

“Get Real!” Assessing for Quantitative Literacy*

by Grant Wiggins†

“OK, people, settle down. It’s time to take out some paper and pencil, we’re going to have a pop quiz today in Quant. Lit. 101. Stop the groaning please! You have 40 minutes only. As always, you can consult any resource, including other people in the room, but budget your time wisely and note all texts and people consulted for each answer. . . . Here are the questions.”

1. What is the meaning of the phrase “statistical tie” in the sentence “The result of the 2000 election in Florida was a statistical tie, even though George Bush was declared the winner”? Extra credit: Sketch out a mathematically sound and politically palatable solution to the problem of close elections.
2. Respond to the following claim, made by a student to his geometry teacher: “Well, you may have proven the theorem today, but we may discover something tomorrow that proves the theorem wrong.”
3. Guesstimate quickly, please: If you want the most money for your retirement, should you (a) invest \$500 per year in an index-based mutual fund from the time you are 16 years old to the time you are 30, or (b) invest \$1,000 per year in a bank savings account from the time you are 25 until you are 65?
4. Is mathematics more like geography (a science of what is really “out there”) or more like chess (whose rules and logical implications we just made up)? Did we “discover” the truth that $1 + 1 = 2$, or did we “invent” it? Based on our work this semester, give two plausible reasons for each perspective. Then give your own view, with reasons.
5. Study the data on the last 10 years of AIDS cases in the United States from the newspaper clipping in front of you. What are two trends for charting future policy?
6. “At current rates of revenue and payout the Social Security fund will be bankrupt by the time you retire.” Explain how this statement could be both true and false, mathematically speaking, depending on the definitions and assumptions used.
7. Comment on this proof, please:¹
Solve $6x - 10 = 21x - 35$ for x .
Solution: $2(3x - 5) = 7(3x - 5)$
Therefore $2 = 7$
8. “Hoops” McGinty wants to donate millions of dollars from his salary and sports-drink earnings toward a special exhibit in the new Rose Planetarium area of the American Museum of Natural History in New York. Hoops wants the exhibit to include a three-dimensional scale model of the solar system in which the size of the planets and the distance of each planet from the sun would be exactly to scale. There is a catch, however: the sun is to be represented by a regulation NBA basketball. The nervous folks in the gifts department of the museum call on you because of your expertise in astronomy and matters of scale. What can you advise them—quickly—about the feasibility of McGinty’s plan? What approach will work best to ensure a basketball-related design in the display?
9. Discuss the following statement, picking a key axiom as an example to support your observations: “The axioms in any mathematical system may *logically* precede the theorems, but it does not follow (and indeed is not true historically) that they were all formulated *prior in time* to the theorems. Axioms are *not* self-evident truths. They may even sometimes be less obvious than theorems, and formulated late in the game. They are necessary ‘givens’, shaped by what we wish to be able to prove.”
10. Write a memo to the House Education Committee on the accuracy and implications of the following analysis:

New York Times, August 13, 2001
Rigid Rules Will Damage School
By Thomas J. Kane and Douglas O. Staiger
As school was about to let out this summer, both houses of Congress voted for a dramatic expansion of the federal role in the education of our children. A committee is at work now to bring the two bills together, but whatever the specific result, the center of the Elementary and Secondary Education Act will be identifying schools that are not raising test scores

* Reprinted, with permission, from *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges*, Bernard L. Madison and Lynn Arthur Steen, Editors, Princeton, NJ: National Council on Education and the Disciplines, 2003, pp. 121-143.

† Grant Wiggins is the President of Grant Wiggins & Associates, an educational organization that consults with schools, districts, and state education departments on a variety of issues, notably assessment and curricular change. Wiggins is the author of *Educative Assessment (1998)*, *Assessing Student Performance (1999)*, and (with Jay McTighe) *Understanding by Design (2000)*. Wiggins' many articles have appeared in such journals as *Educational Leadership* and *Phi Delta Kappan*.

fast enough to satisfy the federal government and then penalizing or reorganizing them. Once a school has failed to clear the new federal hurdle, the local school district will be required to intervene.

The trouble with this law . . . is that both versions of this bill place far too much emphasis on year-to-year changes in test scores. . . . Because the average elementary school has only 68 children in each grade, a few bright kids one year or a group of rowdy friends the next can cause fluctuations in test performance even if a school is on the right track.

Chance fluctuations are a typical problem in tracking trends, as the federal government itself recognizes in gathering other kinds of statistics. The best way to keep them from causing misinterpretations of the overall picture is to use a large sample. The Department of Labor, for example, tracks the performance of the labor market with a phone survey of 60,000 households each month. Yet now Congress is proposing to track the performance of the typical American elementary school with a sample of students in each grade that is only a thousandth of that size.

With our colleague Jeffrey Geppert of Stanford, we studied the test scores in two states that have done well, investigating how their schools would have fared under the proposed legislation. Between 1994 and 1999, North Carolina and Texas were the envy of the educational world, achieving increases of 2 to 5 percentage points every year in the proportion of their students who were proficient in reading and math. However, the steady progress at the state level masked an uneven, zigzag pattern of improvement at the typical school. Indeed, we estimate that more than 98 percent of the schools in North Carolina and Texas would have failed to live up to the proposed federal expectation in at least one year between 1994 and 1999. At the typical school, two steps forward were often followed by one step back.

More than three-quarters of the schools in North Carolina and Texas would have been required to offer public school options to their students if either version of the new education bill had been in effect. Under the Senate bill a quarter of the schools in both states would have been required to restructure themselves sometime in those five years—by laying off most of their staffs, becoming public charter schools or turning themselves over to private operators. Under the more stringent House bill, roughly three-quarters of the schools would have been required to restructure themselves.

Both bills would be particularly harsh on racially diverse schools. Each school would be expected to achieve not only an increase in test scores for the school as a whole, but increases for each and every racial or ethnic group as well. Because each group's scores fluctuate depending upon the particular students being tested each year, it is rare to see every group's performance moving upward in the same year. Black and Latino students are more likely than white students to be enrolled in highly diverse schools, so their schools would be more likely than others to be arbitrarily disrupted by a poorly designed formula. . . .

In their current bills, the House and Senate have set a very high bar—so high that it is likely that virtually all school systems would be found to be inadequate, with many schools failing. And if that happens, the worst schools would be lost in the crowd. The resources and energy required to reform them would probably be dissipated. For these schools, a poorly designed federal rule can be worse than no rule at all.²

11. "It is fair to say that no more cataclysmic event has ever taken place in the history of thought." Even though we have not read the text from which this quote comes, mathematician Morris Kline was referring to a mid-nineteenth-century development in mathematics. To what was he *most likely* making such dramatic reference? Why was it so important in the history of thought?

* * *

In an essay designed to stimulate thought and discussion on assessing quantitative literacy (QL), why not start with a little concrete provocation: an attempt to suggest the content of questions such an assessment should contain? (Later I will suggest why the typical form of mathematics assessment—a "secure" quiz/test/examination—can produce invalid inferences about students' QL ability, an argument that undercuts the overall value of my quiz, too.)

Note that the questions on my quiz relate to the various proposed definitions of QL offered in *Mathematics and Democracy: The Case for Quantitative Literacy* (hereafter "case statement").³ As part of a working definition, the case statement identified 10 overlapping elements of quantitative literacy:

- A. Confidence with Mathematics
- B. Cultural Appreciation
- C. Interpreting Data
- D. Logical Thinking
- E. Making Decisions
- F. Mathematics in Context
- G. Number Sense
- H. Practical Skills
- I. Prerequisite Knowledge
- J. Symbol Sense

to which I would peg my quiz questions categorically as follows:

- | | | |
|-----|---------------------|------------------|
| 1. | Statistical Tie | C, E, F, H |
| 2. | Fragile Proof | A, D, I |
| 3. | Investment Estimate | E, F, G, H |
| 4. | Discover or Invent | A, B, D, I |
| 5. | AIDS Data | C, F, G, I |
| 6. | Social Security | A, B, D, E, G, H |
| 7. | Silly Proof | D, I |
| 8. | Solar System | C, E, F, G, H |
| 9. | Axioms and Truth | D, I, J |
| 10. | Testing Memo | C, D, E, F, H |
| 11. | Cataclysmic | B |

If we wish for the sake of mental ease to reduce the 10 overlapping elements of quantitative literacy to a few phrases, I would propose two: *realistic mathematics in*

context and mathematics in perspective. Both of these can be summed up by a familiar phrase: quantitative literacy is about mathematical understanding, not merely technical proficiency. Certainly, the call for a more realistic approach to mathematics via the study of numbers in context is at the heart of the case for QL. The importance of context is underscored repeatedly in *Mathematics and Democracy*,⁴ and not only in the case statement:

In contrast to mathematics, statistics, and most other school subjects, quantitative literacy is inseparable from its context. In this respect it is more like writing than like algebra, more like speaking than like history. Numeracy has no special content of its own, but inherits its content from its context.⁵

. . . mathematics focuses on climbing the ladder of abstraction, while quantitative literacy clings to context. Mathematics asks students to rise above context, while quantitative literacy asks students to stay in context. Mathematics is about general principles that can be applied in a range of contexts; quantitative literacy is about seeing every context through a quantitative lens.⁶

But what exactly is implied here for assessment, despite the surface appeal of the contrast? To assess QL, we need to make the idea of “context” (and “realistic”) concrete and functional. What exactly *is* a context? In what sense does mathematics “rise above context” while QL asks students to “stay in context”? Does context refer to the content area in which we do QL (as suggested by one of the essays in *Mathematics and Democracy*⁷) or does context refer to the conditions under which we are expected to use mathematical abilities in any content area? If QL is “more like writing,” should we conclude that current writing assessments serve as good models for contextualized assessment? Or might not the *opposite* be the case: the contextual nature of writing is regularly undercut by the canned, bland, and secure one-shot writing prompts used in all large-scale tests of writing? If context is by definition unique, can we *ever* have standardized tests “in context”? In other words, is “assessing performance in context” a contradiction in terms?

