or, one might argue the reverse—that the practices live in the content standards.

Unfortunately, this problem is not limited to the issue of splitting or integrating content and practices. It resides everywhere. It is the same problem as deciding how narrowly to isolate a specific aspect of content knowledge and when to assess a package of related ideas. In large part, this is a problem inherited from the specification of standards. In

⁷This section draws from presentations given by Eva Baker (University of California, Los Angeles) and by Bill McCallum (University of Arizona) and Jason Zimba (Student Achievement Partners).

current efforts related to the Common Core, the standards say that the designers of curriculum and assessment need to connect the practices with content, but are silent about how that is to be done.

The practice standards refer to some of the deepest virtues of education. Socrates would not have recognized adding complex numbers as a learning outcome, but he probably would have recognized perseverance and precision as outcomes of education. These ideas have been around a long time. They are in the air, but they are fragile. These beasts do not live well in captivity. There are any number of pitfalls to assessing them. One is divorcing them from content, which makes practices ephemeral, even as it robs content of the richness it demands. Another is making a bureaucratic matrix with all of the standards and all of the practices, without an empty cell anywhere. The assessment consortia working to develop assessments aligned to the Common Core have made some smart decisions about how to handle the practices. They have had the courage to highlight a couple of really important ones — one about communicating mathematical reasoning and another about modeling. Their approach shows a good absence of matrix thinking.

Splitting is a complicated question. Consider for a moment the idea of splitting a standard. There might be good reasons to do so. But splitting can also be a threat to the rigor, focus, and coherence of the standard. For instance, rigor can suffer when we split, because students learn X and learn Y but never learn to put them together as we would like. Also, splitting leads to laundry lists, which are not helpful for focus. And in terms of coherence, the combining of different things in a standard is sometimes key to conveying a better understanding of ideas being expressed. To see the issue of coherence consider the following high school algebra standard from the Common Core.

Solve simple rational and radical equations in one variable, and give examples showing how extraneous solutions may arise.

It might be tempting to divide this standard into two standards, one about rational equations and another about radical equations. However, this standard falls under the following cluster heading.

Understand solving equations as a process of reasoning and explain the reasoning.

One reason for not splitting the standard is that keeping rational and radical equations together may help maintain a focus on the process of solving and the reasoning for the process, as is implied by the inclusion of “showing how extraneous solutions may arise,” which happens in related ways in both contexts. In writing standards, the problem of deciding when and where to split (or not) is a difficult one. Similarly, measurement requires isolating the “thing” to be measured, but this requires identifying clearly and explicitly what the “thing” is, which requires deciding when to split (or not).

Scaling while maintaining quality

Another challenge, down in the details but consequential, is that good items cannot be rare gems. The creation of reliable and valid measures usable with large numbers of students for high stakes requires that large numbers of items be produced. In efforts to develop new assessments, which necessitate using new formats and assessing aspects of proficiency not assessed in the past, it is crucial to have a steady stream of high-quality items. At present, though, item writing can be a rather illusive art. It requires balancing different kinds of constraints and sensibilities — mathematical, pedagogical, psychometric, psychological, social, and more. Likewise, good item writers are uncommon and viewed as uniquely gifted. While such people are a resource, they should not be made precious and the work they do needs to be professionalized (developed into a shared and technical activity).

To develop high-quality assessments at scale will require increased attention of two kinds. One is that efforts need to be focused as much on developing prototypes and item shells to use as starters for replication and modification as on the production of individual items. Second, the tacit understanding and skill of good item writers needs to be drawn out, made explicit, and used as the basis for assessment-development standards and training.

Assessing learning or ranking people

Unfortunately, principles of test development can exist in tension with important pedagogical and practical impulses. This problem is exacerbated by the fact that tests often have to fulfill multiple, competing purposes. For instance, tests are inevitably used to rank people. Doing so requires a test design that reliably separates people. However, the same tests are also used to make judgments about whether

standards are being met. With the goal that all students reach proficiency, separation is not the gold standard that it is taken to be among assessment experts who focus on producing reliable scores for individual students. Including an item that everyone might get correct provides little discriminating “test information” and takes up limited time and space. Another related tension arises because mathematical proficiency is not uni-dimensional and information that can inform instruction cannot be reduced to a single score. A reliable measure of any one dimension, or any one concept or skill, requires a collection of items. But, separate tests for all of the learning goals that are important to identify pedagogically and to assess in order to inform instruction would be impractical, especially in the context of reliably ranking people.

This tension is about differences in the purpose of assessment. The reality, though, is that any test will be used for multiple purposes, whether intended or not. Designing for every purpose is folly, but making strategic, clear-eyed choices about purposes and identifying appropriate uses of any test are important.

Coping with local variation

Developing assessments based on standards is a major change for assessment in this country. In the past, large-scale summative assessment has been designed to measure general background characteristics — what used to be called intelligence. In an educational system in which what is being taught and learned varies dramatically from school to school, and from teacher to teacher, this is the only option for creating a reliable common measure (Cohen, 2010). As the country moves to standards and shifts the assessment focus from norm-referenced assessment of general characteristics (whether “intelligence” or “aptitude” in a content area) to criteria-referenced assessment of standards, it may be in for a rude awakening. Such tests may reveal, much more dramatically, how little actual mathematics is being effectively taught and learned. Mathematicians and educators are aware of this malaise in general, but this will make it specific (and personal).

Such a test is likely to give much more focused and detailed information. Teachers and schools poorly prepared to use information for making improvements may become disheartened. An advantage of a norm-referenced test, that spreads people out along a general, uni-dimensional factor, is that it guarantees that some do well and some do poorly. A criteria-referenced test may well expose with

great clarity when and where teaching and learning are and are not occurring. On the surface, this may seem like a good thing, but if it threatens the status quo or exposes problems without clear solutions, resistance may be overwhelming. Information about success and failure is vital to improvement, but insufficient capacity can lead to feelings of inadequacy, despondence, and avoidance.

As the country moves to standards and shifts the assessment focus from norm-referenced assessment of general characteristics (whether “intelligence” or “aptitude” in a content area) to criteria-referenced assessment of standards, it may be in for a rude awakening. Such tests may reveal, much more dramatically, how little actual mathematics is being effectively taught and learned.

These tensions suggest several key questions for the development of assessment:

- Will the tests be long enough to assess problem solving and perseverance?
- Will testing formats allow for assessing, producing, and critiquing extended chains of reasoning?
- What reporting formats provide adequate feedback for informing instruction?
- What governs assessment decisions? Mathematics or psychometrics? For example:
 - If the assessment uses computer adaptive testing, how will it maintain attention to and balance of different practices and content?
 - If everything is computer-based, how do students draw mathematical representations?
 - Who makes the big decisions: people in mathematics and mathematics education or those from the testing community?

Norms and structures for productive work across professional communities⁸

Improving the assessment of mathematical proficiency will require a coordinated effort among several disparate professional communities. Bryk, Gomez, and Grunow (2011) argue that such work requires diverse professional expertise and needs to be organized for the task at hand.

Another question is “Who should be doing the work?” If the listing of problem parts above captured even a small part of the problem ecology, then a very diverse collegueship of expertise will be necessary to make progress (Bryk and Gomez 2008). Furthermore, these actors must be organized in ways that enhance the efficacy of individual efforts, align those efforts, and increase the likelihood that a collection of such actions might accumulate toward efficacious solutions. While innovations abound in education, we argue that the field suffers from a lack of purposeful collective action. Instead, actors work with different theories of the same problem, activities are siloed, and local solutions remain local. (pp. 129-130)

Vying for political power and debating issues in the popular press are unlikely to improve outcomes. Nor is gathering a diverse group of people together in a single room and telling them to make a test. Instead, working across professional communities requires the creation of norms and structures for productive work. It requires a shared understanding of a practical problem, a focus on relevant data, structures for the engagement of key perspectives, and respectful collective engagement in making sense of information, including that which takes into account the needs of users and the context of use — what von Hippel (2005) argues is “sticky” information because it is specific to particular situations and cannot be readily transferred to other contexts or glossed with a general label.

This will be much more than the current assembly line model for assessment development, with mathematicians, teachers, policymakers, and psychometricians responsible for injecting their specific expertise into a product handed down the line. It calls for growing a network of percipient

professionals to deliberate thoughtfully about what an item is meant to elicit, whether it does, and reasons for saying it does what is claimed, where these different experts respectfully negotiate and find balance among competing concerns. It is about merging and melding different expertise in the midst of improvement work that attends to overall coherence.

