The *Journal of Mathematics Education at Teachers College* is a publication of the Program in Mathematics and Education at Teachers College Columbia University in the City of New York.

Dr. Elizabeth Hagan was discharged from the US Navy as a Lieutenant in 1946 and completed the PhD in Measurement and Evaluation at Teachers College in 1952. Her collaboration with Robert Thorndike resulted in many papers and books including the influential 1961 John Wiley publication, *Measurement and Evaluation in Psychology and Education*. Dr. Hagan concluded her career at Teachers College in 1976 as Acting Dean of Academic Affairs.

Dr. Stuart Weinberg was the Mathematics Department Chairman at Stuyvesant High School before joining the Teachers College faculty as Director of Student Teaching for the Program in Mathematics. Dr. Weinberg has applied his extensive classroom experience to the development of methods of assessing teachers' classroom performance utilizing belief systems and attitudes.

Aims and Scope

The *JMETC* is a re-creation of an earlier publication by the Teachers College Columbia University Program in Mathematics. As a peer-reviewed, semi-annual journal, it is intended to provide dissemination opportunities for writers of practice-based or research contributions to the general field of mathematics education. Each issue of the *JMETC* will focus upon an educational theme. The themes planned for the 2012 Fall-Winter and 2013 Spring-Summer issues are *Equity* and *Leadership*, respectively.

JMETC readers are educators from pre-K-12 through college and university levels, and from many different disciplines and job positions—teachers, principals, superintendents, professors of education, and other leaders in education. Articles to appear in the *JMETC* include research reports, commentaries on practice, historical analyses, and responses to issues and recommendations of professional interest.

Manuscript Submission

JMETC seeks conversational manuscripts (2,500-3,500 words in length) that are insightful and helpful to mathematics educators. Articles should contain fresh information, possibly research-based, that gives practical guidance readers can use to improve practice. Examples from classroom experience are encouraged. Articles must not have been accepted for publication elsewhere. To keep the submission and review process as efficient as possible, all manuscripts may be submitted electronically at www.tc.edu/jmetc.

Abstract and keywords. All manuscripts must include an abstract with keywords. Abstracts describing the essence of the manuscript should not exceed 150 words. Authors should select keywords from the menu on the manuscript submission system so that readers can search for the article after it is published. All inquiries and materials should be submitted to Ms. Krystle Hecker at P.O. Box 210, Teachers College Columbia University, 525 W. 120th St., New York, NY 10027 or at JMETC@tc.columbia.edu.

Copyrights and Permissions

Library of Congress Cataloging-in-Publication Data

Journal of mathematics education at Teachers College
p. cm.
Includes bibliographical references.
ISSN 2156-1397
EISSN 2156-1400
1. Mathematics—Study and teaching—United States—Periodicals
QA11.A1 J963

More Information is available online: www.tc.edu/jmetc

Call for Papers

The "theme" of the fall issue of the *Journal of Mathematics Education at Teachers College* will be *Equity*. This "call for papers" is an invitation to mathematics education professionals, especially Teachers College students, alumni and friends, to submit articles of approximately 2500-3500 words describing research, experiments, projects, innovations, or practices related to equity in mathematics education. Articles should be submitted to Ms. Krystle Hecker at JMETC@tc.columbia.edu by September 1, 2012. The fall issue's guest editor, Mr. Nathan N. Alexander, will send contributed articles to editorial panels for "blind review." Reviews will be completed by October 1, 2012, and final manuscripts of selected papers are to be submitted by October 15, 2012. Publication is expected by November 15, 2012.

Call for Volunteers

This *Call for Volunteers* is an invitation to mathematics educators with experience in reading/writing professional papers to join the editorial/review panels for the fall 2012 and subsequent issues of *JMETC*. Reviewers are expected to complete assigned reviews no later than 3 weeks from receipt of the manuscripts in order to expedite the publication process. Reviewers are responsible for editorial suggestions, fact and citations review, and identification of similar works that may be helpful to contributors whose submissions seem appropriate for publication. Neither authors' nor reviewers' names and affiliations will be shared; however, editors'/reviewers' comments may be sent to contributors of manuscripts to guide further submissions without identifying the editor/reviewer.

If you wish to be considered for review assignments, please request a *Reviewer Information Form*. Return the completed form to Ms. Krystle Hecker at hecker@tc.edu or Teachers College Columbia University, 525 W 120th St., Box 210, New York, NY 10027.

Looking Ahead

Anticipated themes for future issues are:

Fall 2012	Equity
Spring 2013	Leadership
Fall 2013	Modeling
Spring 2014	Teaching Aids

TO OBTAIN COPIES OF *JMETC*

To obtain additional copies of *JMETC*, please visit the *Journal*'s website www.tc.edu/jmetc. The cost per copy delivered nationally by first class mail is $5.00. Payment should be sent by check to *JMETC*, Teachers College Columbia University, 525 W 120th St., Box 210, New York, NY 10027.