What about assessing for mathematics in perspective, our other capsule summary of QL? As quiz questions 2, 4, 9, and 11 suggest, such an assessment represents a decidedly unorthodox approach to teaching and assessment for grades 10 to 14. Some readers of this essay no doubt reacted to those questions by thinking, “Gee, aren’t those only appropriate for graduate students?” But such a reaction may only reveal how far we are from understanding how to teach and assess for understanding. We certainly do not flinch from asking high school students to read and derive important meaning from Shakespeare’s *Macbeth*, even though our adult hunch might be that students lack the

psychological and literary wisdom to “truly” understand what they read. Reflection and meaning making are central to the learning process, even if it takes years to produce significant results. Why should mathematics assessment be any different?

In fact, I have often explored questions 4 and 9 on the nature of “givens” and proof with high school mathematics classes, with great results, through such questions as: Which came first: a game or its rules? Can you change the rules and still have it be the same game? Which geometry best describes the space you experience in school and the space on the surface of the earth? Then why is Euclid’s the one we study? In one tenth-grade class, a student with the worst grades (as I later found out from the surprised teacher) eagerly volunteered to do research on the history of rule changes in his favorite sports, to serve as fodder for the next class discussion on “core” versus changeable rules. (That discussion, coincidentally, led to inquiry into the phrase “spirit versus letter of the law”—a vital idea in United States history—based on the use of that phrase in a ruling made by the president of baseball’s American League in the famous George Brett pine-tar bat incident 20 years ago.)

I confess that making mathematics more deliberately meaningful, and then assessing students’ meaning making (as we do in any humanities class), is important to me. Although some readers sympathetic to the case statement may disagree, they only need sit in mathematics classrooms for a while (as I have done over the past 20 years) to see that too many teachers of mathematics fail to offer students a clear view of what mathematics *is* and why it matters intellectually. Is it any accident that student performance on tests is so poor and that so few people take upper-level mathematics courses?

Without anchoring mathematics on a foundation of fascinating issues and “big ideas,” there is no intellectual rationale or clear goal for the student. This problem is embodied in the role of the textbook. Instead of being a resource in the service of broader and defensible priorities, in mathematics classes the textbook *is* the course. I encourage readers to try this simple assessment of the diagnosis: ask any mathematics student midyear, “So, what are the few really big ideas in this course? What are the key questions? Given the mathematics you are currently learning, what does it enable you to do or do better that you could not do without it?” The answers will not yield mathematics teachers much joy. By teaching that mathematics is mere unending symbol manipulation, all we do is induce innumeracy.

Quiz question 11 interests me the most in this regard because, whether or not I agree with Kline, I would be willing to bet that not more than one in 100 highly educated people know anything about the development in question—

even if I were to give the hint of “Bolyai and Lobachevski.” More important, most would be completely taken aback by Kline’s language: how can any development in mathematics be intellectually cataclysmic? (I can say without exaggeration that I was utterly roused to a life of serious intellectual work by becoming immersed in the controversies and discoveries Kline refers to. I had no idea that mathematics could be so controversial, so thought provoking, so important.)

Regardless of my idiosyncratic St. John’s College experience, should not all students consider the meaning of the skills they learn? That is what a liberal education is all about: So what? What of it? Why does it matter? What is its value? What is assumed? What are the limits of this “truth”? These are questions that a student must regularly ask. In this respect, quantitative literacy is no different from reading literacy: assessment must seek more than just decoding ability. We need evidence of fluent, thoughtful meaning making, as Peter T. Ewell noted in his interview in *Mathematics and Democracy*.⁸

Talking about quantitative literacy as part of liberal education may make the problem seem quaint or “academic” in the pejorative sense. The QL case statement is in fact *radical*, in the colloquial and mathematical sense of that term. As these opening musings suggest, we need to question the time-honored testing (and teaching) practices currently used in *all* mathematics classes. We are forced to return to our very roots—about teaching, about testing, about what mathematics is and why we teach it to nonspecialists—if the manifesto on quantitative literacy is to be realized, not merely praised.

The result of students’ endless exposure to typical tests is a profound lack of understanding about what mathematics is: “Perhaps the greatest difficulty in the whole area of mathematics concerns students’ misapprehension of what is actually at stake when they are posed a problem. . . . [S]tudents are nearly always searching for [how] to follow the algorithm. . . . Seeing mathematics as a way of understanding the world . . . is a rare occurrence.”⁹ Surely this has more to do with enculturation via the demands of school, than with some innate limitation.¹⁰

Putting it this way at the outset properly alerts readers to a grim truth: this reform is *not* going to be easy. QL is a Trojan horse, promising great gifts to educators but in fact threatening all mainstream testing and grading practices in all the disciplines, but especially mathematics. The implications of contextualized and meaningful assessment in QL challenge the very conception of “test” as we understand and employ that term. Test “items” posed under standardized conditions are decontextualized by design.

These issues create a big caveat for those cheery reformers who may be thinking that the solution to quantitative

illiteracy is simply to add more performance-based assessments to our repertoire of test items. The need is not for performance tests (also out of context)—most teacher, state, and commercial tests have added some—but for an altogether different approach to assessment. Specifically, assessment must be designed to cause questioning (not just “plug and chug” responses to arid prompts); to teach (and not just test) which ideas and performances really matter; and to demonstrate what it means to *do* mathematics. The case statement challenges us to finally solve the problem highlighted by John Dewey and the progressives (as Cuban notes¹¹), namely, to make school no longer isolated from the world. Rather, as the case statement makes clear, we want to regularly assess student work with numbers and numerical ideas in the field (or in virtual realities with great verisimilitude).

What does such a goal imply? On the surface, the answer is obvious: we need to see evidence of learners’ abilities to use mathematics in a distinctive and complicated situation. In other words, the challenge is to assess students’ abilities to bring to bear a repertoire of ideas and skills to a specific situation, applied with good judgment and high standards. In QL, we are after something akin to the “test” faced by youthful soccer players in fluid games after they have learned some discrete moves via drills, or the “test” of the architect trying to make a design idea fit the constraints of property, location, budget, client style, and zoning laws.

Few of us can imagine such a system fully blown, never mind construct one. Our habits and our isolation—from one another, from peer review, from review by the wider world—keep mathematics assessment stuck in its ways. As with any habit, the results of design mimic the tests we experienced as students. The solution, then, depends on a team design approach, working against clear and obligatory design standards. In other words, to avoid reinventing only what we know, assessment design needs to become more public and subject to disinterested review—in a word, more professional.

This is in fact the chief recommendation for improving mathematics teaching in *The Teaching Gap*, based on a process used widely in Japanese middle schools.¹² I can report that although such an aim may at first seem threatening to academic prerogative, for the past 10 years we have trained many dozens of high school and college faculties to engage in this kind of group design and peer review against design standards, without rancor or remorse. (Academic freedom does not provide cover for assessment malpractice: a test and the grading of it are not valid simply because a teacher says that they are.)

Thus the sweeping reform needed to make QL a reality in school curriculum and assessment is as much about the reinvention of the job description of “teacher” and the norms of the educational workplace as it is about

developing new tests. To honor the case statement is to end the policies and practices that make schooling more like a secretive and austere medieval guild than a profession.¹³ The change would be welcome; I sketch some possibilities below.

What We Assess Depends on Why We Assess

Any discussion of assessment must begin with the question of purpose and audience: for what—and whose—purposes are we assessing? What are the standards and end results sought and by whom? What exactly do we seek evidence of and what should that evidence enable us and the learners to do?

These are not simple or inconsequential questions. As I have argued elsewhere, in education we have often sacrificed the primary client (the learner) in the name of accountability.¹⁴ Students' needs too often have been sacrificed to teachers' need for ease of grading; teachers' needs as coach too often have been sacrificed to the cost and logistical constraints imposed by audits testing for accountability or admissions. Rather than being viewed as a key element in ongoing feedback cycles of learning to perform, testing is viewed as something that takes place *after* each bit of teaching is over to see who got it and who did not, done in the most efficient manner possible, before we move on in the linear syllabus, regardless of results.

If there is an axiom at the heart of this argument it is this: assessment should be first and foremost for the learner's sake, designed and implemented to provide useful feedback to the learner (and teacher-coach) on worthy tasks to make improved performance and ultimate mastery more likely.¹⁵ This clearly implies that the assessment must be built on a foundation of realistic tasks, not proxies, and built to be a robust, timely, open, and user-friendly system of feedback and its use. Assessments for other purposes, (e.g., to provide efficiently gained scores for ranking decisions, using secure proxies for real performance) would thus have to be perpetually scrutinized to be sure that a secondary purpose does not override the learner's right to more educative assessment.

We understand this in the wider world. Mathematicians working for the U.S. Census Bureau are paid to work on situated problems on which their performance appraisals depend. We do not keep testing their mathematical virtuosity, using secure items, to determine whether they get a raise based merely on what they know. Athletes play many games, under many different conditions, both to test their learning and as an integral part of learning. I perform in concert once a month with my "retro" rock band the *Hazbins* to keep learning how to perform (and to feel the joy from doing so); a score from a judge on the fruits of my guitar lessons, in isolated exercises, would have little value

for me. The formal challenge is not an onerous extra exercise but the *raison d'être* of the enterprise, providing educational focus and incentive.

Yet, most tests fail to meet this basic criterion, designed as they are for the convenience of scorekeepers not players. Consider:

- The test is typically unknown until the day of the assessment.
- We do not know how we are doing as we perform.
- Feedback after the performance is neither timely nor user friendly. We wait days, sometimes weeks, to find out how we did; and the results are often presented in terms that do not make sense to the performer or sometimes even to the teacher-coach.
- The test is usually a proxy for genuine performance, justifiable and sensible only to psychometricians.
- The test is designed to be scored quickly, with reliability, whether or not the task has intellectual value or meaning for the performer.

In mathematics, the facts are arguably far worse than this dreary general picture suggests. Few tests given today in mathematics classrooms (be they teacher, state, or test-company designed) provide students with performance goals that might provide the incentive to learn or meaning for the discrete facts and skills learned. Typical tests finesse the whole issue of purpose by relying on items that ask for discrete facts or technical skill out of context. What QL requires (and any truly defensible mathematics program should require), however, is assessment of complex, realistic, meaningful, and creative performance.

Whether or not my particular opening quiz questions appeal to you, I hope the point of them is clear: Evidence of "realistic use," crucial to QL, requires that students confront challenges like those faced in assessment of reading literacy: Hmm, what does this mean? What kind of problem is this? What kind of response is wanted (and how might my answer be problematic)? What is assumed here, and is it a wise assumption? What feedback do I need to seek if I am to know whether I am on the right track?¹⁶ Assessment of QL requires tasks that challenge the learner's judgment, not just exercises that cue the learner.