Such work benefits from the development of methods for working together, tools for organizing the work, and boundary objects that meaningfully serve the work within and among communities (Akkerman & Bakker, 2011). For instance, the Common Core standards have potential for playing such a role. In discussing an assessment item, people can refer back to the standards as a common reference point that can coordinate exchanges among teachers, mathematicians, and psychometricians. To date, the work of the assessment consortia has benefited greatly from the use of the document in this way. A problem goes up on the screen. It is being proposed for the item bank. It is supposed to assess a standard. Throat clearing. Eventually people get back to: here is the evidence to be elicited by this item in relation to this standard, does it elicit the evidence and what are reasons for saying it does? Such conversations cannot happen on an assembly line. These conversations coordinate professional expertise and build actionable knowledge.

There are of course inherent difficulties in such work. Take, for instance, language used in an item. Everyone agrees that unnecessary language in tasks should be avoided, but deciding what is necessary and unnecessary can be an important question. An assessment specialist may be inclined to minimize the reading load. This inclination may be shaped by assumptions about uni-dimensionality as well as conceptions about mathematics. However, assessing proficiency with mathematical language and with mapping between language and mathematical symbols is important. So, deep and important questions about the measured construct may surface. There may also be voices in the room that think that unless a problem has a context it is not a good problem. Experts need to negotiate and defend interpretations based on shared information and agreed-upon documentation of the assessment problem being addressed.

⁸ This section draws from presentations given by Kristin Umland (University of New Mexico), Shelbi Cole (Smarter Balance Assessment Consortium), and Doug Sovde (Partnership for Assessment of Readiness for College and Careers).

Working together to develop quality items

At the heart of quality assessments are quality items. Item development is one of the most important contexts for collective work by members of the relevant professional communities. Item development is where we need to get the math right. It sounds so simple and yet making it happen is complex.

Engaging professional communities in writing, reviewing, and discussing items is essential for several reasons. One reason is the need for such large numbers of affordable items. Item writing and reviewing is professional work that involves significant content knowledge and ought to be routinely and efficiently carried out by teams of professionals. More importantly though, as conveyed throughout this document, understanding of mathematical proficiencies is limited, as is understanding of the design of items to measure mathematical proficiencies and the evidence-based arguments that link item performance to claims. In addition, as assessments explore the potential for a much wider range of item formats afforded by technology it will be important to have the insights and concerns of different professional communities contributing to the work. Engagement in the production of quality items is an important activity for developing the understanding that will be needed for significant improvement of assessments and ultimately of teaching and learning. It will require a combination of expertise held by different professional communities and will require productive engagement and exchange.

Establishing clear structures for engagement of different professional perspectives can help coordinate expertise, support productive exchange, and lead to quality products. As a case of this, the Illustrative Mathematics Project (www.illustrativemathematics.org) has created an online community engaged in writing mathematics tasks to illustrate the Common Core standards. Contributors have been recruited from different professional communities. Each proposed task is accompanied by potential solutions and a commentary that explicitly states the intended pedagogical use, the intended mathematical focus, the standard being illustrated, and relationships to other standards. Every task must pass two reviews, one from a pedagogical perspective and one from a mathematical perspective, with explicit criteria and trained reviewers for each. In this work, people's participation is organized to match specific expertise and to coordinate with others. The structured commentary and explicit review criteria support productive exchange across professional boundaries.

Similar approaches are needed to develop assessments. For instance, a useful commentary to accompany an assessment item might address the following.

- *Aspect of proficiency* that the item is designed to assess, including reference to important Common Core content and practices.
- *Purpose of the item* and the *context in which* it is to be used.
- *Likely student responses* (correct, incorrect, and partial proficiency) and how they should be interpreted.
- *Evaluation criteria*, for student responses, including clear, defensible, correct answers and/or rubrics.

Another useful structure for coordinating work on creating assessments aligned to the Common Core has been developed by members of the Partnership for Assessment of Readiness for College and Career. As they develop their assessment, they consider the claims the assessment should support. Independent from the extent of agreement about their existing claim structure, it is important to note that the generation of a public claim structure contributes to ongoing dialog about both the content and purpose of the assessment. Then, as an intermediate step linking the claim structure to the development of items, they generate evidence statements about what students might know and be able to do (the evidence) that would permit making the claims described in the claim structure. As items are produced, arguments need to explain how the item will elicit compelling evidence. This structure draws directly from the evidence-centered-design approach developed by Mislevy and his colleagues (Huff, Steinberg, & Matts, 2010; Mislevy, 2006).

The development of structured commentaries about key features and evidentiary arguments that pass muster with key professional communities, for instance teachers, mathematicians, and assessment experts, can improve the overall quality of items by harnessing the expertise of different professionals. Investing in organized, cross-community production, including explicit criteria, evidentiary arguments, profession-specific roles, and procedures for collective vetting, provides important ground for improving assessment.

Developing summative assessment item prototypes

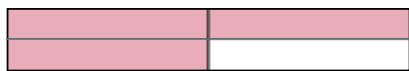
Given the challenges discussed earlier, high-quality large-scale high-stakes assessment will require collaboration and innovation. It will require the development of item types that both measure knowledge and support instruction (at least avoid derailing it), that size up what students know and do not know, and reveal ways in which students think about mathematical content. New designs will be needed for assessing mathematical practices. Prototypes will need to be structured in ways that attend to rigorous mathematical content and to pedagogical sensibilities. And all of this will need to be done with regard for validity, reliability, fairness, and affordability.

This work will require respect for different perspectives, perhaps even occasional suspension of one's convictions and experimentation with the adoption of perspectives different from one's own. Many individuals, on first approaching this work, are inclined to think that they know what a good mathematics problem (or assessment item) is and is not. Initially, heavy-handed opinions are common. Typically, though, as one provides a review, hears responses, and begins to take in the deliberations and improvements made, one begins to get a sense of one's own expertise, other's expertise, and the role of different expertise in the improvement of items.

Combined professional expertise is essential for the development of items for large-scale high-stakes assessments that effectively measure mathematical proficiencies. Unfortunately, there is currently limited understanding of both the proficiency and approaches for measuring it. For example, what types of items might assess the cornerstones of constructing mathematical arguments or solving problems using modeling? To measure proficiency with a practice of constructing mathematical arguments, simply adding "explain your answer" to existing items is ineffective.

As a simplified example of these dynamics, consider the following (poor) item as a starting point.

What fraction is represented by the shaded area?

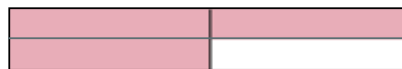


This item has many problems. One issue is that it makes an assumption out of the very thing that is to be taught, learned, and assessed — the definition of a fraction as relative size to a given whole. Without specifying the whole, or unit, it is not reasonable to ask what fraction is represented. If one of the four rectangles constituting the larger figure is the whole, then the shaded area is $3/1$. Not stating the

assumption that the larger rectangle comprised of the four smaller rectangles is the whole, gives students who are just learning this content the impression that mathematics is an arbitrary game and it fosters instruction that undermines mathematical meaning.

To support someone proposing such an item, one could respond with additional questions that might help to highlight issues.

What fraction is represented by the shaded area?



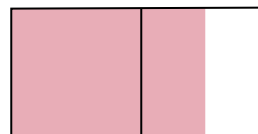
Four students give responses. Explain what must be true for each student to be correct.

- Student A: $3/4$
- Student B: $3/2$
- Student C: $3/1$
- Student D: $30/1$

Mathematical expertise, including that represented in the Common Core standards, provides a more precise basis for considering both the content and what needs to be taught, learned, and assessed. Here is an excerpt from *Progressions for the Common Core State Standards in Mathematics* (Common Core Standards Writing Team, 2013, p. 3).

3.NF.1 Understand a fraction $1/b$ as the quantity formed by 1 part when a whole is partitioned into b equal parts; understand a fraction a/b as the quantity formed by a parts of size $1/b$.

The importance of specifying the whole




Without specifying the whole it is not reasonable to ask what fraction is represented by the shaded area. If the left square is the whole, the shaded area represents the fraction $3/2$; if the entire rectangle is the whole, the shaded area represents $3/4$.

It gives a standard that provides a definition for a fraction and describes the importance of attending to the whole. It points out that the task of specifying the whole provides opportunities for the mathematical practices of attending to precision (MP6) and constructing viable arguments and critiquing the reasoning of others (MP3).


The Smarter Balance Assessment Consortium has explored an item addressing this mathematics in a drag-and-drop environment where students can create representations.