Journal of Mathematics Education at Teachers College

Spring – Summer 2012

A Century of Leadership in Mathematics and its Teaching

TABLE OF CONTENTS

Preface

Articles

Assessment Notes from the Field

TABLE OF CONTENTS, continued

Journal of Mathematics Education at Teachers College
Spring–Summer 2012, Volume 3

Assessment for the Common Core Mathematics Standards

Hung-Hsi Wu
University of California at Berkeley

This article makes two simple observations about high-stakes assessments. The first is that, because mathematics is a very technical subject, an assessment item can be mathematically flawed regardless of how elementary it is. For this reason, every assessment project needs the active participation of high level mathematicians. A second point is that high-stakes assessments are inherently a very blunt instrument because they are incapable of accurately measuring the most important aspect of mathematics achievement: sustained sequential thinking. Because the general public and policy makers are not aware of this fact, they tend to read more into such assessment scores than such a limited instrument can deliver. If we want high-stakes assessments to have a positive influence on mathematics education, this article suggests that we should reorient our thinking about how much student achievement such assessments can reliably measure, which is "not very much."

Keywords: mathematical integrity, sequential thinking, driver's license test, basic competence, ingenuity, assessing excellence.

The adoption of the *Common Core State Standards for Mathematics* (2010), hereafter referred to as CCSSM, by forty-five states[1] sets the stage for the real battle: how to implement these standards successfully. There are three major players in this battle: textbooks, professional development for teachers, and assessment. The need for adequate school mathematics textbooks that are compatible with the CCSSM is acute and far from being met. Because commercial interests are the dominant factor here, textbooks are probably not a good subject for an intellectual discussion. As to teachers' need for assistance in acquiring the content knowledge for teaching a CCSSM curriculum, it is no less acute but it is as yet unclear if our nation has the commitment and the resolve to meet this enormous challenge (cf. Wu, 2011b). There remains assessment. This article will try to make some comments on the most glaring pitfalls that await the assessments of the CCSSM and, at the end, some suggestions on how to avoid them.

It is a fact that assessment directly influences or, as some would say, drives the curriculum. At a time when the CCSSM tries to infuse the school curriculum with some mathematical integrity (see, e.g., Wu, 2011a), they also raise our expectations for the *mathematical quality* of the assessment items. The first part of this article will give a fairly detailed examination of some released assessment items from the point of view of their possible impact in the school classroom. These items are taken from National Assessment of Educational Progress (NAEP) assessments, state tests, and two complete summative assessment samples from the Mathematics Assessment Project (MAP). Such an examination, no doubt, will be considered a waste of time in those quarters that call for the elimination of all assessments. This extreme view stems from a misunderstanding of the role that high-stakes assessments ought to play, and it must be said that the whole education establishment has to carry the blame for contributing to this misunderstanding. In the second part of this article, we will point out the inherent limitations of what can be learned about student achievement from high-stakes assessments and hope that, with this understood, we can make such assessments a positive influence in mathematics education.

If we want a good curriculum based on the CCSSM, then we would want the CCSSM assessments to be as good as possible. At the very least, the CCSSM assessment items should be free of obvious mathematical flaws. If the past is any guide, even this lowered expectation may be too high. According to one report (Daro, Stacavage, Ortega, Stefano, & Linn, 2007) as quoted in National Mathematics Advisory Panel, 2008, (p. 8–3 of Chapter 8), mathematically flawed assessment items have been a fact of life thus far:

> Five percent of NAEP items were found to be seriously flawed mathematically at Grade 4, and 4 percent were designated seriously flawed at Grade 8. The state [assessment] items were classified as 7 percent seriously flawed in fourth grade and 3 percent seriously flawed in eighth grade. For [marginally flawed] items, NAEP had 28 percent at Grade 4 and 23 percent at Grade 8, while the state sample had 30 percent at Grade 4 and 26 percent at Grade 8.

It would be natural to speculate that the presence of these flawed assessment items is correlated with the frequent absence of high-quality mathematicians in the education

[1] As of March 15, 2012.

conversation of the recent past. Indeed, the very technical nature of mathematics makes it nearly impossible to uphold mathematical integrity without the help of such mathematicians. If the forthcoming assessment consortia for the CCSSM, the Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortium, hope to avoid past pitfalls, they would do well to engage knowledgeable mathematicians in every phase of their work.