The same holds true for assessing students' understanding of mathematics in perspective. Students may be able to prove that there are 180 degrees in any triangle, but it does not follow that they understand what they have done. Can they explain why the proof works? Can they explain why it matters? Can they argue the crucial role played by the parallel postulate in making the theorem possible, the 2000-year controversy about that postulate (and the attempts by many mathematicians to prove or alter it), and the eventual realization growing from that controversy that there could

be other geometries, as valid as Euclid's, in which the 180-degree theorem does *not* hold true?

As it stands now, almost all students graduate from college never knowing of this history, of the existence of other valid geometries, and of the intellectual implications. In other words, they lack perspective on the Euclidean geometry that they have learned. When they do not really grasp what an axiom is and why we have it, and how other systems might and do exist, can they really be said to understand geometry at all?

What is at stake here is a challenge to a long-standing habit conveyed by a system that is not based on well-thought through purposes. This custom was perhaps best summarized by Lauren Resnick and David Resnick over 15 years ago: "American students are the most tested but the least examined students in the world."²⁷ As the case statement and the Resnick's remark suggest, what we need is to probe more than quiz, to ask for creative solutions, not merely correct answers.¹⁸

What Is Realistic Assessment and Why Is It Needed?

Regardless of the nettlesome questions raised by the call for improved quantitative literacy, one implication for assessment is clear enough: QL demands evidence of students' abilities to grapple with realistic or "situated" problems. But what is unrealistic about most mathematics tests if they have content validity and tap into skills and facts actually needed in mathematics? The short answer is that typical tests are mere proxies for real performance. They amount to sideline drills as opposed to playing the game on the field.

The aims in the case statement are not new ones. Consider this enthusiastic report about a modest attempt to change college admissions testing at Harvard a few years back. Students were asked to perform a set of key physics experiments by themselves and have their high school physics teacher certify the results, while also doing some laboratory work in front of the college's professors:

The change in the physics requirement has been more radical than that in any other subject. . . . For years the college required only such a memory knowledge of physical laws and phenomena as could be got from a . . . textbook. . . . [U]nder the best of circumstances the pupil's thinking was largely done for him. By this method of teaching . . . his memory was loaded with facts of which he might or might not have any real understanding, while he did very little real thinking. . . . This was a system of teaching hardly calculated to train his mind, or to awaken an interest in [physics].

How different is the present attitude of the college! It now publishes a descriptive list of forty experiments, covering the elementary principles of mechanics, sound, light, heat, and electricity. These, so far as possible, are quantitative experiments; that is, they require careful measurements from which the laws and principles of physics can be reasoned out. Where, for any reason, such measurements are impossible, the experiments are merely illustrative; but even from these the student must reason carefully to arrive at the principles which they illustrate. The student must perform these experiments himself in a laboratory, under the supervision of a teacher. He must keep a record of all his observations and measurements, together with the conclusions which he draws from them. The laboratory book in which this record is kept, bearing the certificate of his instructor, must be presented for critical examination when he comes to [the admissions office]. In addition to this, he is tested by a written paper and by a laboratory examination.¹⁹

This account was written about Harvard in the *Atlantic Monthly*—in 1892! We know what happened later, of course. The College Board was invented to make admissions testing more streamlined and standardized (and thereby, it must be said, more equitable for students around the country, as well as less of a hassle for colleges), but at another cost, as it turns out.

Although the century-old physics test may not have been situated in a real-world challenge, it was a noble attempt to see if students could actually *do* science. This is surely where assessment for QL must begin: Can the student do mathematics? Can the student confront inherently messy and situated problems well? That is a different question from "does the student know various mathematical 'moves' and facts?"

Some folks have regretted or resented my long-time use of the word "authentic" in describing the assessments we need.²⁰ But the phrase remains apt, I think, if readers recall that one meaning of authentic is "realistic." Conventional mathematics test questions are not authentic because they do not represent the challenges mathematicians face routinely in their work. As noted above, a mathematics test is more like a series of sideline drills than the challenge of playing the game. In fact, mathematics tests are notoriously unrealistic, the source of unending jokes by laypersons about trains heading toward each other on the same track, and the source of the wider world's alienation from mathematics. (Research is needed, I think, to determine whether simplistic test items are so abstracted from the world as to be needlessly hard for all but the symbolically inclined.)

How should we define "realistic"?²¹ An assessment task, problem, or project is realistic if it

is faithful to how mathematics is actually practiced when real people are challenged by problems involving numeracy. The task(s) must reflect the ways in which a person's knowledge and abilities are tested in real-world situations. Such challenges

- *ask us to “do” the subject.* Students have to use knowledge and skills wisely and effectively to solve unstructured problems, not simply grind out an algorithm, formula, or number.
- *require judgment and innovation.* Instead of merely reciting, restating, or replicating through demonstration the lessons taught and skills learned, students have to explore projects in mathematics, using their repertoire of knowledge and skills.
- *reflect the contexts in which adults are tested in the workplace, in civic life, and in personal life.* Contexts involve specific situations that have particular constraints, purposes, and audiences.
- *allow appropriate opportunities to rehearse, practice, consult resources, solicit feedback, refine performances, and revise products.* Secrecy, enforced quiet, solitary work, and other artificial constraints imposed by large-scale testing are minimized.

Nothing new here. Benjamin Bloom and his colleagues made the same point almost 50 years ago, in their account of application and synthesis:

[S]ituations new to the student or situations containing new elements as compared to the situation in which the abstraction was learned . . . Ideally we are seeking a problem which will test the extent to which an individual has learned to apply the abstraction in a practical way.²² . . . [A] type of divergent thinking [where] it is unlikely that the right solution to a problem can be set in advance.²³

In later materials, Bloom and his colleagues characterized synthesis tasks in language that makes clearer what we must do to make the assessment more realistic:

The problem, task, or situation involving synthesis should be new or in some way different from those used in instruction. The students . . . may have considerable freedom in redefining it. . . . The student may attack the problem with a variety of references or other available materials as they are needed. Thus synthesis problems may be open-book examinations, in which the student may use notes, the library, and other resources as appropriate. Ideally synthesis problems should be as close as possible to the situation in which a scholar (or artist, engineer, and so forth) attacks a problem he or she is interested in. The time allowed, conditions of work, and other stipulations, should be as far from the typical, controlled examination situation as possible.²⁴

Researcher Fred Newmann and his colleagues at the University of Wisconsin have developed a similar set of standards for judging the authenticity of tasks in assessments and instructional work and have used those standards to study instructional and assessment practices around the country.²⁵ In their view, authentic tasks require:

Construction of Knowledge

1. Student organization of information (higher-order skills)
2. Student consideration of alternatives

Disciplined Inquiry

3. Core disciplinary content knowledge required
4. Core disciplinary processes required
5. Elaborated written communications required to expand understanding

Value Beyond School

6. Problems are connected to the world beyond the classroom
7. An audience beyond the school is involved

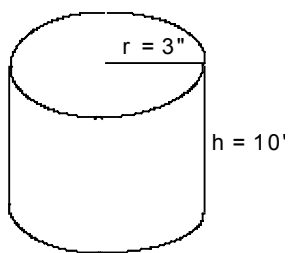
What do such tasks look like? Compare Figures 1 and 2 below. Figure 1 shows various test questions on an eighth-grade state mathematics test. Six test “items” (the four below and two others) make up the entire set of questions used to assess against the state standard for the students’ knowledge of volume. Figure 2 shows an example of a situated performance that requires students to use their understanding of that same knowledge effectively. (The second test does not replace the first test; it supplements it.)

Let us cast this contrast in terms of validity of inference. We are being asked to consider: what can we infer from good performance on the four test items? I would certainly grant the conventional premise that a student who gets most of these questions right is more likely to have control over the discrete skills and facts of this sub-domain than a student who gets most of them incorrect.

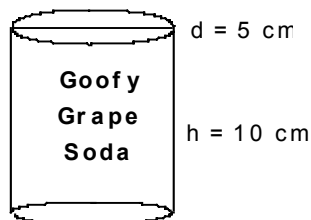
What about evidence of QL? Can we infer the likelihood that a student who got all the state test questions correct will likely do well on the second contextualized problem? Not a very plausible inference, I would claim, backed in part by data from a pilot mathematics portfolio project we ran for the Commission on Standards and Accountability in 15 districts that took the state test that had these test items.²⁶ The scores on the task in Figure 2 were low across the board—averaging 2 on a rubric scale of 6, with little range in scores within and across quite varied districts. This is not what we would expect, and it underscores the validity problems lurking in an exclusive reliance on conventional test items.

Figure 1: State Test Items, Eighth-Grade Geometry:

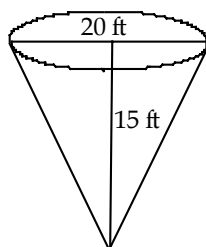
34. What is the surface area of the cylinder shown below?



- A. 13π square inches C. 60π square inches
 B. 18π square inches D. 78π square inches
35. A can of Goofy Grape Soda has a diameter of 5 cm and a height of 10 cm. What is the volume of the can of soda?



- A. 78.50 cm^3 C. 196.25 cm^3
 B. 157.00 cm^3 D. 392.50 cm^3
36. What is the volume of the cone?



- A. 4,710 cu ft
 B. 1,570 cu ft
 C. 942 cu ft
 D. 300 cu ft
37. The area of the base of a triangular pyramid was doubled. How does the new volume compare with the old volume?
- A. one-fourth
 B. one-half
 C. two times
 D. four times

Figure 2: A More Realistic Performance Task

That's a Wrap. You are in charge of the gift wrapping of purchases in a large department store. On average, 24,000 customers make clothing purchases in your store each year. About 15 percent of the customers want their purchases gift wrapped. In a month, the store typically sells 65 jackets, 250 shirts, 480 pairs of pants, and 160 hats. All boxes cost the same price per square foot of flat cardboard—\$1, and wrapping paper costs 26 cents per yard. Each roll of gift wrap is one yard wide and 100 yards long.

As the manager of gift wrap, you naturally want to plan for the year's gift wrapping costs and you want to save money where possible. What would be the best box shape for pants, shirts, jackets, and hats that requires the least amount of box and gift wrap?