Look at the fraction model shown



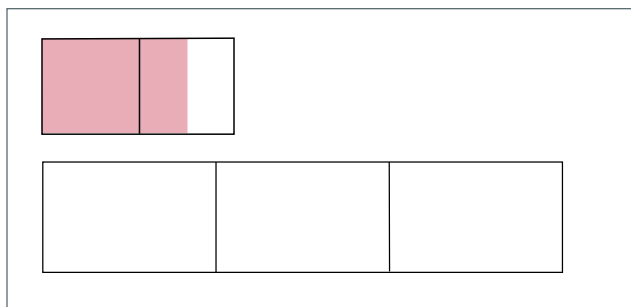
The shaded area represents $\frac{3}{2}$. Drag the figures below to make a model that represents $3 \times \frac{3}{2}$.

A B C D



This item assesses students’ skill in modeling the computation, but it also provides additional information that can inform a picture of what students know and do not know. For instance, consider the work of three different students.

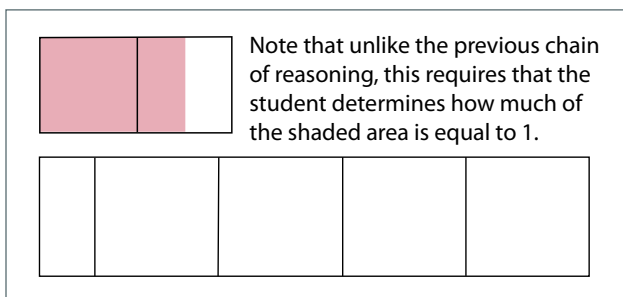
Student 1 drags three of shape B, which is equal in area to the shaded region. This student probably has good understanding of cluster 5.NF.B — *apply and extend previous understandings of multiplication and division to multiply and divide fractions*.



Such a student can say, “Well I’m simply multiplying three times the area equivalent to three halves.” He or she can drag three figures that are equal to three halves into the space to create a model for three times three halves. The student knows that $3 \times \frac{3}{2}$ is equal to 3 iterations of $\frac{3}{2}$. Calculation of the product is not necessary because of the student’s understanding of whole-number multiplication and its extension to fractions.

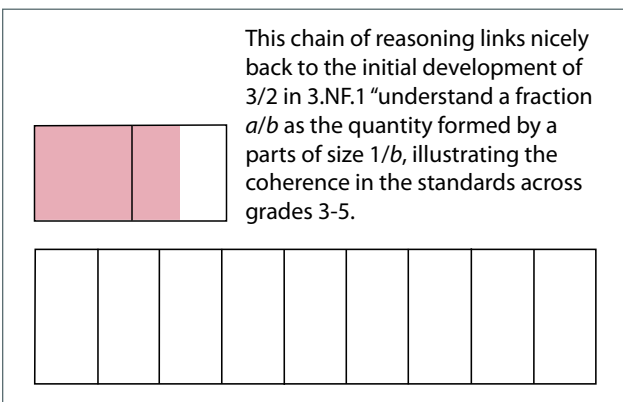
Student 2 might go through a fifth-grade process. This student might know how to calculate 3 and $\frac{3}{2}$. The

student might convert the result to a mixed number to get four and a half. But the reasoning cannot stop there because the result must still be interpreted. It is four and a half, but what is one? The item requires further reasoning, where the student would have to figure out what the “one” is and how to build the model.



Note that unlike the previous chain of reasoning, this requires that the student determines how much of the shaded area is equal to 1.

In contrast, student 3 below builds an argument based on the definition. Even though the item is aligned to the fifth-grade standards, given the way a unit fraction is developed in third grade, a third-grade student could reason that if the original figure is three halves, then one half is known and can be applied to represent three halves and then three copies of three halves, or nine halves.



This chain of reasoning links nicely back to the initial development of $\frac{3}{2}$ in 3.NF.1 “understand a fraction $\frac{a}{b}$ as the quantity formed by a parts of size $\frac{1}{b}$, illustrating the coherence in the standards across grades 3-5.

There are multiple ways to reason about this item. It does not align neatly to any one single standard, but it aligns nicely to the idea of multiplication and division of fractions as an extension of prior work and to the notion of coherence across standards. As an assessment item, it suggests reliable uni-dimensional score information as well as information about student thinking. In addition, it suggests a class of items that provide valuable information and could be produced at scale.

Another prototypic item is useful for assessing student understanding of functions and considering functions in terms of input and output. This “simulation item” available on the Smarter Balance website asks students to match sales-tax calculators to states with known sales taxes by entering a purchase price and using the result to determine which match.

Different states have different sales tax rates. Three States have online calculators to compute sales tax on a purchase. Use the following steps to match each calculator with the correct state.

- Select Calculator A, B, or C.
- Enter a Purchase Price.
- Then select “Find Sales Tax” to compute the sales tax for that purchase price.

You may use the calculators as many times as you need to solve the problem to the right.

Different states and their sales tax rates are shown.

Drag each calculator into the correct row to show which state can use it to calculate sales tax.

State	Sales Tax Rate	Calculator
Illinois	6.250%	
Indiana	7.000%	
Kansas	6.300%	
Maine	5.000%	
Maryland	6.000%	
Minnesota	6.875%	

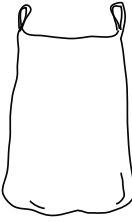
A wise student might put in 100 because it will give the sales tax on \$100.00.

This is a seventh-grade simulation item but similar items having students think about the function underlying the technology could be used across the grades. Given the ubiquity of apps, it would be helpful if students understood whether the technology is functioning as intended.

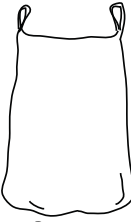
Here is another fourth-grade item using drag and drop technology. It asks students to fill in bags with juice bottles that weigh $3\frac{5}{8}$ lb each so the weight is within the given interval.

Jared is testing how much weight a bag can hold. He plans to put juice bottles into three bags. He wants each bag to have a total weight within the given range.

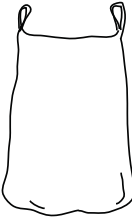
- Drag juice bottles into each bag so that the weight is within a given range.
- Leave the bag empty if the given range is not possible using juice bottles.



Between 6 lb and 7 lb



Between 10 lb and 11 lb



Between 14 lb and 15 lb

This item reflects an important goal of the standards, which is to recognize that a fraction is a number. The item is not just about getting an answer from multiplying a fraction by a whole number but recognizing that the resulting number is located somewhere on the number line. Another important feature of this item is that English learners or struggling readers can get an understanding of the item and what it is asking simply by looking at the diagram and the picture. A critical piece is the last part about leaving the bag empty if the given range is not possible.

The items above could probably be improved, but they give a sense of what it might mean to develop prototype items that address important standards in deep and compelling ways and could be used to produce high-quality large-scale assessments. They also give a sense of potential innovations for item formats that technology might afford and of the thinking and collective problem solving that might be needed to produce a wide range of prototypes that might populate large-scale assessments aligned with the Common Core standards.

Items can often be tweaked for improvement. Consider the following standards in domain 3.MD.

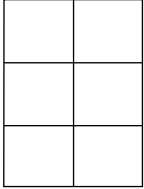
Standards in domain 3.MD

Geometric measurement: understand concepts of area and relate area to multiplication and to addition.

5. Recognize area as an attribute of plane figures and understand concepts of area measurement.
 - a. A square with side length 1 unit, called “a unit square,” is said to have “one square unit” of area, and can be used to measure area.
 - b. A plane figure which can be covered without gaps or overlaps by n unit squares is said to have an area of n square units.
6. Measure areas by counting unit squares (square cm, square m, square in, square ft, and improvised units).

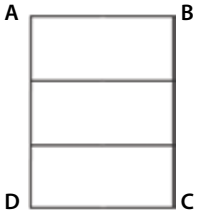
Here is a simple item that might be proposed to assess these standards.

Give the area of the figure in square units.



The issue with this item is that it focuses on the idea of area as counting square units but without attending to the underlying ideas or representing them with mathematical precision. Here is a revision.

The area of Rectangle ABCD is 24 square units. Draw a picture of 1 square unit.



This revision shifts the focus to the coordination between the specification of a unit and the determined area. Now it is precise enough to give a response. It engages a different line of thinking about the idea of using a square unit to measure a two-dimensional surface, one more in line with the standards. It also involves (implicitly) ideas about equivalent fractions (proportions) and how area behaves with rescaling.