We begin the examination of assessment items released by various education agencies with problems on patterns. These were once extremely popular on state and national assessments but are now rarely found as a result of the urging of many mathematicians, especially James Milgram (undated). Figure 1 shows a typical problem of this type taken from a state standardized test for grade eight. The *expected* correct answer is of course D, but the correct answer is "*cannot be determined*." For example, the output is, for all we know, the following sequence that repeats the four numbers 10, 13, 16, 19, *ad infinitum*:

10, 13, 16, 19, 10, 13, 16, 19, 10, 13, 16, 19, etc.

In this case, we cannot say what the output below n will be because it all depends on whether $n = 4k + 3$ for some integer k (in which case the output is 10), or $n = 4k$ (in which case it is 13), or $n = 4k + 1$ (in which case it is 16), or $n = 4k + 2$ (in which case it is 19).

Mistakes in pattern problems like this are part of the **Wishful Thinking Syndrome** in school mathematics education: give out partial information and students will automatically fill in the missing information to achieve a complete conceptual understanding on their own. For example, tell students that a fraction is a piece of pizza and they will understand that it is actually a number and will learn to add, subtract, multiply, and divide fractions with ease, including how to invert and multiply. For the case at hand, surely giving out the first four terms 10, 13, 16, 19 is sufficient for student to see that $10 = (3 \times 3) + 1$, $13 = (3 \times 4) + 1$, $16 = (3 \times 5) + 1$, $19 = (3 \times 6) + 1$ and, therefore, the output for *n must be* $3n + 1$. The assessment people who made up this pattern problem forgot that mathematics is WYSIWYG, *what you see is what you get*, so that what is not forbidden is permitted. Since there is no statement in the problem that forbids other patterns, there is an infinite number of possibilities for the output. Here is another one:

10, 13, 16, 19, 11, 14, 17, 20, 12, 15, 18, 21, etc.

Would many students choose D as the answer? Probably. Do we want our students to know that the most natural way to continue the pattern is from n to $3n + 1$? Definitely. But do we want students to believe that D is the only correct answer? No, because this would require that we teach students to do mathematics by routinely making unwarranted assumptions (in this case, the assumption that the output is a *linear* function of the input). Assessment, especially high-stakes assessment, sends a powerful signal to the classroom. All educators are obligated to ensure that this signal is the correct one. It is for this reason that we have no choice but to ban such pattern problems from every part of school mathematics education.

Use the chart below to answer question 4.

Input	3	4	5	6	. . .	n
Output	10	13	16	19	. . .	?

4. If the input is n, what will the output be?

A. $n + 3$
B. $n + 7$
C. $3(n + 2) + 1$
D. $3n + 1$

Figure 1.

Because of the increasing awareness that such pattern problems are mathematically untenable, most assessment developers—states, NAEP, TIMSS, etc.—have pulled such items from their webpages. But not entirely. In the 2005 grade 4 NAEP assessment, there is the following question (see "Determine next number in given pattern" in National Assessment of Educational Progress (NAEP)):

3, 6, 5, 8, 7, 10, 9, ?

In the number pattern above, what number comes next?

Answer: _____

This question is worth discussing because it reveals what is wrong with pattern items from a different angle. First of all, avoiding the multiple choice format is a step forward: the test item would have been a perfectly good question had it been changed into a constructed response item such as the following:

In the number pattern above, what number comes next and why?

This would the give students a chance to specify the pattern and then deduce what the next number must be. As it is, however, the expected answer is 12, and the reasoning, which is also a derivative of the Wishful Thinking Syndrome, is probably the following: seeing 3, 5, 7, 9 in the odd-numbered positions and 6, 8, 10 in the even-numbered positions, students should automatically be able to fill in the blanks. Therefore, the sequence can be thus described:

The sequence

$$a_1, a_2, a_3, a_4, a_5, \ldots$$

satisfies $a_1 = 3$ and $a_2 = 6$, and for all nonnegative integers n, $a_{2n+3} = 2 + a_{2n+1}$ and $a_{2n+2} = 2 + a_{2n}$

Granting this rule, then indeed 12 is the correct answer. Unfortunately, this is not the only way to generate the sequence 3, 6, 5, 8, 7, 10, 9. Here is another one: it is known that there is a polynomial $p(x)$ of degree 5 (the Lagrange Interpolation polynomial) so that

$$p(3) = 6,\ p(6) = 5,\ p(5) = 8,\ p(8) = 7,$$
$$p(7) = 10,\ p(10) = 9.$$

Now let k be a fixed number and let $q(x)$ be the polynomial so that

$$q(x) = p(x) + k(x-3)(x-6)(x-5)(x-8)(x-7)(x-10).$$

It is clear that

$$q(3) = 6,\ q(6) = 5,\ q(5) = 8,\ q(8) = 7,$$
$$q(7) = 10,\ q(10) = 9.$$

We define a sequence $b_1, b_2, \ldots$ so that

$$b_1 = 3, \text{ and } b_{n+1} = q(n) \quad \text{for all } n = 1, 2, 3, \ldots$$

The first six members of this sequence then coincide with 3, 6, 5, 8, 7, 10, 9, but the next number is $q(9) = p(9) - 144k$. Since k is arbitrary, the next number cannot be determined on the basis of the information given in the problem.