Your task: Recommend to the purchasing agent in a written report, with graphs and models:

- What size boxes should be ordered for pants, shirts, jackets, and hats when ordered separately;
- The number of rolls of wrapping paper; and
- The approximate cost of wrapping paper for a year's worth of sales of pants, shirts, jackets, and hats.

Points to consider:

1. When the clothes are folded, how big does the box need to be? Of course, the way you fold makes a difference as to what shape box you could use without messing up the clothes.
2. Experiment with measuring, folding, and boxing clothes in the typical light-cardboard boxes clothes come in (or make boxes out of large pieces of paper for the experiment).
3. Some package shapes are easier than others to wrap, with minimal waste. Yet, some easy-to-wrap shapes require more paper to wrap, although less is wasted. Are there any rules or generalizations you can come up with about the amount of paper a box shape ideally requires versus the waste of paper that might be eliminated if a different-shaped box were used? Or are the savings in using the easier-to-wrap box offset by the increased costs in wrapping the new shape?
4. No one can wrap a package without wasting some paper. Figure in the cost of the extra paper and the unused or wasted paper on the roll required, given the needs of real-world wrappers in your sales force.

Your work will be judged against the following criteria:

- Effectiveness in meeting the challenge set
- The appropriateness of the mathematical and logical reasoning used
- The clarity of your communication
- The accuracy of your work

Four scoring guides for each of the criteria are used to judge your work.

The second approach permits us to see evidence of the student's thoughtful use of knowledge and skill. It does *not* obviate the need for the traditional items. But it is simply untrue, as many people defending the status quo claim, that in light of the first test, the second is unnecessary—and especially not worth the hassle and expense.

Realism and Context

Ultimately, realism hinges on situational fidelity—context—not merely whether the task is open-ended or “hands on.” QL assessment asks: Can you draw on a rich repertoire to address this complicated problem, mindful of the particular—perhaps unique—features of context? The packaging example in Figure 2 (and QL aims more generally) seek evidence of students' understanding in situations, not their technical skills in isolation. “Do you understand what to do

here, in this new situation, and why that approach might work?” “Do you grasp the significance of both the problem and the answer?” “Can you generalize from your experience, and how might that generalization be flawed or too sweeping?” These are the kinds of questions we must ask in the assessment of QL. It is, after all, what we mean by “transfer,” the Holy Grail in education.

Our failure to attend to contextualization in mathematics education can lead to humorous results. Consider the infamous National Assessment of Educational Progress (NAEP) bus problem: “An army bus holds 36 soldiers. If 1128 soldiers are being bused to their training site, how many buses are needed? Only 24 percent of eighth graders could answer it correctly. Alas, about the same percentage of the respondents answered “31 remainder 12.” No story better illustrates the habits of teaching, testing, and learning whereby mathematics floats in a netherworld of unmoored abstractions.

Context matters, so it must come to matter in day in and day out assessment. Let us have unending problems that force students to ponder its impact. Consider, for example, how the particulars of the situation affect the use of mathematics in the following problem:

Manufacturers want to spend as little as possible, not only on the product but also on packing and shipping it to stores. They want to *minimize* the cost of production of their packaging, and they want to *maximize* the amount of what is packaged inside (to keep handling and postage costs down: the more individual packages you ship the more it costs).

Suppose you are an engineer for M&M's. The manager of the shipping department has found the perfect material for shipping (a piece of poster board you will

be given). She is asking each engineering work group to help solve a problem: *What completely closed container, built out of the given materials, will hold the largest volume of M&M's for safe and economical shipping?*

You will need to prove to company executives that the shape and dimensions of your group's container idea maximize the volume. You will need to turn in a convincing written report to the managers, making your case and supplying all important data and formulas. Build multiple models out of the material to illustrate your solution. The models are not proof; they will illustrate the claims you will offer in your report. Your group will also be asked to make a three-minute oral report at the next staff meeting. The reports will be judged for accuracy, thoroughness, and persuasiveness.

Merely providing the correct mathematical answers is beside the point here. We might say without stretching the truth too much that the seemingly correct mathematics answer (the sphere) is the wrong answer here. In fact, I have seen seventh graders provide better answers to this problem than calculus students, with more insight yet limited tools.

Realism thus is not merely about access to performance-based test questions. Realism refers more to verisimilitude of content-process-situation-task-goal constraints. Most performance-based test questions are typically unrealistic because they strip context to a bare minimum, in the service of isolating the skills to be tested. To focus on context is to do justice to the fluid, ambiguous, and ill-structured situations in which typical adult performance invariably occurs, in which the answer is often “Well, it depends . . . or “Well, if . . . then . . . else . . . then . . .”

The humor in the NAEP bus problem should not blind us to the harm of failing to assess in contexts. Decontextualized training and assessment lead to unfortunate, even fatal results. Consider this complaint, made a few years back in a federal report criticizing the testing program of a national organization: “These programs are lacking in ‘real world’ scenarios and result in non-thinking performance, where the ability of the student to demonstrate a mastery of complex problems, good judgment, situational awareness, . . . and leadership skills have all been removed.” The sobering fact is that this is not an accrediting report about a school's program but a Federal Aviation Administration (FAA) report concerning deficiencies in the annual pilot testing and rectification program for a major U.S. airline. It is even more sobering to realize that the FAA is criticizing the airline for its use of airplane simulators in annual re-certification testing—a challenge more realistic than almost all school testing in mathematics today.²⁷

How then should we proceed with our design work?

Context is usefully addressed in assessment by reference to the mantra at the heart of process writing: worry about specific purpose and audience. Realistic challenges always have real (or virtual) purposes and real (or virtual) audiences. We can add other criteria and set out questions that can be used as design standards for the building of context in QL assessment:

- Is there an overriding performance goal to guide action that is obvious to the student?
- Is there a distinct audience for the work whose needs and feedback can appropriately focus the work and adjustments en route?
- Are the options and constraints in the task realistic or arbitrary?
- Are appropriate resources available? Does the task require an efficient as well as effective use of notes, materials, and repertoire of skills and concepts?
- Is secrecy concerning performance goals, possible strategy, criteria, and standards minimized?
- Is the setting realistically noisy and messy—sufficiently ill-structured and ill-defined that the learner must constantly consider what the question really is and what the key variables are?
- Are there apt opportunities to self-assess, to get lots of feedback, and to self-adjust en route as needed?

In *The Understanding by Design Handbook*, we summarize these design questions by the acronym GRASPS:

- What is the performer's *goal* in this scenario? What must he or she accomplish?
- What *role* does the performer play in this situation?
- Who is the primary *audience* for the performer's work?
- What is the *situation*? What conditions/opportunities/constraints exist?
- What are the particular *performances/products* that must be produced?
- Against what *standards* and criteria will the work be judged?²⁸

We also developed a six-faceted view of how understanding (in context) manifests itself.²⁹ For example, when we truly understand, we

- *Can explain, make connections, offer good theories:* We can make sense of what we experience. We can “show our work” and defend it. We can provide thorough, supported, and justifiable accounts of phenomena, facts, and data. We can answer such questions as: Why is that so? What explains such events? What accounts for such an effect? How can we prove it? To what is this connected? How does this work? What is implied? Why do you think so?

- *Can interpret:* Tell meaningful stories; offer apt translations; provide a revealing historical or personal dimension to ideas and events; make it personal or accessible through images, anecdotes, analogies, models. We can answer such questions as: What does it mean? Why does it matter? What of it? What does it illustrate or illuminate in human experience? How does it relate to me? What does and does not make sense here?
- *Can apply:* Effectively use and adapt what we know in diverse contexts. We can answer such questions as: How and where can we use this knowledge, skill, process? In what ways do people apply this understanding in the world beyond the school? How should my thinking and action be modified to meet the demands of this particular situation?
- *Have perspective:* See multiple points of view, with critical eyes and ears; see the big picture. We can answer such questions as: From whose point of view? From which vantage point? What is assumed or tacit that needs to be made explicit and considered? How is this justified or warranted? Is there adequate evidence? Is it reasonable? What are the strengths and weaknesses of the idea? Is it plausible? What are its limits?
- *Can empathize:* Get inside, find value in what others might find odd, alien, or implausible; perceive sensitively, enter the mind and heart of others. We can answer such questions as: How does it seem to you? What do they see that I do not? What do I need to experience if I am to understand? What was the artist, writer, or performer feeling, seeing, and trying to make me feel and see?
- *Show self-knowledge:* Perceive the personal style, prejudices, projections, and habits of mind that both shape and impede our own understanding; we are aware of what we do not understand, why it is so hard to understand. What are my blind spots? What am I prone to misunderstand because of prejudice, habit, style? How does who I am influence how I understand and do not understand?

As this summary suggests, the idea of “context” is inseparable from what it means to understand. Three of the six facets described above explicitly warn us to attend to context: application, perspective, and empathy. Other colloquial language also makes this clear. We want students to develop *tact* in the older sense of that term as used by William James in his *Talks to Teachers*: “sensitivity to the demands of the particular situation”³⁰ Thus, I would argue, understanding is a usefully ambiguous term: it properly asks us to consider both the intellectual content and the interpersonal wisdom needed here, now, in this case.

Thus the goal in assessing QL is not merely to determine whether students can use mathematical knowledge in problems but whether they can communicate with others

about that understanding and be held accountable for the consequences of the use of their knowledge. Our failure to assess using real-world consequences of the work itself and to assess students' defense of their choices and results (as opposed to giving an answer and waiting for scores returned by teachers or test companies) may explain a ubiquitous phenomenon that greatly angers the general public: students' failure to take an adequate interest in work quality and results.

Here is a vivid example of the problem. A graphics design teacher at a vocational high school brought in a real potential client for design jobs, and the students were asked to bid on the work by providing mock-ups and price quotes. They listened to the man explain his product needs, they went to work (without asking questions), and they worked very hard, much harder than usual, to produce what they thought he wanted. He came back the next week, inspected the work, and politely turned down all the efforts. The students' reaction? Anger. "We worked so hard!..." Yes, but did they ever check with the client to make sure they were on the right track? Did they put a variety of design styles before the client to tease out his tastes, as all good designers do? No. The teacher had taught these students technical skills but not how to accomplish results in the marketplace. This illustrates the need to take seriously all six facets of contextual understanding. To find evidence of students' understanding, we need problems that *require* such understanding.