Another consideration in the design of items for summative assessment considers ways of attending to content coherence across the grades. Here is a standard from first grade in domain 1.NO.

7. Understand the meaning of the equal sign, and determine if equations involving addition and subtraction are true or false. *For example, which of the following equations are true and which are false?*
 $6 = 6$, $7 = 8 - 1$, $5 + 2 = 2 + 5$, $4 + 1 = 5 + 2$.

The examples are simple, but anyone who has taught algebra will recognize the continued importance of this content into high school. If this grade one standard falls out of the curriculum after first grade and is not embedded in ongoing mathematical work across the grades, we are in trouble. By eighth grade, students will need to be facile with the meaning of the equal sign in the context of algebra and the solving of equations. They should not be surprised by expressions on the right-hand side of the equation that are not simply the answer. The development of understanding of the equal sign needs to start early, but it cannot disappear. And, it should probably be assessed across the grades.

A third-grade item might address fluency, such as one asking students to determine which equations are true.

$$3 \times 8 = 20 + 4$$

$$50 \div 10 = 5 \times 1$$

$$9 \times 9 = 8 \times 10$$

This item addresses the computational fluency central to third grade while maintaining regard for understanding of the equal sign.

Here is a similar item at the fifth-grade level in the context of fraction computation, again where students are to determine which equations are true.

$$\frac{1}{2} \times \frac{1}{3} = \frac{3}{6} \times \frac{1}{3}$$

$$\frac{2}{2} \times \frac{1}{3} = \frac{3}{6} \times \frac{1}{3}$$

Alternatively, as a way of making the item less computational and engaging understanding of the equal sign in different way, the following item asks students to fill in the box in a way that makes the equation true.

$$\frac{1}{2} \times \frac{1}{3} = \frac{\square}{6} \times \frac{1}{3}$$

This item requires that students use the structure of the equation. They must use the fact that one-half times one-third equals three-sixths times one-third without actually multiplying. The design of the item is appealing because there are not many good ways for a fifth grader to do this without reasoning about the structure and meaning of the equation. As an assessment item, its influence on instruction — its WYTIWYG factor — is likely to be more positive than the previous. It suggests that paying attention to structure and looking thoughtfully at the problem before you start solving really matter.

In continuing to explore the idea of items developed to address coherence of content across the grades, here is an eighth-grade item.

Solve for x.

$$3x + 17 = 3x + 12$$

It might well be good if eighth-graders were able to respond that this equation has no solution, not because they have been taught to keep an eye out for problems that do not have a solution, but because they understand the equal sign and are inclined to consider structural features of equations.

Here is another eighth-grade item that incorporates regard for the equal sign in the context of other content.

$$x^4 - 5x^3 + x^2 + 2x + 1 =$$

Drag the correct expression to make a true equation.


$$x^3 + (x+1)^2 + x^4 - 6x^3$$

$$x^4 - 3x^3 + 2x^3 + x^2 + 2x + 1$$

$$x^4 - 5x^3 + x + x + 2x + 1$$

Although students should be expected to retain what they have learned in the past, it is not the case that all preceding standards should be assessed at each grade level. However, as the professional community continues to design and scrutinize new assessments aligned with the Common Core standards, it needs to consider what coherence means and ways to more fully specify it. The items above provide a glimpse of one way of pursuing this work. In addition, these items lend themselves to spinning off replicas and variations that would support assessment at scale.

Another important issue for the development of summative assessment prototypes is getting smarter about ways of efficiently processing the full range of what can increasingly easily be captured electronically. Here is an item where there is a tent 8 feet by 10 feet and sleeping bags 3 feet by 6 feet. A hypothetical student, Dru, calculates the area of four sleeping bags to be 72 square feet and concludes that the tent will fit 4 people because 72 is less than 80, which is the area of the tent. A second student, Teller, says they will not fit. Students are asked to decide which is correct and explain why.


Dru and Teller had a tent that is 8-feet by 10-feet. Each adult has a sleeping bag that is 3-feet by 6-feet. 

a. Dru said that four adults would fit in the tent. Each adult needs 18 square feet of floor space. $18 + 18 + 18 + 18 = 72$. The tent is 80 square feet, so there is room to spare. Teller said that he tried and could not get four adults to fit in Tent C.

Who is right?

b. Explain.

Teller is right because you can fit three but there is more sq. ft. left over enough to equal 18 just it wasn't a 3x6 it was a 9x2

 *Gr0 9x2*
Red = 3x6

Garin, M. K., Ciro, T. M., Chapin, S., Copley, J. V., & Sheffield, L. J. (2008). Project M2: Using Everyday Measures: Measuring with the Meerlots from Project M2: Mentoring Young Mathematicians series.

The above item is a nice example of practice MP.3 about constructing viable arguments and critiquing the reasoning of others. It also raises the issue of needing to figure out effective ways of having students import drawings and computations that are natural to do in a paper-and-pencil context, perhaps with a drawing program or using tablet technology. Eventually, it would be good to have scoring software that can do more than simply look for 9 x 2 and can consider multiple patterns in student responses that would be deemed correct. The data mining techniques and other approaches used by artificial intelligence scoring algorithms are becoming more sophisticated because of the increased demand. Imagining what might be, in ways that explore the interactive design space of item structure, technology, and mathematical content, will be essential for making significant progress in the development of useful assessments, greater understanding of mathematical proficiencies, and the improvement of mathematics teaching and learning.

Views of assessment in and from the classroom⁹

One important perspective to incorporate into deliberation on the design of assessments is that of teachers. This section provides a dialog between two teachers, Eyal and Mel, as they describe ways in which they work to implement high-level tasks and help their student learn to demonstrate mathematical practices as characterized in the Common Core standards. Eyal and Mel begin by setting the stage with a particular problem.

Eyal: Mel, there's a problem I've been thinking about. I love it. It's adapted from the book, *Innumeracy*. I think you're going to like it. Think about this. If we took all of the blood from everyone in the world and poured it over Central Park, what depth would the blood reach?

Investigate



If all the blood from everyone in the world was poured over Central Park, what depth would the blood reach?

– Adapted from, *Innumeracy* by John Allen Paulos

Melanie: Oh my gosh, Eyal, that's disgusting, and I love it! I want to send this to the substitute teacher who's teaching my math class tomorrow because I think this problem is so engaging my students will go crazy. There are so many points of entry. There's no way that if I give this problem to my students my class will be anything other than a success. I just know it. ... Or, maybe not. ...

Teachers face predictable challenges in implementing complex, open-ended tasks. What are potential pitfalls? One is that students may give up and classroom management may become an issue. Another is that students need constant reassurance that their answer is right or that their solution path is okay. Little cues, such as nodding or saying thank you can shape student engagement. The Common Core standards include the mathematical practice of perseverance in problem solving. However, perseverance does not just happen. Telling students to persevere is not the same thing as teaching them what it is and how to do it.

Teachers need to have effective approaches to help cultivate it in students. Perseverance needs to be taught and figuring out how to do so is demanding professional work.

Eyal and Mel go on to describe how they teach the first mathematical practice given in the Common Core standards.

Standards for mathematical practice

CCSS.Math.Practice.MP1 Make sense of problems and persevere in solving them.

Mathematically proficient students start by explaining to themselves the meaning of a problem and looking for entry points to its solution. They analyze givens, constraints, relationships, and goals. They make conjectures about the form and meaning of the solution and plan a solution pathway rather than simply jumping into a solution attempt. They consider analogous problems, and try special cases and simpler forms to gain insight into its solution. They monitor and evaluate their progress as they work on a problem. They adjust their plan as they go along. They persevere in solving a problem. They check their answers to problems using a different method, and they continually ask themselves, "Does this make sense?" They can understand the approaches of others to solving complex problems and identify correspondences between different approaches.

Make sense of problems and persevere in solving them

Eyal: So what does that really look like? What are things we do in our classrooms to think about this idea and to help cultivate it?

Melanie: All of the things we are going to talk about are ways we give constant feedback to our students about their progress towards mastering this standard. So this is an example of a teamwork rubric we use in our geometry class.

Teamwork rubric

Math Thinking

Everyone demands to understand
If you listen to the team, you hear
people explaining their thinking

Fix #4 before moving on. Also, I really love how well you work together but make sure all the work is shown on everyone's packet!