We should put this discussion in perspective. We are not saying that an average fourth grader would know the preceding reasoning and recognize that this NAEP item is not well-posed, nor are we saying that its presence did significant damage to the 2005 NAEP scores. What we *are* saying is that, so long as the NAEP scores impact a state's national image, the NAEP assessments items will affect state assessments and, indirectly, how mathematics is taught in the school classroom. This item will contribute to the general disregard of the WYSIWYG characteristic of mathematics. We may add that this disregard seems to be shared equally by teachers, students, and school textbooks. At a time of disarray in school mathematics education, we cannot afford to add to the disarray by legitimizing the teaching of blatantly incorrect mathematics to impressionable minds.

The next two types of items have to do with definitions and conventions. They are not mathematically flawed in the literal sense, but their more-than-infrequent appearance in standardized assessments will certainly promote defective school mathematics education. We accept the fact that mathematics needs definitions and conventions, but we also know that these are means to an end rather than an end in themselves. If summative assessments are to assess mathematical learning truly, a slight nod in the direction of definitions and conventions already goes a long way. Unhappily, many states seem to insist on making such items a staple of their standardized assessment, thereby encouraging brute force memorization of facts in order to get some easy points. Note that we are not talking about the definitions of substantive concepts such as congruence, similarity,[2] average speed in a fixed time interval s to t, or division of fractions. The kind of definitions being tested on many state assessments are illustrated by the following items taken from the standardized assessments of two states:

Example 1. What is the inverse of the statement "If Mike did his homework, then he will pass this test"?

1. If Mike passes this test, then he did his homework.
2. If Mike does not pass this test, then he did not do his homework.
3. If Mike does not pass this test, then he only did half his homework.
4. If Mike did not do his homework, then he will not pass this test.

Example 2. Which property is illustrated by $2(2x + 4) = 4x + 8$?

A. Additive Identity
B. Distributive Property
C. Associative Property of Addition
D. Commutative Property of Addition

Example 3. Which equation illustrates the multiplicative identity element?

(1) $x + 0 = x$
(2) $x - x = 0$
(3) $x \cdot \frac{1}{x} = 1$
(4) $x \cdot 1 = x$

Instead of testing students on whether they know the terminology of the distributive law, would it not be more educational to assess, given four numbers a, b, c, and d satisfying $a + c = 4$ and $b + d = 7$, whether they recognize that $ba + cd + da + cb$ is equal to 28? And what is the purpose of learning what a "multiplicative identity element" is without any meaningful mathematical context? As to Example 1, perhaps the best commentary I can supply is to furnish what the state claims is the correct answer—it is (4)—because none of the few colleagues at Berkeley that I quizzed knew the answer.

A good example of the kind of convention too often honored by standardized tests is the so-called *Rules for the Order of Operations* taught around the sixth grade:

(1) Evaluate all expressions with exponents.
(2) Multiply and divide in order from left to right.
(3) Add and subtract in order from left to right.

These rules are of course inspired by the need in algebra to write polynomials with minimum notation. In

[2]We do not have in mind "same size and same shape" and "same shape but not necessarily the same size" for congruence and similarity, respectively. Rather, "a composition of translations, rotations, and reflections" and "a composition of a dilation and a congruence," respectively, as demanded by CCSSM.

mathematical terms, if we take into account that subtraction is *adding* an additive inverse and dividing is just *multiplying* a multiplicative inverse, then the rules amount to nothing more than: "Exponents first, then multiplications, then additions." (See Wu, 2004, for an extended discussion.) In mathematics, a convention is usually put in use when it contributes to clarity as well as brevity. Take the convention in algebra of denoting multiplication by a dot, e.g., $5 \cdot 12$ means 5 times 12. This is a good convention under most circumstances because it avoids the confusion of the letter x with the multiplication sign $\times$. But if by chance one has to multiply decimals, one would ignore the convention of writing, e.g., $3.1416 \cdot 20.4$, and write 3.1416×20.4 instead. With this in mind, we return to Order of Operations and consider the following item from a state assessment for sixth grade:

> The steps Quentin took to evaluate the expression $3m - 3 \div 3$ when $m = 8$ are shown below.
>
> $3m - 3 \div 3$ when $m = 8$
> $3 \times 8 = 24$
> $24 - 3 = 21$
> $21 \div 3 = 7$
>
> What should Quentin have done differently in order to evaluate the expression?
>
> A. divided $(24 - 3)$ by (24×3)
> B. divided $(24 - 3)$ by $(24 - 3)$
> C. subtracted $(3 \div 3)$ from 24
> D. subtracted 3 from $(24 \div 3)$

Faced with an item like this, the question is why would anyone write $3 \times 8 - 3 \div 3$ if all that is intended is $(3 \times 8) - (3 \times \frac{1}{3})$? It would be good mathematics education to ask teachers to explain to students that, for the purpose of communication, $(3 \times 8) - (3 \times \frac{1}{3})$ is to be preferred. Test items like the preceding one, from this point of view, do nothing but promote bad mathematics education. They inspire school textbooks to make heavy weather of the Rules of Order of Operations and related conventions. Even reference books on school mathematics take the hint and drill students on problems such as the following:

> Evaluate $4 + 5 \times 6 \div 10$.

(See Kaplan et al., 1998.) Such instruction time would be more profitably spent on explaining why, in a sixth grade classroom, one *should* always write $4 + (5 \times 6 \times \frac{1}{10})$ in place of $4 + 5 \times 6 \div 10$.

In 2004, I served on the Content Review Panel of CST (California Standards Tests) and in an April session, items on Rules of Order of Operations showed up. I suggested that such items should be used sparingly and I explained the reason at length. The explanation was apparently not convincing to the rest of the panel, and an entrepreneur from Silicon Valley who was on the panel then rallied the other members on the panel to pass the resolution that one or two such questions should be on *every* sixth grade CST. As a result of that meeting, I came to the realization that, until teachers, educators, and administrators become better informed mathematically, good mathematics education would never be a reality. That was the genesis of the article Wu, 2004, which I wrote essentially in one sitting after the meeting. That experience also illustrates very well why the active participation of high-level mathematicians is essential to maintaining mathematical integrity in standardized assessment and, therewith, also in school mathematics education itself.

Next, we take up a type of assessment items that are not flawed by present day school mathematics standards, i.e., what is called *Textbook School Mathematics* (TSM) in Wu 2011a and 2011b, but which should nevertheless be eliminated from standardized tests because they have a deleterious effect on the current effort to improve school mathematics education. There are major areas in the school curriculum that have been problematic for so long that any improvement would require substantive changes. For example, fractions, negative numbers, and geometry in middle and high school readily come to mind. In these areas, the CCSSM do a reasonable job, and if the curriculum is faithful to the CCSSM, improvement will likely follow, at least in theory. But there are also smaller areas such as percent, ratio, and rate in middle school in which student achievement has always been low. Two items in NAEP on rate and ratio that will be discussed next serve to remind us where things go wrong.

Let us begin with the concept of "speed," which is an example of "rate." While it is recognized that rate problems inspire apprehension and fear among students and teachers, it is not generally recognized that the concept of rate has never been defined in the K–12 literature. The latter tries to make rate into a single quantity that can be derived from a comparison-by-division of two different kinds of measures of a single situation, e.g., a time interval and the distance traveled in that interval. This is impossible in the middle grades when such problems arise of course, because rate is the derivative of a function that describes work in terms of time but calculus is not taught in middle school. Surprisingly, it seems never to have occurred to people in school education that *we should not insist on teaching a concept when we cannot explain what it is.* Recently, there was a major professional development impact study (National Center of Education Evaluation, 2011) on, among other things, students' ability to learn about rate, but because this basic contradiction was not recognized and dealt with from the beginning, the conclusion of the study was that professional development had no effect. (To be sure, there are other factors that contributed to this unsatisfactory conclusion, some of which are detailed in Wu, 2011b.) Therefore, any real improvement in students' learning of this topic would require a major rethinking.

There are some aspects of rate that can be taught in K–12, such as *average rate in a time interval* and *constant*

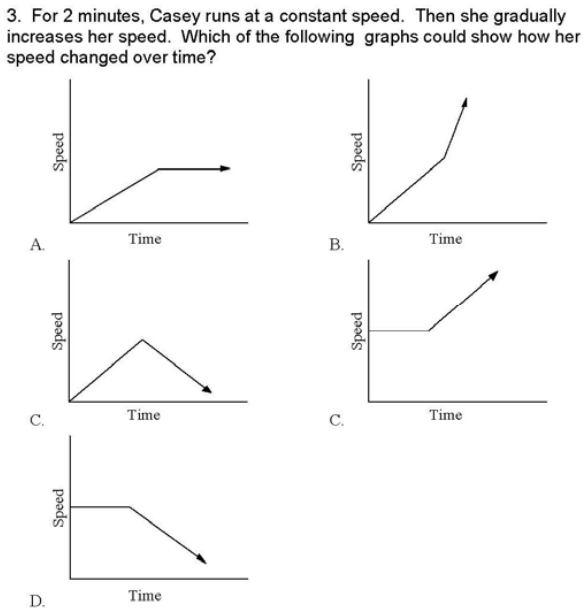


Figure 2.