Consider the following transformations of a conventional high school unit, based on a well-known textbook, to see how assessment of course content can be approached more contextually without sacrificing rigor:

The original unit, summarized:

Topic: Surface Area and Volume (geometry)

Knowledge and skill sought:

- How to calculate surface area and volume for various three-dimensional figures
- Know and use Cavalieri's Principle to compare volumes
- Know and use other volume and surface area formulas to compare shapes

Assessments, all derived from the University of Chicago School Mathematics Project geometry textbook.

- Odd-numbered problems in Chapter 10 Review, pp. 516–519
- Progress self-test, p. 515
- Homework: each third question in the sub-chapter reviews
- Completion of one "exploration"
- Exploration 22, p. 482—"Containers holding small amounts can be made to appear to hold more than they do by making them long and thin. Give some examples."

- Exploration 25, p. 509—"Unlike a cone or cylinder, it is impossible to make an accurate two-dimensional net for a sphere. For this reason, maps of earth are distorted. The Mercator projection is one way to show the earth. How is this projection made?"

The assessments revised:

- Consult to the United Nations on the least controversial two-dimensional map of the world, after having undertaken Exploration 22.
- Investigate the relationship of the surface areas and volume of various containers (e.g., tuna fish cans, cereal boxes, Pringles, candy packages, etc.). Do they maximize volume? Minimize cost? If not, why not? Consider what nonmathematical variables determine container shape and size.

What we are really seeking evidence of in context-bound assessment is a combination of technical skill and good judgment in its use. A "good judge," said Dewey, "has a sense of the relative values of the various features of a perplexing situation," has "horse sense," has the capacity to "estimate, appraise, and evaluate," and has "tact and discernment." Those who judge well, whether it be in matters of numeracy or human interaction, bring expertise to bear intelligently and concretely on unique and always incompletely understood events. Thus, merely "acquiring information can never develop the power of judgment. Development of judgment is in spite of, not because of, methods of instruction that emphasize simple learning. . . . [The student] cannot get power of judgment excepting as he is continually exercised in forming and testing judgments."³¹

It is noteworthy that fields with the incentive to better merge theory and praxis (engineering, medicine, business, law, etc.) have gravitated to the case- or problem-based learning method of instruction and assessment, in which context is key. Yet, it is still rare in mathematics testing to ask students to confront questions such as quiz questions 1, 3, 5, and 6 until very late in a mathematics career, if at all.

Why is this so? I believe part of the answer is a tacit (and false) learning theory that dominates mathematics education, which might be vocalized as follows: "First, you have to learn all the basics, in the logical order of the elements (thus disconnected from experiential and historical context), using only paper and pencil; *then* you can ask important questions and confront 'real' problems." By that argument, of course, we would never allow little kids to play the difficult game of soccer until years of desk-bound study or let medical students make rounds to see patients. This is simply un-thought through pedagogy—a bad habit—abetted by overreliance on textbooks, built as they are on the logic of mathematics ideas instead of the logic of pedagogy to maximize the understanding of those ideas.³²

In no area of human performance is it true that years of drills and facts must precede all attempts to perform. That view is truly premodern. And as noted above, the validity of tests that follow from this assumption is open to question: evidence of competence cannot be had from exclusive reliance on results from “sideline drills,” for the same reason that ability to cite the textbook meaning of a symptom is not an accurate predictor of performance ability in medicine.

Assessment of QL requires challenges that are essentially not well structured or even well defined; problems that are, well, *problematic*. As in book literacy, evidence of students’ ability to play the messy game of the discipline depends on seeing whether they can handle tasks without specific cues, prompts, or simplifying scaffolds from the teacher-coach or test designer. In QL, students confront situations that have no signs pointing to the right algorithm or solution path. This raises havoc with traditional psychometrics because it is the exact opposite of an effective test “item.”

Because real problems are messy and not amenable to unequivocal final answers, we need to see how students respond to such uncertainties metacognitively. A realistic assessment would thus always ask for a self-assessment. “How well did your proposed answer or solution work? What were the strengths and weaknesses of that approach? What adjustments need to be made, based on your approach and the resultant effects?” Unless they confront such questions, students will continue to exit formal schooling with the belief that merely giving back what was taught is a sufficient indicator of numeracy, or believing that performance is a ritual response to an academic prompt. This is why Norm Frederiksen, a former senior researcher at Educational Testing Service, declared that the real bias of the SAT was not related to content but to format: the neat and clean character of test items versus the messy and uncertain character of the challenges put before us in life.³³

Genuine fluency always demands creativity, not “plug and chug.” It is worth recalling that Bloom and his colleagues (who developed the *Taxonomy of Educational Objectives*) described “synthesis” as always “creative,” and “application” as requiring “novel situations or problems.”³⁴ Fifty years on, the *Taxonomy* still has not achieved its purpose. Two generations (or more) of mathematics educators have misunderstood what Bloom meant. Understanding *why* must be part of the QL agenda, too. Usiskin speculated in *Mathematics and Democracy* that the problem is training. Most teachers think of “application” as “word problems”:

The many examples of the need for quantitative literacy offered in the case statement can easily lead us to wonder why so little has been

accomplished. I believe the problem relates in part to a perception by the majority of mathematics teachers about the “word problems” or “story problems” they studied in high school. . . . These problems have little to do with real situations and they invoke fear and avoidance in many students. So it should come as no surprise that current teachers imagine that “applications” are as artificial as the word problems they encountered as students, and feel that mathematics beyond simple arithmetic has few real applications.³⁵

We see the urgency of this issue more clearly now in the fallout from the disputed Florida election results. The idea of “margin of error” was made real and high stakes; the consequences of failing to consider the lessons of past, close local elections have come back to haunt us. (In retrospect, it seems amazing that there was no procedure for considering a very close election as a statistical tie that then would call forth additional means for fairly filling the office in question.) The consequences of a person’s work need to be *felt in context*—in the same way they are on the playing field or in the auditorium. A challenge, then, is to engineer assessment so that students have far more direct and powerful experiences of the actual effects of their work—on themselves, on other people, and on situations, be they real or virtual. We need an intellectual Outward Bound, as I like to call it, or better assessment software that is more like *Oregon Trail* than *Reader Rabbit*.

Core Assessment Tasks

Because so few current mathematics assessments seem to pass muster, I would propose that we commission a national team of experts to develop the equivalent of the Olympic decathlon in mathematics: 100 key tasks, forming the mathematics “centalon.” The parallel is apt. Around the time Harvard was trying out its new physics performance tests, the modern Olympics were being invented. One of the challenges faced was an assessment problem related to our concern. The task was to design a performance test, over a few days, for well-rounded “athleticism” in all its manifestations, with no favoritism given to speed, strength, hand-eye coordination, or stamina. Thus, the decathlon.

What would be the equivalent in mathematics? We need to know: too few teachers can today answer these questions: What are the most important and representative problems in mathematics? What types of complex challenges should we seek evidence of to deem a graduate quantitatively literate? What might be 100 performances at the heart of QL in which high school students and college underclassmen should be certified, along the 1892 Harvard model? What genres of problems (again to use the parallel with reading and writing literacy) sum up the domain of real performance that a highly literate person should be able to master? Too much of the reform debate has been cast in

terms of mathematics content. In light of the bad habits and lack of imagination in mathematics assessment, we need a sage account of priority performances and situations.

Lauren Resnick and her colleagues in the New Standards project developed an answer a few years back. Beyond the mix of open-ended and constructed-response questions developed for the standardized examinations, there is a mathematics portfolio to be assembled by students locally. Each student submits evidence of work in conceptual understanding of number and operation, geometry and measurement, and functions and algebra; and exhibits in problem solving, data study, mathematical modeling, design of a physical structure, management and planning, pure mathematical investigation, and history of a mathematical idea. In addition, students provide evidence of 12 discrete skills (e.g., “know how to write a simple computer program to carry out computations to be repeated many times”).³⁶ Apropos the issue of context, under design of a physical structure the guidelines say: “Show that you can design a physical structure that meets given specifications . . . [and] explain how your design meets its purpose and how it can be built within constraints (physical, functional, budgetary, aesthetic).”

An ironic weakness in this approach is that the evidence sought was framed by a very broad-brush and context-less set of guidelines, with no reference to any specific big ideas, core content, or key situations. For example, the guidelines under problem solving say, “Show that you can formulate a variety of meaningful problems . . . use problem-solving strategies to solve non-routine and multi-step problems . . .”—without offering any examples or criteria as to what such problems are. We also gain no sense here of the key questions that lie at the heart of thoughtful numeracy.

Questions offer a key doorway into identifying bigger ideas and more meaningful work. In *Understanding by Design*,³⁷ we coach faculties in developing “essential questions” for all units, courses, and programs as a way to focus their teaching on more justified intellectual priorities. Sadly and predictably, mathematics teachers have the hardest time of any group doing the work. Indeed, more than a few mathematics teachers have told me there are no big ideas in mathematics. It is a “skills” discipline, they say.³⁸

Here, for example, is a draft curricular unit in algebra from a veteran teacher who was trying for the first time to organize his lessons around big ideas:

Course Title: Integrated Mathematics I
Topics: Linear Equations, Linear Graphs
Understandings:
 First degree equations give linear graphs.
 Intercepts are zeros.
 Slope intercept form ($y = mx + b$).
 Standard form ($Ax + By = C$).

Essential Questions:

- What does slope measure?
- How is slope calculated?
- How do you graph a line from an equation?
- How do you write an equation of a line from a graph?

We look in vain here for meaningful issues and challenges that might stimulate student thought or inquiry. But change is occurring. Consider the following responses to a design exercise by a small group of high school mathematics teachers:

Unit goal, with reference to big idea: Students will understand measures of tendency (and other methods for grading, voting, and ranking)

Thought-provoking questions on unit goal:

1. By what mathematical method should grading and rankings be done to be most fair?
2. Should voters be allowed to rank order candidates? Are there defensible alternatives to our voting system?
3. Is the mathematically sound solution always the most objectively fair solution?

Predictable misunderstandings:

1. Computing the “average” or “majority” is the only fair method
2. Mathematics cannot help us resolve differences of opinion about fairness

Interesting investigations:

1. Saylor point system in soccer standings
2. Distributive and rank order voting
3. New grading systems for students (median/standard deviation/throw out high and low, etc.)