Melanie: The expectations are really explicit at the top. Students know that we are really looking for their math thinking. "Everyone demands to understand." That's the language we use in our

⁹ This section is adapted from a presentation given by Eyal Wallenberg and Melanie Smith (The Urban Assembly School for Law and Justice), much of it taken verbatim in hopes of conveying an unfiltered practitioner perspective.

classrooms. If we walk by and listen to the team, we're going to hear people explaining their thinking. And, so, students know that if I walk by and I see these things and that's evidence they're progressing towards mastery of this standard. And then, on a daily basis we give them really specific feedback about their progress. So, how did your group do, what do you need to do to improve upon it, giving some "warm" feedback, some "cool" feedback, and we expect students then to implement the changes that we suggest the very next day.

Eyal: And what I really like about this is — it's all written feedback. Another way we give in-the-moment, teacher-to-student feedback is by live-tweeting our class, but it's not really tweeting. We just walk over to our laptop, which is connected to a smart board, and we'll write down quotes that we hear or observations that we make, while students are working. So, for example, if I'm listening to team 3, I might hear somebody say, "I dunno, Kimberly, what do you think?" Or I might note that the group is really persisting, even though one strategy they used didn't work. And, sadly, in team 3, somebody was talking out of the group, so that shows up in red.

Live "tweeting" of quotes and observations

The screenshot shows a list of six teams with their work and observations:

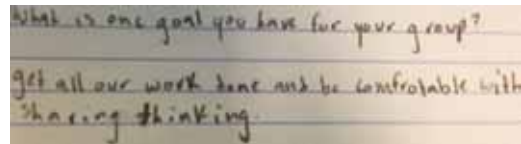
- Team 1:** "Well - how did you get that?" "Is everyone ready to go on?" using multiple strategies.
- Team 2:** great debate about what strategy to use
good discussion about points of reference for number of students in NYC
- Team 3:** "I dunno - Kimberley, what do you think?"
group is persisting, even though first strategy did not work. *talk outside the group*
- Team 5:** lovely tone - very supportive *talk on other topics*
"If that's the case, then could we use the other answer choices..."
- Team 6:** "I don't get what you said - can you explain again?"
"Why did you multiply there?"

Eyal: Another way we'll have feedback is student-to-student feedback. And one way we'll do that is by using video. So we'll take really short clips, with our phones, and we'll put it on our document reader so we can quickly show it to the class, like 5 minutes after it happened. And we'll have students give "warm" specific feedback to other students about things that they heard. When they're listening to their peers, they might say, "Tywan,

I really heard him explaining his thinking," or "I heard that Jaycee was playing the skeptic."

Melanie: And last we ask students to give themselves feedback on their progress. And this comes in the form of goal setting. So, working in teams is really difficult and we think that working in teams is also the best way for students to develop this skill of perseverance. So at the beginning of a unit, we'll ask them to set a goal for their group. So, Zyairah's goal for her group, she wanted to

"...be comfortable with sharing thinking."



"...make sure everyone understands."

be sure that everybody was comfortable sharing their thinking. And Iyonna wanted to make sure everyone understands before they move on.

Melanie: We're doing all of these things to make sure that students see the importance of perseverance and to realize that that's something real mathematicians do. Real mathematicians have to persevere when they solve a problem. They don't solve a problem in 30 seconds, they don't solve in one hour. They don't even solve it in two days. Maybe it takes two years.

Eyal: So far, we looked at the different teacher strategies we use in the classroom to address this idea that students will give up completely when working on complex problems, and we'll now go to our second one which is students need constant reassurance that their answer is right or that what they're doing is okay. We think this one connects really strongly with the standard for mathematical practice around critiquing the reasoning of others. And the reason we see that connection is that, if students are in the habit of asking if other people's arguments make sense, they'll start asking themselves, does my argument make sense, is my line of reasoning logical?

Standards for mathematical practice

CCSS.Math.Practice.MP3 Construct viable arguments and critique the reasoning of others.

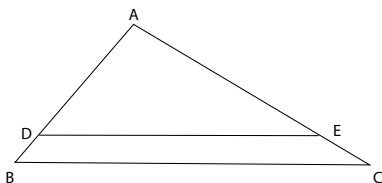
Mathematically proficient students understand and use stated assumptions, definitions, and previously established results in constructing arguments. They make conjectures and build a logical progression of statements to explore the truth of their conjectures. They are able to analyze situations by their conclusions, reason inductively, and justify their conclusions. They justify their conclusions, reason inductively, and justify their conclusions. They justify their conclusions, reason inductively, and justify their conclusions. They justify their conclusions, reason inductively, and justify their conclusions.

Construct viable arguments and critique the reasoning of others.

Melanie: And that's a really important skill for our students to master in order to do well on the standardized tests that exist right now. So, for example, this is an item from the New York State Geometry Regents exam from last June. Students are asked to solve the length of EC.

NYS Geometry Regents, June 2012

16 In the diagram of $\triangle ABC$ shown below, $DE \parallel BC$.



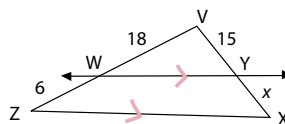
If $AB = 10$, $AD = 8$, and $AE = 12$, what is the length of EC ?

- (1) 6
- (2) 2
- (3) 3
- (4) 15

Melanie: This problem doesn't explicitly tell them to construct a viable argument and critique the reasoning of others, so when we are teaching this concept, we want them to be thinking about that as they're solving this problem. So, we'll present this same problem to them in a slightly different way. We'll tell them about Alan and Noah, who did their homework last night and they both had to solve this problem, but they did it in totally different ways. What Alan's doing looks pretty familiar to me, I think I've seen that in class, we've talked about it, but I've never seen anything like what Noah's doing. And, I'm not sure if either of them is right, maybe they're both right, maybe neither of them is right. And we'll ask students to examine both of these and to think about who's right and to justify why.

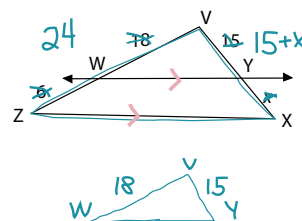
And that way, when they read the problem on the regents, they are comfortable and know what to do.

Alan's Answer: $\frac{18}{6} = \frac{15}{x}$



Noah's Answer:

$$\frac{24}{18} = \frac{15+x}{15}$$



Eyal: Students really dig this idea of playing the skeptic with other people. And, another thing we'll do, not a lot but every once in a while, we'll have students develop a mathematical argument that they then present to a small group and they'll purposely make one part of their reasoning not quite right. And then it really creates this need to listen, because you know that something being presented to you doesn't quite make sense. The main theme in all of this is that we want to build student capacity around critiquing other people's reasoning, so they can critique their own reasoning. We want them to be really hungry around that habit.

Playing the skeptic



Melanie: So this is a type of work we are doing on a daily basis with our students to develop their capacity to persevere and problem solve, to critique the reasoning of others, and to construct viable arguments. And, all of this is formative assessment. We don't assign numbers or letter grades, and all of our feedback is given in a written form, or it's an activity designed to have them practice this type of work without necessarily getting any direct feedback from the teacher at all. Then the next part of the teacher's job is to actually do some grading, right? We have to give our students grades. And often with summative assessments we have to assign a letter grade or a number grade, and that's where things get a little more challenging.

Eyal: So, here's an example, just a little task that a colleague of ours gives, where students tie knots in ropes and look at the relationship between how many knots and the new length of the rope. Eventually they're going to come up with ideas about when it's possible to tie the same number of knots and get the same length.

Knots and ropes

Imagine that you and your friend each own a dog and you like to take them for walks together. You and your friend would like to shorten your dogs' leashes so that it is easier for you to pull the dogs apart when they tangle.

The leashes are different lengths and thicknesses.

Can you and your friend tie the same number of knots in each leash and end up with leashes that are the same length?



Eyal: Here is just a small excerpt from Jabari's work. These are all his words. So he made a table and he looked at how much the rope shrank every time we tied a new knot. He referred to the shrinking as the "drop rate" and eventually found an answer by finding the average drop rate.

Jabari's work

I discovered that the **best way to solve this was to find a steady drop of the average so I found the averages** of both ropes.

Drop Rate	Thick & Long	Drop Rate	Thin & Short
3.5	55.5	3.5	35.5
5	52	4	32
5	47	3	25
4	38	4	22

After finding the drop rate of each number I **took all the drop rates added them up and then divided them. . .**

Now that I found the average for both ropes **the next step was to choose a length to start from for both ropes. . .** I picked the numbers 80.1 and 74.5 as my starting points.