rate, and a first step may be to teach these carefully to middle school students and leave the general concept to calculus. We are far from doing that because the concept of constant rate—the mainstay of all school rate problems—has not yet been given a precise definition in textbooks and the education literature. One possible definition is the following: a motion has **constant speed**, by definition, if its average rate over any time interval is a fixed number (see, for example, Wu, 2011c, Chapter 22). Constant rate *is* a concept that can be taught as a single quantity.

Given this background, the concept of "speed" has to be handled with extreme care in middle school. If it is felt that the general concept of "nonconstant speed" is too important to be left out of school instructions, then of course it can be discussed *intuitively* by way of examples to illustrate its complexity (e.g., Wu, 2011d, pp. 74–81). However, when taught this way, the intuitive concept of speed should not be part of *summative* assessment whose aim, of course, is to assess what has been learned. Intuition is fragile and we do not know yet how to assess it *directly*.

Now we come to the NAEP item on speed given in 2011 for grade 8 ("Identify a graph that shows how speed changed (calculator available)" in NAEP) (Figure 2). Here the use of the word "speed" in "she gradually increases her speed" is improper because speed is no longer constant in this instance. If we are serious about making a positive impact on the school mathematics culture, a NAEP item like this would do nothing but legitimize the mistaken belief among textbook publishers and many mathematics educators that "speed" is a precise *mathematical* concept in school education. That would be a step backward.

There is actually a way to change this item into a mathematically correct one with a little extra effort. Assuming that time is measured in hours, we can introduce an *ad hoc* definition of **average speed at time *t***, for any time t, as the *average speed in the time interval* $\left[t-\frac{1}{360}, t\right]$. That is, we measure the average speed of the motion in the 1 second interval before time t for each t. It can be argued very persuasively that this is almost as good as the intuitive notion of "instantaneous speed." In fact, some may even feel that this is a concept that should be taught in schools (of course, then it should be made clear that the unit of "1 second" can be changed to fit the occasion). Be that as it may, the NAEP item can now be amended to read:

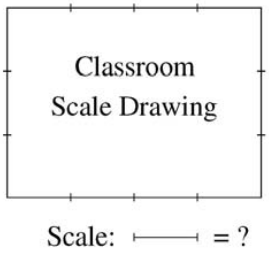
> For 2 minutes, Casey runs at a constant speed. Then she gradually increases her average speed. Which of the following graphs could show how her average speed changed over time?

And, of course, the word "speed" along the vertical axis in each graph will also have to be changed to "average speed." This is now a reasonable item to assess fourth graders' understanding of graphs.

Another concept that has been consistently abused in school mathematics is that of "scale drawing." Let us look at the NAEP item that originally inspired the present discussion. This is the item, "Determine scale used in drawing (calculator available)" in grade 4 of the 2011 NAEP assessment (NAEP) (Figure 3). It is understandable that there is no precise definition of "scale drawing" in grade 4 and fourth graders have to understand this item using their experiences with photographs or drawings they have seen. All the same, when Jackie talks about a "scale drawing of her classroom," a fourth grader would think of a room with height, length, and width, and with chairs and friends, whereas the given picture is a rectangle. The fourth grader now has to concentrate to see that it is only the *shape of Jackie's classroom floor* that this item is talking about! Therefore, would it not be better if the item is rephrased as follows?

> The picture shows Jackie's scale drawing of the shape of her classroom floor. Which scale did she use?

If the preceding discussion seems to be nothing more than nitpicking, let me point out that there is apparently no precise definition of what a "scale drawing" is in school textbooks. They talk about scale drawings of 3-dimensional objects in common-sensical terms and expect students to know how to do problems about these drawings. Unhappily they don't, and anyone who needs proof of this can consult *Windows on Teaching* (Merseth, 2003). In Case 3, a tenth grade class was asked to do a problem about the scale drawing of a staircase and it is



10. The picture shows Jackie's scale drawing of her classroom. Which scale did she use?

1. ⟷ = 1 inch.
2. ⟷ = 10 feet.
3. ⟷ = 100 feet.
4. ⟷ = 1 mile.