Or, consider these thought-provoking and perspective-providing questions generated by the mathematics department at the Tilton School, based on a multiyear effort focused on *Understanding by Design*.³⁹

Do mathematical ideas exist separate from the person who understands and can communicate them?

1. What is a quantifiable idea? Is any idea quantifiable? If not, when and why not?
2. How do we determine what makes mathematical ideas true, proven, and/or usable?
3. How do we use mathematical ideas to decipher and explain the world we live in?
4. In what ways is mathematical thinking and logic applicable outside the realm of mathematics? What are the limits or dangers of that extension, if any?
5. In what ways is mathematics rigid, fixed, and systematic? In what ways is mathematics aesthetic, elegant, flexible, and ever expanding?

The Tilton mathematics faculty have committed to address these questions in all their courses. The *Understanding by Design* model requires that the questions be matched by assessments and enabling lessons to ensure that the questions are not idle but integral.

It Is Not the Problems But What We Score that Ultimately Matters

Even the most wonderful, realistic challenges are ruined by foolish scoring and grading practices. Perhaps nothing reveals the need for fundamental reform in mathematics more than the propensity of teachers to use simplistic scoring on one-time test items: the answer is either right or wrong (with partial credit sometimes granted, although with no explicit criteria for points taken off). But if all understanding is a matter of degree, existing along a continuum and subject to disagreement (we can have different understandings of the same problem), there is much teacher baggage to throw overboard in giving grades and scores. If we seek evidence of a student's explanation of confusing data, what does a range of explanations look like—from the most simplistic to the most sophisticated on any of my 11 quiz questions? For example, what does a novice understanding of Hoops McGinty's basketball planetarium problem of scale or the retirement funds problem look like compared with a sophisticated answer?

We can ask such questions meaningfully precisely because the proposed test questions are by design not technique-unique, unlike almost all current mathematics test questions. QL questions can be asked of high school sophomores or college seniors, with profit, just as writing prompts and soccer games can be used from K–16. We find that when we assess this way, some students with less technical skill propose better solutions than students with more advanced technical knowledge, as in the M&M's problem. We would expect such results if we were assessing for QL, just as we expect and actually find some younger performers in writing whose language is less developed but more powerful than that of older students with more schooling and vocabulary under their belts.

But are teachers of mathematics ready to think this way? Consider the following two student answers to the same problem to see how even an experienced teacher of twelfth-grade mathematics can have difficulty escaping our common habits of testing and grading:

Consider an ice cream sugar cone, 8 cm in diameter and 12 cm high, capped with an 8 cm in diameter sphere of luscious rich triple-chocolate ice cream. If the ice cream melts completely, will the cone overflow or not? How do you know—explain your answer.

Answer 1: We must first find the volume of the cone and the ice cream scoop:

$$\begin{aligned} V_{\text{cone}} &= (1/3)\pi r^2 h \\ &= (1/3)\pi 4^2 \times 12 \\ &= 201.06 \text{ cm}^3 \end{aligned}$$

$$\begin{aligned} V_{\text{scoop}} &= (4/3)\pi r^3 \\ &= (4/3)\pi \times (4)^3 \\ &= (4/3) \times 201.06 \text{ cm}^3 \\ &= 268.08 \text{ cm}^3 \end{aligned}$$

We now see that the scoop of ice cream has a volume that is well over 50 cm more than the cone's volume. Therefore it is unlikely that the melted ice cream could fit completely inside the cone. However, as all ice cream lovers like myself know, there is a certain amount of air within ice cream [therefore experiments would have to be done].

Answer 2: Obviously, the first thing to do would be to plug in the values in the equations for the volume of a cone and sphere [the student performs the same calculations as above]. From this we can see that the ice cream will not fit in the cone.

Now I will compare the two formulas:

$$\begin{aligned} (4/3)\pi r^3 &= (1/3)\pi r^2 h \\ 4\pi r^3 &= \pi r^2 h \\ 4\pi r &= \pi h \\ 4r &= h \end{aligned}$$

From this final comparison, we can see that if the height of the cone is exactly 4 times the radius, the volumes will be equal . . . [The student goes on to explain why there are numerous questions about ice cream in real life that will affect the answer, e.g., Will the ice cream's volume change as it melts? Is it possible to compress ice cream?, etc. He concludes by reminding us that we can only find out via experiment.]

The second explanation is surely more sophisticated, displaying a number of qualities we seek in QL (e.g., attention to context, comfort with numbers and data). The student's analysis is mature in part because it subsumes the particular mathematics problem under a broader one: under what conditions are the volumes of different shapes equal? In the first case, all the student has done is calculate the volumes based on the formulas and the given numbers. The second explanation is mature and indicative of understanding by showing perspective: the student has written a narrative as if he were explaining himself to his teacher—mindful, in a humorous way, of audience and purpose.

Nonetheless, the teacher in question gave these two papers the same grade. Both papers gave the correct mathematical answer, after all. Even more alarming, the second paper was

given lower grades than the first paper by a slight majority of middle school mathematics teachers (who seemed to take offense at the student's flippancy) in a national mathematics workshop a few years ago.

Of course, when scoring criteria are unclear, arbitrariness sets in—usually in the form of scoring what is easiest to see or scoring based on the teacher's unexamined and semiconscious habits. That is why learning to score for inter-rater reliability (as is done with Advanced Placement essays and in state and district tests) is such a vital part of any successful reform effort. Yet over 50 percent of teachers of mathematics in our surveys argued that rubrics are “not needed” in mathematics and that, in any event, such scoring is “too subjective.”

What if mathematics teachers routinely had to use multiple criteria with related rubrics in the assessment of performance? Here are five possible criteria, with the top-level descriptor from each rubric (used in the pilot statewide performance assessments in North Carolina, mentioned above)⁴⁰:

- *Mathematical Insight.* Shows a sophisticated understanding of the subject matter involved. The concepts, evidence, arguments, qualifications made, questions posed, and/or methods used are expertly insightful, going well beyond the grasp of the topic typically found at this level of experience. Grasps the essence of the problem and applies the most powerful tools for solving it. The work shows that the student is able to make subtle distinctions and to relate the particular problem to more significant, complex, and/or comprehensive mathematical principles, formulas, or models.
- *Mathematical Reasoning.* Shows a methodical, logical, and thorough plan for solving the problem. The approach and answers are explicitly detailed and reasonable throughout (whether or not the knowledge used is always sophisticated or accurate). The student justifies all claims with thorough argument: counterarguments, questionable data, and implicit premises are fully explicated.
- *Contextual Effectiveness of Solution.* The solution to the problem is effective and often inventive. All essential details of the problem and audience, purpose, and other contextual matters are fully addressed in a graceful and effective way. The solution may be creative in many possible ways: an unorthodox approach, unusually clever juggling of conflicting variables, the bringing in of unobvious mathematics, imaginative evidence, etc.
- *Accuracy of Work.* The work is accurate throughout. All calculations are correct, provided to the proper degree of precision/measurement error, and properly labeled.
- *Quality of Communication.* The student's performance is persuasive and unusually well

presented. The essence of the research and the problems to be solved are summed up in a highly engaging and efficient manner, mindful of the audience and the purpose of the presentation. There is obvious craftsmanship in the final product(s): effective use is made of supporting material (visuals, models, overheads, videos, etc.) and of team members (when appropriate). The audience shows enthusiasm and/or confidence that the presenter understands what he/she is talking about and understands the listeners' interests.

Note that these criteria and rubrics provide more than a framework for reliable and valid scoring of student work. They also provide a blueprint for what the assessment tasks should be. Any assessment must be designed mindful of the rubrics so that the criteria are salient for the specifics of the proposed task. That compels teachers and examination designers to ground their designs in the kinds of complex and nonroutine challenges at the heart of QL. Rather than requiring a new array of secure tests with simplistic items, we should be requiring the use of such rubrics in all assessment, local and statewide.

This approach has an important parallel in literacy assessment. In miscue analysis, we make readers' strategies and renderings explicit, helping them see where they succeeded and where they did not and why, and where a misreading is plausible and sensible and where not, so that both learner and teacher come to better understand reading performance. But we rarely do such miscue analysis at higher grades in any subject, despite its power for the learner.

Here is what happened when some mathematics students were taught to self-assess their work through an error analysis after a test: “After we graded their tests, students were asked to evaluate their own performance. . . . Each student was required to submit a written assessment of test performance that contained corrections of all errors and an analysis of test performance. . . . We directed our students to pay particular attention to the *types* of errors they made. . . . They were to attempt to distinguish between conceptual errors and procedural errors.” The teachers found this process extremely useful: “Student self-assessment and the resulting student-teacher dialogue were invaluable in drawing a clear picture of what students were thinking when errors were made.” But they also reported that students found the task demanding. In particular, teachers in training “had difficulty weighing the seriousness of an error” and seemed “to have difficulty assigning grades to their work. . . . Many had a tendency to weigh effort heavily regardless of the quality of the product.”⁴¹

The previously cited example from North Carolina included a rubric for assessing mathematical insight. Not only is

insight vital to mathematics but it can and must be taught, hence assessed, as part of quantitative literacy.⁴² This is yet another “radical” aspect of the QL agenda: even though we typically flinch from assessing insight because of fear or ignorance, we must assess what we value highly, including mathematical intuition or insight. *Of course* insight is measurable: anyone who can see through messy contexts, unfamiliar situations, or ill-structured and seemingly intractable problems to an effective solution has insight. We think insight is impossible to assess because we rarely use test items that require insight.

An article on the Third International Mathematics and Science Study (TIMSS) results by a *New York Times* reporter makes the point clearly:

Consider one problem on the test. . . . It shows a string wound around a circular rod exactly four times, creating a spiral from one end of the rod to the other. The problem was asked only of those who had studied advanced mathematics: what is the string’s length, if the circumference of the rod is 4 centimeters and its length is 12 centimeters?

The problem is simply stated and simply illustrated. It also cannot be dismissed as being so theoretical or abstract as to be irrelevant. . . . It might be asked about tungsten coiled into filaments; it might come in handy in designing computer chips. . . . It seems to involve some intuition about the physical world. . . .

It also turned out to be one of the hardest questions on the test. . . . [Only] 10% solved it completely. But the average for the United States was even worse: just 4 % The rate of Swedish students’ success was 6 times greater than that of the Americans; Swiss students did more than four times as well. . . .