I steadily subtracted the average drop rate for the Thick & Long (4.3) and the average drop rate for the Thin & Short (3.5) until I got to an equal number of knots and an equal length.

Melanie: So this is a standardized assessment and it's aligned with a mathematical practice from the Common Core, and because it's a standardized assessment, to grade it we're given a rubric by the company that developed this assessment. It's given to both the teachers and the students. It's not specific to math. One page addresses five domains of student work: problem formulation, researching the problem, interpreting results, communicating the solution, and precision and accuracy. Another page addresses characteristics of the work product: insight, efficiency, idea generation, concept formation, integration, and solution seeking.

Melanie: Zooming in on the lowest three levels for the characteristics of idea generation and concept formation, we see that we need to determine whether to evaluate Jabari's idea generation and concept formation as an emerging novice, a novice, or an accomplished novice.

<p>Accomplished Novice – 3</p> <p>Work product shows consistent evidence of proper use of conventional ideas</p> <p>Work product uses and incorporates concepts in a limited fashion to organize and explain findings</p>	
<p>Accomplished Novice – 3</p> <p>Work product shows some evidence of proper use of conventional ideas</p> <p>Work product does not use or incorporate concepts and/or does not explain findings coherently</p>	
<p>Accomplished Novice – 3</p> <p>Work product shows little or no evidence of proper use of conventional ideas</p> <p>Work product organizes and explains findings in a way that does not use concepts in any significant fashion</p>	

Characteristics of work product					
Insight	Efficiency	Idea Generation	Concept Formation	Integration	Solution Seeking
Emerging Expert – 7					
Work product shows strong evidence of an intuitive sense of subject-area rules to demonstrate insight	Work product treats task highly efficiently, few ways it could be done more efficiently	Work product shows strong evidence of novel or creative use of controversial ideas and/or strong evidence of unique or innovative ideas	Work product shows strong evidence of conscious design around a set of core concepts to organize and explain findings	Work product uses integration and connection among its elements in a highly efficient fashion that is readily apparent	Work product shows strong evidence of a cogent, coherent solution strategy for the problem
Accomplished Strategic Thinker – 6					
Work project shows evidence of a more intuitive use of subject-area rules to gain insight beyond literal application of rules	Work product treats task efficiently with a few minor or inconsequential inefficiencies	Work product shows strong evidence of novel or creative use of convention ideas and/or clear evidence of original ideas	Work product is purposely and intentionally structured around a set of core concepts to organize and explain findings	Work product is integrated and connected in an effective fashion	Work product shows evidence of a cognitive, coherent solution strategy for the problem
Strategic Thinker – 5					
Work product shows evidence of applying subject-area rules in an insightful fashion beyond literal application of rules	Work product is predominantly efficient in its treatment of the task, but some inefficiency may still be apparent	Work product shows strong evidence of proper use of conventional ideas and some evidence of original or novel ideas	Work product uses and incorporates a set of core concepts to organize and explain findings	Work product shows convincing evidence of imagination or connection among all its elements	Work product shows evidence of a full and complete solution strategy for the problem
Emerging Strategic Thinker – 4					
Work product shows some evidence of applying subject-area rules in an insightful fashion beyond literal application of the rules	Work product shows evidence of efficiencies in its treatment of the task, but has several areas where efficiency could be improved	Work product shows consistent evidence of proper use of convention ideas and at least some evidence of original ideas or novel variations on conventional ideas	Work product uses and incorporates concepts to organize and explain findings, but with some inconsistency	Work product shows evidence of integration or connection among all its elements with some pieces that are not well integrated or connected	Work product comes very close to a complete solution strategy
Accomplished Novice – 3					
Work product applies subject-area rules correctly and uses rules to demonstrate limited insight into subject area	Work product has areas of efficiency in its treatment of the task, but also contains significant inefficiencies	Work product shows consistent evidence of proper use of conventional ideas	Work product uses and incorporates concepts in a limited fashion to organize and explain findings	Work product shows limited evidence of integration or connection among all its elements and one or more places where lack of integration or connection is a problem	Work product falls short of complete solution strategy
Novice – 2					
Work product applies subject area rules in procedural (literal) fashion	Work product is inefficient in its treatment of the task	Work product shows some evidence of proper use of conventional ideas	Work product organizes and explains findings in a way that does not use concepts in any significant fashion	Work product shows little evidence of integration or connection among all its elements and many places where lack of integration or connection is a problem	Work product falls well short of a complete solution strategy for the problem
Emerging Novice – 1					
Work product applies wrong rules. Applies rules inefficiently, or not at all	Work product is highly inefficient, redundant or confused in its treatment of the task	Work product shows little or no evidence of proper use of conventional ideas	Work product does not use or incorporate concepts and/or does not explain findings coherently	Work product shows essentially no evidence of integration or connection among all its elements	Work product fails to show a solution strategy for the problem

Melanie: We find it hard to use this rubric in any useful way. When we tried to use this rubric in our math department and had a morning session, it took us over an hour just to grade Jabari's work, and we never actually reached a consensus. So we took a step back and tried to look at Jabari's work through the lens of the Common Core. Because this is a modeling task, we looked specifically at the Common Core modeling standard.

Modeling standards

CCSS.Math.Practice.MP4 Model with mathematics.

Mathematically proficient students can apply the mathematics they know to solve problems arising in everyday life, society, and the workplace. In early grades, this might be as simple as writing an addition equation to represent a situation. In middle and high school, a student might use geometry to design a product or make a scale drawing. In high school, a student might use algebra to model a situation and solve for an unknown. In college and beyond, students use mathematics to model situations in the physical and biological sciences, engineering, and economics. They routinely interpret and reflect on the results of their modeling.

- Making assumptions to simplify a situation
- Representing relationships between quantities with tables, graphs & equations
- Generalizing

Melanie: We looked at three different things. The first was whether he made assumptions to simplify a situation. We saw evidence of that. He decided that every time he tied a knot that the rope would shrink by the same amount. That's his drop rate. And he also simplified the situation by assigning a value to the length of the rope initially, because that wasn't given to him.

Eyal: So in terms of whether he was able to represent relationships between quantities. He was. He made a table. But we were maybe hoping that he would do some other representations also and he didn't.

Melanie: And last, he never generalized. He didn't develop other representations, an equation maybe or a graph. He never got to that point, and we are not even sure Jabari knew that he was supposed to generalize in this problem. It wasn't really clear to him. He did have a solution that worked. But, we had questions about the level of the sophistication.

Eyal: Yes, that leaves us in a funny space. The Common Core gives us wonderful language to talk about mathematical sophistication, but also leaves us wondering how to give it a score and realizing that assessing mathematical practices is messier than we first might have thought.

Melanie: We found the rubric was clunky. It wasn't useful to us as teachers or useful to Jabari or the other students in that class. So, this leaves us with some questions. What kinds of training and support are teachers going to be given, in particular for grading mathematical practices? How will expectations be made clear to teachers and students? For instance, what are we asking them to demonstrate? They aren't mind readers; they need guidance. Last, standardizing these standards will require a standardized test. Can we think outside the box, here? Can we go beyond a timed paper and pen test? Is there another way — a better way — for us to assess these hard-to-assess competences?

These teachers point compellingly at the need to understand mathematical proficiencies better so that they can more effectively be taught and learned. They also provide insight into ways in which our collective understanding of mathematical proficiencies can be furthered through thoughtful deliberation of core assessment issues — what needs to be assessed, do items elicit evidence, and why believe the results?

Fair assessments in a diverse society¹⁰

It is unconscionable to accept the status quo of an education system that undervalues diversity and produces achievement gains predictable by race, class, and language background. Certainly, broader factors in society are the source of much inequity, for example discrimination and the social construction of poverty. Schools inherit these ills. Such factors shape what students bring to school, the distribution of educational opportunity, and out-of-school supports and obstacles. However, much of the difference in learning (namely achievement gains) is directly attributable to schools and to teaching. Differences are evident, even when controlling for broader societal effects. In addition, American society is becoming increasingly diverse and its ability to prosper is tied to its ability to harness that diversity.

Sources of inequity in schooling are perpetuated by default patterns of instruction, but also by default patterns in assessment practices. Several scholars are working to identify these patterns and to propose principles to remedy them. These include, for instance, being explicit about what is expected, using multiple representations, and giving credit to correct, non-standard approaches.