Figure 3.

clear that nobody could articulate what a "scale drawing" was. In fact, nobody seemed to be aware that if they didn't know what it was, they could not do the problem. This is another manifestation of the Wishful Thinking Syndrome on the part of teachers, textbook developers, and standards writers. Let us attempt a definition: Given a 3-dimensional object, imagine a *real-life size* photograph has been taken of the object and the pictorial image of the object is F.[3] Note that F is a 2-dimensional figure. A **scale drawing** of the object is a *2*-dimensional geometric figure F' similar to F. In particular, a scale drawing is a 2-dimensional figure similar to a *2-dimensional* figure and not the original 3-dimensional object. Of course this begs the question of what it means for two geometric figures to be *similar*. This is where the CCSSM excels as they explain the meaning of similarity, first intuitively in eighth grade and then precisely in high school geometry. Simply put, if F and F' are similar, then there is a one-to-one correspondence between the two figures and there is a fixed positive constant s (called the **scale factor**) so that if P and Q are points in F and P' and Q' are the corresponding points in F', respectively, the distance between P' and Q' is s times the distance between P and Q. For a fuller discussion of this issue, see for example Wu, 2012, pp. 44–48.

It would make a good topic for a pedagogical discussion in Methods classes to debate how much improvement in students' performance would have resulted in Case 3 of Merseth (2003) if the students had been given this definition.

For most students, it is far too complex cognitively to be able to think *precisely* about a 2-dimensional scaled-down version of a 3-dimensional object, to the point that they can solve mathematical problems. See Case 3 of Merseth (2003) again. So the preceding definition is not only necessary and correct, but it very likely would improve student achievement.[4] But naturally one would not want to burden a fourth grader with the concepts of "one-to-one correspondence" and "similarity." Therefore, for the NAEP item, the only option is to minimize the cognitive load as much as possible by being simple, clear, and precise. In this case, by saying that it is "Jackie's scale drawing of the shape of her classroom floor" seems to meet most of the requirements. This explains the apparent nitpicking.

For better or for worse, NAEP will have to function as an education institution in addition to being the nation's report card. Ideally, its assessments should do no harm (therefore, no mathematical flaws) and they should exemplify good mathematical practices rather than follow the existing defective ones. Given the impact the NAEP scores exert on the nation's school mathematics education, this is the responsibility for NAEP that comes with the territory and this is why NAEP ought to be more careful in phrasing items such as Jackie's scale drawing. Again, allow me to repeat the recommendation that there should be knowledgeable mathematicians assisting in every phase of assessment.

Above and beyond the technical details of individual test items, each high-stakes summative test sends an overall, *global* message about what is important in the curriculum, how school districts should write their next pacing guides, and how teachers should approach their lessons. But, as is well-known, it is very difficult to obtain complete test forms from state agencies (Massachusetts seems to be an exception), test vendors, or NAEP. For the purpose of this article, it is most fortunate that the Mathematics Assessment Project (MAP) offers two complete 3-hour summative test forms together with their scoring rubrics (MAP-PST1 and MAP-PST2). On the website MAP, it is stated that "The Mathematics Assessment Program (MAP) aims to bring to life the *Common Core State Standards* (CCSS) in a way that will help teachers and their students turn their aspirations for achieving them into classroom realities." Therefore, these test forms fall completely within the scope of this article. *In order to simplify the following discussion, we will treat these test forms as if they were standardized tests that had already been administered nationwide, but of course we have to remind ourselves from time to time that this is just make-believe.*

The format of both test forms is easy to describe: There are ten one-step or two-step Short Tasks on skills worth 10 points total, and there are 10 constructed response items worth 10 points each. Thus, the maximum score of each test is 110 points. Students have to write

[3]Think of some of the gigantic pictures of cars on billboards.

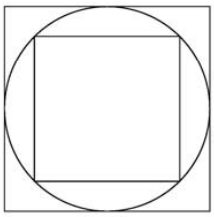
[4]The needed research for its validation should be relatively simple.

their answers to the Short Tasks, and one infers from the scoring rubric that no partial credit will be given. All but two of the constructed response items (Best Buy Tickets and Propane Tank, both from the first form) are broken up into smaller steps, presumably for easy grading.

These tests are a refreshing alternative to other standardized assessments that, to a large extent, consist of multiple choice items. In order to concentrate on the global message of these test forms, I will not mention the mathematical flaws of the individual items in the test forms or the rubrics except when it is absolutely necessary. I will also leave the Short Tasks alone in order to focus on the 20 constructed response items (10 in each test form).

These items do not ask for explanations of known facts in school mathematics such as "State and prove the Pythagorean Theorem." Rather, they assess students' ability to apply what they know to solve problems, and 12 out of the 20 are provided with a real-world context. Most likely, students are seeing these items for the first time. The level of difficulty of the items varies from moderate to high (such as the item Circle Pattern in the second form). For example, Figure 4 shows one of the more straightforward ones.