What is so interesting about this particular example . . . is that it requires almost no advanced mathematics at all. It does not require memorization of an esoteric concept or the mastery of a specialty. It requires a way of thinking. If you cut the cylinder open and lay it flat, leaving the string in place, you get a series of four right triangles with pieces of string as their diagonals. The length of the string is calculated using a principle learned by every ninth grader. . . .

Nothing could be a better illustration of the value of teaching a mathematical way of thinking. It requires different ways of examining objects; it might mean restating problems in other forms. It can demand a playful readiness to consider alternatives and enough insight to recognize patterns.⁴³

Here are some indicators (“look-fors”) of insight, derived in part from analysis of the six facets of understanding mentioned earlier. These indicators can be used as guidelines for designing tasks that require such abilities. Insight is revealed by the ability to show:

- Other plausible ways to look at and define the problem;
- A potentially more powerful principle than the one taught or on the table;
- The tacit assumptions at work that have not been made explicit;
- Inconsistency in current versus past discussion;
- Author blind spots;
- Comparison and contrast, not just description;
- Novel implications; and
- How custom and habit influence the views, discussion, or approach to the problem.

The basic blueprint for tasks that can help us assess insight was provided by a grumpy workshop participant many years ago: “You know the trouble with kids today? They don’t know what to do when they don’t know what to do!” But that is because our assessments are almost never *designed* to make them not know what to do.

Longitudinal Rubrics

sophistication. Of a person: free of naiveté, experienced, worldly-wise; subtle, discriminating, refined, cultured; aware of, versed in, the complexities of a subject or pursuit.⁴⁴

As suggested in our discussion of rubrics and criteria, understandings are not right or wrong. They exist on a continuum running from naïve to expert. To find evidence of QL, we need something more than scores on new test questions. We need a whole different way of charting progress over time. We need to validly and reliably describe a student’s degree of understanding of core tasks over time, just as we have done for a few decades in English literacy. QL requires that we discriminate between naïve, developing, competent, and expert performance (to suggest four key points on a continuum of ability).

Some such rubrics already exist in mathematics, with greater refinement. Consider the following from Great Britain representing what might be termed the British rubric for QL:

Attainment Target 1: MA1. Using and Applying Mathematics:

- *Level 1.* Pupils use mathematics as an integral part of classroom activities. They represent their work with objects or pictures and discuss it. They recognise and use a simple pattern or relationship.
- *Level 2.* Pupils select the mathematics they use in some classroom activities. They discuss their work using mathematical language and are beginning to represent it using symbols and simple diagrams. They explain why an answer is correct.
- *Level 3.* Pupils try different approaches and find ways of overcoming difficulties that arise when they are solving problems. They are beginning to organise their

work and check results. Pupils discuss their mathematical work and are beginning to explain their thinking. They use and interpret mathematical symbols and diagrams. Pupils show that they understand a general statement by finding particular examples that match it.

- *Level 4.* Pupils are developing their own strategies for solving problems and are using these strategies both in working within mathematics and in applying mathematics to practical contexts. They present information and results in a clear and organised way. They search for a solution by trying out ideas of their own.
- *Level 5.* In order to carry through tasks and solve mathematical problems, pupils identify and obtain necessary information. They check their results, considering whether these are sensible. Pupils show understanding of situations by describing them mathematically using symbols, words and diagrams. They draw simple conclusions of their own and give an explanation of their reasoning.
- *Level 6.* Pupils carry through substantial tasks and solve quite complex problems by independently breaking them down into smaller, more manageable tasks. They interpret, discuss and synthesise information presented in a variety of mathematical forms. Pupils' writing explains and informs their use of diagrams. Pupils are beginning to give mathematical justifications.
- *Level 7.* Starting from problems or contexts that have been presented to them, pupils progressively refine or extend the mathematics used to generate fuller solutions. They give a reason for their choice of mathematical presentation, explaining features they have selected. Pupils justify their generalisations, arguments or solutions, showing some insight into the mathematical structure of the problem. They appreciate the difference between mathematical explanation and experimental evidence.
- *Level 8.* Pupils develop and follow alternative approaches. They reflect on their own lines of enquiry when exploring mathematical tasks; in doing so they introduce and use a range of mathematical techniques. Pupils convey mathematical or statistical meaning through precise and consistent use of symbols that is sustained throughout the work. They examine generalisations or solutions reached in an activity, commenting constructively on the reasoning and logic or the process employed, or the results obtained, and make further progress in the activity as a result.
- *Exceptional Performance.* Pupils give reasons for the choices they make when investigating within mathematics itself or when using mathematics to analyse tasks; these reasons explain why particular lines of enquiry or procedures are followed and others rejected. Pupils apply the mathematics they know in familiar and unfamiliar contexts. Pupils use

mathematical language and symbols effectively in presenting a convincing reasoned argument. Their reports include mathematical justifications, explaining their solutions to problems involving a number of features or variables.⁴⁵

The phrasing throughout nicely indicates that students must gain not merely more specialized knowledge of mathematics per se but also increased understanding of mathematics in use. Again, these rubrics do more than show how we should score. They properly serve as criteria for the development of tasks requiring such abilities.

A Sad Irony: Our Ignorance About How Assessment Works

Why is the problem of innumeracy so difficult to solve if many of the ideas cited above are now in use in some high schools and colleges worldwide? Because too many mathematics educators have not been forced to defend their assessments or challenge their habits. Typical tests would not pass muster if held up to the light.

Please suspend disbelief. (If you are still reading, you are *not* among the educators I worry about.) Consider such common unthinking and questionable grading habits in mathematics classes as forcing a small sample of scores to fit a bell curve or computing grades via a calculation of the mean score, as if both were the only possible or defensible practices. In fact, both practices are ironic examples of the *thoughtless use of algorithms*—in the context of classrooms and educational goals. Yet when surveyed, many mathematics teachers claim that the grades they assign are *more* valid than those in other fields because “mathematics is inherently more precise and objective,” as one teacher put it. Many of the same teachers are surprisingly uninterested in relevant educational research. “That may be true in their study,” is the common refrain, “but not in my class.”

A more serious misconception has to do with the relation of QL to state tests. Consider the universal excuse offered by mathematics educators when many of these reform ideas are presented in workshops. “That’s great, we would love to do this, but we cannot. We have to teach to the test.” In our surveys, the argument for being unable to teach for understanding because of tests comes more from mathematics teachers than any other group. Yet this “teach to the test” mantra ironically turns out on closer inspection to be an example of a *form of innumeracy* described by John Paulos in his deservedly best-selling book on the subject: namely, the educator is often confusing causality with correlation.⁴⁶

The extended lament makes this conflation clearer: “Well, we’re told that we must get test scores up. So, clearly we cannot teach for the kind of QL and understanding being discussed here. We have no choice but to teach to the test,

to use in our testing the kinds of items the state test uses. We have to cover so much content superficially. . . .” Two quick arguments will get the speaker and listeners to stop and consider the reasoning implied:

1. Let me see if I understand. Aren’t you saying, then, that the way to raise test scores is to teach less competently (since you admit to being forced to use a superficial and scattered approach, as opposed to teaching for understanding)?
2. If you’re right, then by analogy I should practice the physical examination all year if I want to have the best possible result on my annual medical checkup.

The bewildered looks speak volumes. They do not protest my analysis; it clearly puzzles them. Then, I march out the data: National Assessment of Educational Progress (NAEP), the Second International Mathematics and Science Study (SIMSS), the Third International Mathematics and Science Study (TIMSS), and other credible test data all suggest that mathematics instruction is not working for the vast majority of American students. More generally, in an exhaustive meta-analysis Paul Black and Dylan Wiliam have shown that improving the quality of classroom feedback offers the greatest performance gains of any single teaching approach: “There is a body of firm evidence that formative assessment is an essential component . . . and that its development can raise standards of achievement. We know of no other way of raising standards for which such a strong prima facie case can be made.”⁴⁷ A score, by itself, is the least useful form of feedback; rubrics provide a great deal more, and specific comments in reference to the rubrics—*contextual* feedback—provide still more.⁴⁸

The National Academy of Sciences recently released a set of summary findings on how people learn, in which they concluded:

Students develop flexible understanding of when, where, why, and how to use their knowledge to solve new problems if they learn how to extract underlying principles and themes from their learning exercises. [But m]any assessments measure only propositional (factual) knowledge and never ask whether students know when, where, and why to use that knowledge.⁴⁹

The most telling research stems from the TIMSS teaching study, summarized in *The Teaching Gap*.⁵⁰ J. Stigler and J. Hiebert present striking evidence of the benefits of teaching for understanding in optimizing performance (as measured by test scores). When correlated with test results, the data clearly show that although Japanese mathematics teachers in middle school cover fewer topics, they achieve better results. Rather than defining themselves as teachers of “skills” in which many discrete skills are covered (as American teachers identify themselves and as videos of their classes reveal), the primary aim in Japanese middle school classes is conceptual understanding. Other data

confirm this view: Japanese teachers teach fewer topics in greater depth than do American teachers. They emphasize problem-based learning, in which rules and theorems are typically derived, not merely stated and reinforced through drill.

How did Japan develop such a sophisticated conception of instruction in mathematics? The authors answer: the key is a continuous-progress group professional development research process called Lesson Study. The phrase sums up a long-standing tradition in Japan whereby K–8 teachers work all year in small teams to study student performance and develop, refine, and teach exemplary lessons to improve results. The process involves constant experimentation and peer review:

1. Group defines problem and plans lesson
2. A group member teaches the lesson
3. Group evaluates and revises the lesson
4. Another member teaches revised lesson
5. Evaluation by entire faculty
6. Share results in print and at “lesson fairs”

With respect to the next to the next-to-last step, the authors note:

Evaluating and Reflecting, Again. This time, it is common for all members of the school faculty to participate in a long meeting. Sometimes an outside expert will be invited to attend as well. . . . Not only is the lesson discussed with respect to what these students learned, but also with respect to more general issues raised . . . about teaching and learning.⁵¹

Is it any wonder, then, if this process is customary, that the typical Japanese teacher would develop more sophisticated curricular and assessment designs? Stigler and Hiebert ironically note that one reason the Japanese process works is that the teachers’ research, unlike university research, is done *in their context*, i.e., research leading to “knowledge that is immediately usable.”⁵² Interestingly, many educators in our workshops readily admit that local faculty are not yet ready for such a system, given the American norm of individual isolationism.