A confusion that often arises when reviewing such recommendations is that principles of *equitable assessment*

Websites of groups addressing issues related to fair assessments in a diverse society:

- Understanding Language (<http://ell.stanford.edu/>)
- The Algebra Project Inc. (<http://www.algebra.org/>)
- BUENO Center for Multicultural Education (<http://buenocenter.org/>)
- World-Class Instructional Design and Assessment (<http://www.wida.us/>)
- Center for the Mathematics Education of Latinos/as (<http://math.arizona.edu/~cemela/english/>)
- Educational Testing Service (http://www.ets.org/s/achievement_gap/)



Work on one of three equitable assessment practices draws on a growing understanding of the ways diverse students use mathematical practices to advance their learning.

practice can appear to be nothing more than principles for “good” assessment, independent of a concern for fairness. Equitable assessment practice and good assessment practice have some common ground and common spirit, but they are not the same. The focus makes a difference, and confounding them runs the risk of minimizing attention to important features of fair assessments. There are several ways in which equitable assessment practices may be distinct. Some are about attending to and accommodating different groups of students, for instance by avoiding cultural contexts familiar to some groups but not to others. Some are good ideas to do regardless of fairness, but are ideas that deserve heightened attention because they matter differently for different groups and their use reduces bias, such as minimizing unnecessary language complexity of an item.

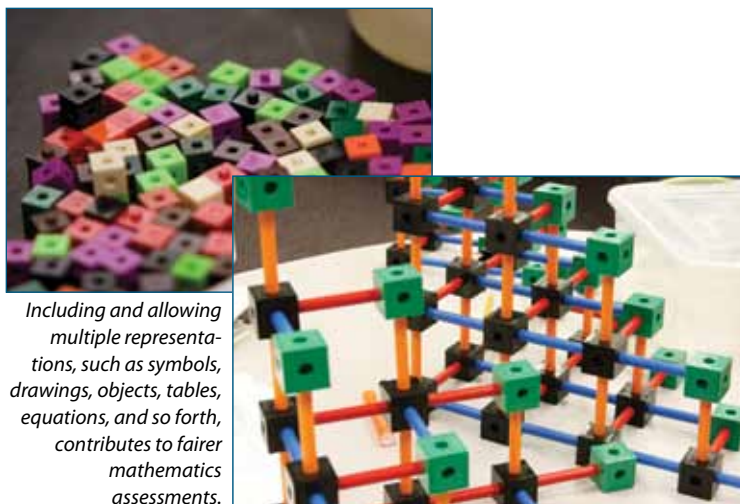
Current work on the development of more equitable assessment practice is proceeding along three related lines of work. One draws on current research on equitable approaches to teaching to consider implications for assessment. Another analyzes differences in the way items function across populations of students to identify and reduce sources of test bias. A third draws on a growing understanding of the ways diverse students use mathematical practices to advance their learning.

¹⁰ This section is adapted from presentations given by Judit Moschkovich (University of California, Santa Cruz), Maria Martiniello (independent scholar), Guillermo Solano-Flores (University of Colorado, Boulder), and Marcus Hung (Algebra Project).

Equitable teaching — equitable assessment

In part, equitable teaching is specific to the learners and the subject matter being taught. However, some general guidelines are evident. Equitable teaching makes use of prior knowledge, cultural assets, and home languages; teachers need to familiarize themselves with these aspects of students' backgrounds and identify resources that can be drawn on to support student engagement and learning. Equitable teaching maintains high expectations and provides appropriate support for meeting expectations. Two well-documented sources of inequity are, on the one hand, a self-fulfilling prophecy of low expectations, reduced demand, and collusion related to low performance and, on the other hand, high expectations coupled with no support, which sets students up for disappointment and failure. Equitable teaching establishes and maintains demanding engagement in content, provides instruction in discipline-specific practices, and is explicit about academic language and how to use the instruction being provided to support student learning. The cornerstone of equitable teaching, then, is attention and responsiveness to students and their meanings and skills, combined with thoughtful and rigorous treatment of content.

Assessment needs to be designed to be consonant with such guidelines. It needs to recognize the linguistic, cultural, and conceptual resources that students may bring to a task and accommodate students' use of these different resources. At the same time, it needs to address high, yet realistic standards. Different assessments with different formats should be used, such as short-answer tests, written text, oral presentations, drawing, using tools, and so forth. Including and allowing multiple representations, such as symbols, drawings, objects, tables, equations, and so forth, contributes to fairer mathematics assessments. In addition, attention needs to be given to the quality and usability of the information provided to students, so that they have the information they need to focus their efforts and so that they understand what others have understood (and not) about their understanding. These guidelines apply to both formative and summative assessments. Short oral presentations, for instance, may be used midway in an instructional sequence to gather information about what students have taken up and any lingering confusions. Presenting to parents, other students, or community members may also happen at the end of a unit, as a summative performance, providing a different format for demonstrating command of a set of ideas taught and learned.



Including and allowing multiple representations, such as symbols, drawings, objects, tables, equations, and so forth, contributes to fairer mathematics assessments.

Reducing test bias

A second line of work examines demonstrable differences in the ways items function for different groups of students. In particular, this work looks at items for which individuals from different socially identified groups of students with equivalent mathematics proficiency perform differently. As an example, consider the item below, which asks about combinations of inside and outside chores (Massachusetts Department of Education, as cited in Martiniello, 2008).

Every Saturday in the fall, Martin has to do 1 inside chore and 1 outside chore. The chores are listed below.

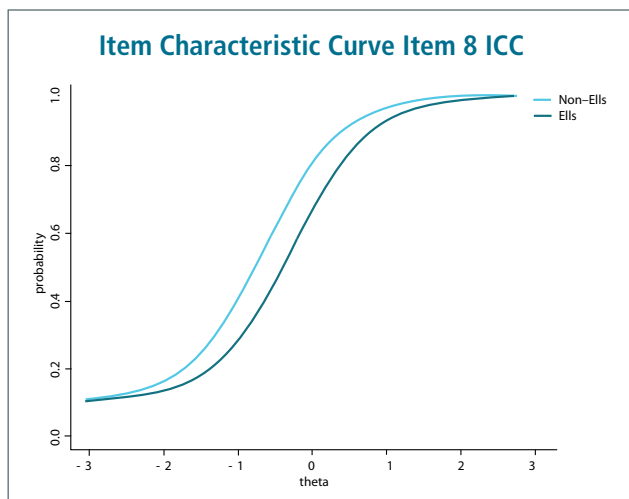
Inside Chores	Outside Chores
vacuum	rake
wash dishes	weed
dust	

How many different combinations of 1 inside chore and 1 outside chore can Martin make?

A. 3
 B. 5
 C. 6
 D. 9

Based on sample responses, a characteristic curve for a particular item shows the probability of getting the item correct for individuals of different abilities in a given population (where score on the overall test is used as a proxy for “ability”). The graph below indicates the probabilities for two populations, English learners and those who are not classified as English learners (Martiniello, 2008).

The “gap” between the two curves is called the item DIF (for differential item functioning), which for this item is relatively large. English learners of a given ability level are much less likely to get this item correct than are other students of the same overall ability level. From interviews with students, Martiniello (2008) found that English learners had difficulty with the vocabulary of this item, both the word “chore” and the chores listed. She found



that home-related vocabulary often introduced test bias in ways that school-related vocabulary did not. In other words, modifying the item to refer to combinations, for example, of math activities and reading activities would likely reduce the item’s DIF because school is a more common experience with more uniform language. It is important to note here that the issue is not simply a matter of lexical complexity as it is often judged by using word-frequency lists. In this example, word-frequency lists do not offer any insights as to the lexical challenge for English learners because all words in this item (except vacuum) are high-frequency words. The words “rake” and “weed” are high-frequency words, but the idea of them as chores (perhaps even as verbs) is likely cultural.

Understanding the source of test bias in items such as this one is an important area of research. Martiniello has identified several issues for English learners: multiple clauses, long noun phrases, unfamiliar non-mathematical words, words with multiple meanings (polysemy), home-related words, cultural references and background knowledge, item layout, and a lack of non-linguistic representations. This last one is particularly important in relation to students with diverse cultural and linguistic backgrounds. Including multiple mathematical representations in an item can reduce the reliance on text and can provide multiple entry points, affording more students the opportunity to demonstrate what they know mathematically.

Understanding mathematical practices

A third important focus for addressing equity is the development of assessments of mathematical practices. Current articulations of mathematical practices are underdeveloped, at least for the purpose of informing mathematics teaching and learning. The Common Core standards take an important step by combining disciplinary and pedagogical sensibilities, but it remains unclear what current versions of the practices mean, what would count as having taught them, and how they would be assessed. Efforts to define, refine, teach, and assess them are likely to contribute to understanding them better. In this effort, educators and scholars focused on teaching underserved students are a step ahead. They have noticed and begun working on practices that are either unacknowledged mathematical assets of underserved students or identifiable roadblocks key to student success.

For example, educators who focus on English learners have noticed that the use of visual representations can be used to advantage by English learners when their use makes the relationships between the visual representations and text or spoken language explicit, and even more so when students are taught to engage in making such relationships visible in mathematical talk of the classroom. The corollary is that excellent visual representations can be quite ineffective when the mathematical words that they imply are not routinely made explicit.

For example, researchers in Texas have observed local teachers using a “four-corner” model, presented as a “foldable” graphic organizer made from construction paper: the four corner flaps are labeled symbolic, tabular, pictorial, and verbal, and open to reveal different representations



It is particularly informative to have students think aloud as they work items, probing their assumptions and reasoning to understand how an item is functioning and to identify potential problems.

for a single mathematical concept or problem. Although this conceptual organization tool has the potential to help English learners connect the mathematical words with pictorial and symbolic expressions, the exercise often reduces to a procedural task with little meaning. The result is that students do not see how the pictorial expression relates to symbolic expression. They do not see how the symbolic expression captures and makes the verbal description easier to interpret and use. This requires the development of clear mappings of the parts and relationships of one representation to the others. Assessment items that present information and questions using multiple representations, especially ones that are relatively explicit about the connections among the representations, allow more students to demonstrate the mathematical knowledge and skill they possess.

In general, delegating (consciously or unconsciously) the teaching of mathematical practices to tacit processes of socialization systematically disadvantages students outside the dominant group and less familiar with the social cues of that group, including teachers, as well as curriculum and assessment developers. As in the examples conveyed by Eyal and Melanie above, equitable teaching pays particular attention to being explicit about what is involved in proficient performance of mathematical practices. Mathematical reasoning and justification is a particularly important mathematical practice and deserves particular attention in this regard.

In assessing mathematical proficiency, formatively or summatively, written or verbal, it is important to focus on mathematical content and students’ mathematical reasoning, not on language accuracy. An overemphasis on correct vocabulary limits seeing and hearing student competencies. It confounds the assessment of mathematical practices with non-mathematical language proficiency. This matters both in the design of scoring rubrics and in the subjective scoring common in some assessments.

Two approaches for gaining greater insight into the development of equitable assessment and for reducing test bias are worth mentioning. One is the use of interviews, as described earlier. It is particularly informative to have students think aloud as they work items, probing their assumptions and reasoning to understand how an item is functioning and identifying potential problems. Another approach that can help to improve the quality of assessments and reduce bias is to structure the involvement of people who study these issues in the development of assessments, as one of the professional communities crucial to development.

Mathematical proficiencies and future progress

In the wake of two decades of each individual state developing its own set of content standards and a political process that increasingly led to including all content at all grade levels, the Council of Chief State School Officers and the National Governors Association teamed up to support the creation of a set of standards with greater focus at each grade level, coherence across the grades, and a rigorous combination of conceptual understanding, fluency, and applied modeling.

To support the Common Core standards, two large assessment initiatives were funded: Partnership for Assessment of Readiness for College and Career (www.parcconline.org) and Smarter Balance Assessment Consortia (www.smarterbalanced.org). The charge is to produce valid, reliable, and fair assessments. Placing fairness at the same level with validity and reliability is noteworthy in the context of assessment. At least it promises that fairness be part of the conversation. In addition, tests are to be designed to document both achievement and growth. They each include a collection of assessments to be given at different

times of the year, for different purposes. They include both selected and constructed response items, performance assessments, and extensive use of technology.

The language of “progress toward college and career readiness” suggests that the tests will need to not only reliably discriminate levels of achievement, but also address benchmarks for readiness. Client states are to agree on common cut scores. This is promising because it creates a need for ongoing exchange about what mathematical proficiencies constitutes college and career readiness.

In these efforts and beyond, there will be much work to do, many opportunities for learning more about mathematical proficiencies and evidence for them, developing new and better assessments and engaging in more productive collaborations. Done smartly, these activities afford the possibility for real improvement. Each of us has something to contribute to this collective work, something in keeping with the professional expertise we have and with our appropriate role in the effort. Educating ourselves and learning to work across professional divides are important first steps.

The language of “progress toward college and career readiness” suggests that the tests will need to not only reliably discriminate levels of achievement, but also address benchmarks for readiness.



Readings and references

- Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of Educational Research*, 81(2), 132-169.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Black, P., & Wiliam, D. (2006). *Inside the black box: Raising standards through classroom assessment*. Granada Learning.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn*. Washington, DC: National Academy Press.
- Brennan R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.
- Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In Hallinan, M. T. (Ed.), *Frontiers in sociology of education* (pp. 127-162). The Netherlands: Springer.
- Bryk, A. S., Yeager, D. S., Hausman, H., Muhich, J., Dolle, J. R., Grunow, A., LeMahieu, P., & Gomez, L. (2013). Improvement research carried out through networked communities: Accelerating learning about practices that support more productive student mindsets. A White Paper prepared for the White House meeting on *Excellence in Education: The Importance of Academic Mindsets*. Retrieved from www.carnegiefoundation.org.
- Cohen, D. K. (2010). Learning to teach nothing in particular. *American Educator*, Winter, 44-54.
- Common Core Standards Writing Team. (2013). *Progressions for the common core state standards in mathematics (draft): Grades 6–8, the number system; high school, number*. Tucson, AZ: Institute for Mathematics and Education, University of Arizona. Retrieved from ime.math.arizona.edu/progressions/.
- Common Core State Standards Initiative. (2010). *Common core state standards for mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf
- Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, 23(4), 310-324.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Klein, F. (1939). *Elementary mathematics from an advanced standpoint: Geometry*. New York, NY: Dover. (Original work published 1908.)
- Mancosu, P., Jørgensen, K. F., & Pedersen, S. A. (Eds.). (2005). *Visualization, explanation and reasoning styles in mathematics* (Vol. 327). New York, NY: Springer.
- Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, 78(2), 333-368.
- Martiniello, M. (2010). Linguistic complexity in mathematics assessments and the performance of English language learners. In R. S. Kitchen, & E. Silver (Eds.), *Assessing English language learners in mathematics* (A Research Monograph of TODOS: Mathematics for All), 2(2), Washington, DC: National Education Association.
- Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education/Praeger.
- Pólya, G. (1957). *How to solve it*. Garden City, NY: Doubleday.
- Thurston, W. P. (1995). On proof and progress in mathematics. *For the Learning of Mathematics*, 15(1), 29-37.
- von Hippel, E. (2005). *Democratizing innovation*. Cambridge: The MIT Press.

The Mathematical Sciences Research Institute (MSRI), located in Berkeley, California, fosters mathematical research by bringing together the foremost mathematical scientists from around the world in an environment that promotes creative and effective collaboration. MSRI's research extends through pure mathematics into computer science, statistics, and applications to other disciplines, including engineering, physics, biology, chemistry, medicine, and finance. Primarily supported by the U.S. National Science Foundation, the Institute is an independent nonprofit corporation that enjoys academic affiliation with nearly 100 leading universities as well as support from individuals, corporations, foundations, and other government and private organizations.

MSRI's major programs, its postdoctoral training program, and its workshops draw together the strongest mathematical scientists, with approximately 2,000 visits over the course of a year. At any time, about eighty-five mathematicians are in residence for extended stays. Public outreach programs and VMath, the largest mathematical streaming video archive in the world, ensure that many others interact with MSRI throughout the year.

MSRI created the Critical Issues in Mathematics Education Workshop Series in 2004. This series of workshops addresses key problems in education today and is designed to engage mathematicians, mathematics education researchers, and K-12 teachers. The workshops provide participants a unique opportunity to learn about research and development efforts in this area. In addition participants develop ideas about methods for working on these problems and get to analyze and evaluate current or proposed programs. These workshops offer a space to make connections and exchange ideas with others concerned with the same issues in their fields.

Most workshops are held at MSRI and last for a few intensely secluded days. Each workshop attracts approximately 200 participants. Workshop organizers make sure to ensure diversity and relevant expertise by reaching out to mathematicians from a broad cross-section of colleges and universities.

For more information visit www.msri.org



Main Office 510-642-0143 • Fax 510-642-8609

Mailing Address: 17 Gauss Way • Berkeley, CA 94720-5070

www.msri.org