An unflinching view of the situation suggests that many of the problems that have led to the current concern with QL are of the Pogo variety: we have met the enemy and it is us. But that is also good news, because it is in our power as educators to change things. Let us begin in an obvious place: with explicit design standards for assessment and curriculum to counter habit, many models of exemplary design, and policy-based incentives to honor the standards.

Who, then, might drive this reform, if we are in danger of having the blind leading the blind? My Swiftian modest proposal, based on the Pogo caution: Do not let mathematics teachers and professors dominate the discussion. If realistic assessment is what we want and

school rarely offers it, let us go beyond school and college walls to the obvious source of contextual insight: people who work and live in the many QL contexts of the wider world. Such a strategy is implied in the case statement in which the “expressions of quantitative literacy” sketch out dozens of venues from which assessment could be taken.⁵³

Interestingly, this is what all vocational teachers in New York State must do. They assemble a team of people in their trade, in what is termed a Consultant Committee. The group discusses the teacher’s curriculum and assessments twice a year (as well as job placement issues). Why not require all academic departments to do this? In mathematics, let us assemble teams composed of engineers, mapmakers, software designers, baseball statisticians, machine workers, accountants, lab technicians, etc., to advise mathematics educators on how their subject is really used and on what persistent problems they encounter concerning their workers’ abilities. Mathematics educators then could tweak draft designs based on these findings into useful assessments. Finally, to complete the process, we could turn to teams of academics and practitioners for peer review of the designers’ work.

The core message of *The Teaching Gap*⁵⁴ is that ongoing teacher research and development is the key to local improvement in all facets of teaching. According to Stigler and Hiebert, the typical Japanese teacher is now far more sophisticated and focused than is the typical American teacher. Why? Because in Japan the culture of the school and the daily demands of the profession make research and development part of the job. Let us base the process of reform on this axiom: to be valid, of high quality, and credible to all key constituencies, assessment requires collaboration in design, the making public of all design work, and peer review against design standards.

Finally, let us never forget that although the issues here seem technical and political, at bottom they are moral. The aim in any realistic assessment process is to gather *appropriate* evidence and render a *considered* judgment, much like what the judge has to do in a civil trial. The analogy is useful because such a judgment is always fraught with uncertainty; it is never neat and clean; it is always in context. The evidence is weighed, considered, argued about. The standard for conviction is that there be a preponderance of evidence of the right kind. To “convict” a student of understanding similarly requires compelling and appropriate evidence and argument: the student should be considered innocent of understanding unless proven otherwise by a preponderance of evidence. That is a high standard, and appropriately so—even though impatient teachers, testers, and policymakers may wish it were otherwise. Alas, for too long we have gotten away with verdicts in mathematics education using highly circumstantial, indirect evidence. It is high time we sought justice.

Notes

1. From N. Movshovitz-Hadar and J. Webb, *One Equals Zero, and Other Mathematical Surprises* (Emeryville, CA: Key Curriculum Press, 1998).
2. *New York Times*, 13 August 2001, A35.
3. Lynn Arthur Steen, ed., *Mathematics and Democracy: The Case for Quantitative Literacy* (Princeton, NJ: National Council on Education and the Disciplines, 2001), 1–22.
4. Steen, *Mathematics and Democracy*.
5. Steen, *Mathematics and Democracy*.
6. Deborah Hughes-Hallett, “Achieving Numeracy: The Challenge of Implementation,” in *Mathematics and Democracy*, Steen, ed., 93–98.
7. Hughes-Hallett, “Achieving Numeracy,” in *Mathematics and Democracy*, Steen, ed.
8. Peter T. Ewell, “Numeracy, Mathematics, and General Education,” in *Mathematics and Democracy*, Steen, ed., 37–48.
9. Howard Gardner, *The Unschooled Mind: How Children Think and How Schools Should Teach* (New York, NY: Basic Books, 1991), 165.
10. See, for example, the theme issue “The Mathematical Miseducation of America’s Youth” in *Phi Delta Kappan* 80:6 (February 1999), and Schoenfeld, A. 1992. “Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics.” In D. A. Grouws, ed., *Handbook of Research on Mathematics Teaching and Learning*. New York: Macmillan, 334–371..
11. Larry Cuban, “Encouraging Progressive Pedagogy,” in *Mathematics and Democracy*, Steen, ed., 87–92.
12. J. Stigler and J. Hiebert, *The Teaching Gap: Best Ideas from the World’s Teachers for Improving Education in the Classroom* (New York, NY: Free Press, 1999).
13. Cf. Dan Kennedy, “The Emperor’s Vanishing Clothes,” in *Mathematics and Democracy*, Steen, ed., 55–60.
14. Grant Wiggins, *Assessing Student Performance* (San Francisco: Jossey-Bass, 1996).
15. See Grant Wiggins, *Educative Assessment: Assessment to Improve Performance* (San Francisco: Jossey-Bass, 1998).
16. P. T. Ewell, “Numeracy, Mathematics, and General Education,” in *Mathematics and Democracy*, Steen, ed., 37:

... the key area of distinction [between QL and mathematics] is signaled by the term literacy itself, which implies an integrated ability to function seamlessly within a given community of practice. Literacy as generally understood in the verbal world thus means something quite different from the kinds of skills acquired in formal English courses.
17. Daniel Resnick and Lauren Resnick, “Standards, Curriculum, and Performance: A Historical and Comparative Perspective,” *Educational Researcher* 14:4 (1985): 5–21.
18. A cautionary note, then, to professional development providers: information, evidence, and reason will not change this habit, any more than large numbers of people quit abusing cigarettes, alcohol, and drugs because they read sound position papers or hear professionals explain why they should quit. Habits are changed by models, incentives, practice, feedback—if at all.
19. “The Present Requirements for Admission to Harvard College,” *Atlantic Monthly* 69:415 (May 1982): 671–77.

20. See, for example, G. Cizek, "Confusion Effusion: A Rejoinder to Wiggins," *Phi Delta Kappan* 73 (1991): 150–53.
21. Earlier versions of these standards appeared in Wiggins, *Assessing Student Performance*, 228–30.
22. Benjamin Bloom, ed., *Taxonomy of Educational Objectives: Book I: Cognitive Domain* (White Plains, NY: Longman, 1954), 125.
23. B. S. Bloom, G. F. Madaus, and J. T. Hastings, *Evaluation to Improve Learning* (New York, NY: McGraw-Hill, 1981), 265.
24. Bloom, et al., *Evaluation to Improve Learning*, 268.
25. Newmann, W. Secada, and G. Wehlage, *A Guide to Authentic Instruction and Assessment: Vision, Standards and Scoring* (Madison, WI: Wisconsin Center for Education Research, 1995).
26. Wiggins, G. & Kline, E. (1997) *3rd Report to the North Carolina Commission on Standards and Accountability*, Relearning by Design, Ewing, NJ.
27. Federal Aviation Report, as quoted in "A Question of Safety: A Special Report," *New York Times*, Sunday 13 November 1994, section 1, page 1.
28. Jay McTighe and Grant Wiggins, *The Understanding by Design Handbook* (Alexandria, VA: Association for Supervision and Curriculum Development, 1999).
29. Grant Wiggins and Jay McTighe, *Understanding by Design* (Alexandria, VA: Association for Supervision and Curriculum Development, 1998).
30. William James, *Talks to Teachers* (New York, NY: W. W. Norton, 1899/1958).
31. John Dewey, *The Middle Works of John Dewey: 1899-1924* (vol. 15) (Carbondale, IL: Southern Illinois University Press, 1909), 290.
32. A historical irony. For it was Descartes (in *Rules for the Direction of the Mind*) who decried the learning of geometry through the organized results in theorems presented in logical order as needlessly complicating the learning of geometry and hiding the methods by which the theorems were derived. See Wiggins and McTighe, *Understanding by Design*, 149 ff.
33. Norm Frederiksen, "The Real Test Bias," *American Psychologist* 39:3 (March 1984): 193–202.
34. Bloom, *Taxonomy of Educational Objectives: Book I: Cognitive Domain*.
35. Zalman Usiskin, "Quantitative Literacy for the Next Generation," in *Mathematics and Democracy*, Steen, ed., 84.
36. National Center on Education and the Economy, *New Standards: High School Mathematics Portfolio* (Washington, DC: National Center on Education and the Economy, 1995).
37. Wiggins and McTighe, *Understanding by Design*.
38. This narrow-minded comment was echoed repeatedly by American teachers in the TIMSS teaching study, discussed below, with unfortunate results for student performance.
39. Wiggins and McTighe, *Understanding by Design*.
40. Wiggins, G. & Kline, E. (1997) *3rd Report to the North Carolina Commission on Standards and Accountability*, Relearning by Design, Ewing, NJ.
41. Stallings, Virginia and Tascione, Carol "Student Self-Assessment and Self-Evaluation." *The Mathematics Teacher*. NCTM. Reston, VA. 89:7, October 1996. pp. 548-554.
42. Cf. Hughes-Hallett, "Achieving Numeracy," in *Mathematics and Democracy*, Steen, ed., 96–97.
43. Edward Rothstein, "It's Not Just Numbers or Advanced Science, It's Also Knowing How to Think," *New York Times*, 9 March 1998, D3.
44. From the Oxford English Dictionary CD-ROM.
45. From the *National Curriculum Handbook for Secondary Teachers* in England, Department of Education and Employment, 1999. Also available on the Internet at www.nc.uk.net. Note that these rubrics are called "Attainment Targets."
46. Paulos, John Allen. 1990. *Innumeracy: Mathematical Illiteracy and Its Consequences* (New York: Vintage Books).
47. Black, P. J. and D. Wiliam. 1998. "Assessment and Classroom Learning." *Assessment in Education* 5(1) 7-74.
48. See Wiggins, *Educative Assessment*, on the importance of feedback to assessment systems and reform.
49. J. Bransford, A. Brown, and R. Cocking, eds., *How People Learn: Brain, Mind, Experience, and School* (Washington, DC: National Academy Press), 37.
50. Stigler and Hiebert, *The Teaching Gap*.
51. Stigler and Hiebert, *The Teaching Gap*, 112 ff.
52. Stigler and Hiebert, *The Teaching Gap*, 122.
53. Steen, *Mathematics and Democracy*, 9 ff.
54. Stigler and Hiebert, *The Teaching Gap*.