Circles and Squares This diagram shows a circle with one square inside and one square outside.



1. What is the ratio of the areas of the two squares? Show your work

_____

2. If a second circle is inscribed inside the smaller square, what is the ratio of the areas of the two circles? Explain your reasoning. _____

Figure 4.

Two statistics concerning these items are of interest. First, 63 out of the 100 points of the constructed items in the first test form MAP-PST1 are on the mathematics of grades K–8 according to the CCSSM, while the corresponding figure for the second test form MAP-PST2 is a bit better: 57 points out of 100.

For the record, here are the details of how these figures were arrived at. The following gives the number of points in each problem awarded to answers using mathematics in K–8:

> [MAP-PST1]: Multiplying cells 10, Patchwork 10, Square 10, Circles and Squares 10, Funsize Cans 10, Multiple Solutions 7, Propane Tanks 6.
>
> [MAP-PST2]: Leaky Faucet 10, Golden Crown 10, Birds' Eggs 10, Floor Pattern 5, Strawberry Boxes 10, Sidewalk Patterns 8, Circle Pattern 4.
>
> **Comments:** *Sometimes part of an item could be about high school mathematics according to the CCSSM, but if the rubrics indicate that answers using only eighth grade mathematics together with guess-and-check also get full credit, then that part of the item will be classified as K–8. A good example is Funsize Cans of the first test form and Strawberry Boxes of the second test form.*

A second statistic is that if a student only knows the following topics of the K–12 curriculum, then he or she will be able to score 106 points out of the 110 maximum in the first test form, and 96 points out of the 110 maximum in the second test form:

> the mathematics of grades K–8,
> the concept of a function,
> linear equations and functions,
> quadratic equations and functions (but *not* including the quadratic formula), and
> finite probability pertaining to standard permutations and combinations.

Precisely, that student would be able to do everything in the first test form except item 9 of the Short Tasks and part 2 of Multiple Solutions, and would be able to do everything in the second test form except items 1, 2, and 8 of the Short Tasks, Cubic Graphs (except (*ii*) and (*iii*) of part 2b), and part 2 of Floor Pattern.

I believe these tests—with their statistics above—will throw high school mathematics education out of kilter as there would be little incentive for many teachers to teach anything beyond factoring quadratic polynomials in high school. Specifically, teachers may be tempted not to teach any of the following high school topics in the CCSSM:

> all of high school geometry including the precise definition of similarity and the effect of similarity on length, area, and volume,
> finite geometric series,
> rational expressions,
> the arithmetic of complex numbers,
> the quadratic formula for real and complex quadratic polynomials,
> basic facts about exponential, logarithmic, and trigonometric functions, and
> high school statistics.

Such a temptation may be further strengthened by the general tendency in the scoring rubrics to downplay the mathematical sophistication that is appropriate for high school in favor of more elementary considerations such as

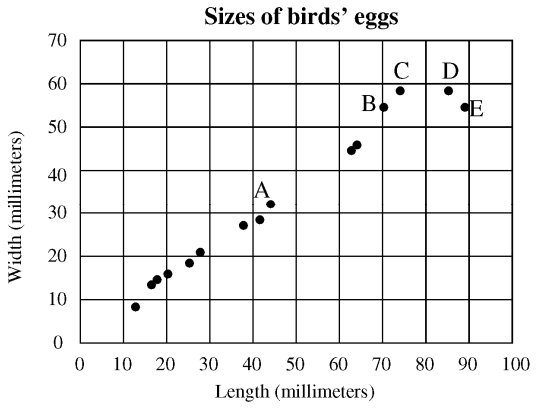


Figure 5.

guess-and-check. Perhaps two examples will suffice to make the point.

First, in the item Best Buy Tickets in the first test form, the problem boils down to comparing two linear functions $C_1 = \frac{2}{25}n$ and $C_2 = 10 + \frac{1}{25}n$: find the value n_0 at which the two functions are equal, decide which function is greater for $n < n_0$, and show that it becomes the smaller function for $n > n_0$. This is entirely routine middle school mathematics and the rubric duly outlines this solution. So far so good. *But*, in the same breath, the rubric also says:

May decide to solve arithmetically.

It then proceeds to outline a solution by evaluating in succession the values of C_1 and C_2 at $n = 50$, 100, 150, 200, 250, 300 and noting that the functions are equal at $n = 250$ and that their relative order of magnitude at the remaining five values is interchanged before and after 250. The rubric then states that if a student concludes that, on the basis of these six numerical values, C_1 is smaller than C_2 for $n < 250$ but bigger than C_2 of $n > 250$, that student will get full credit (10 points). In other words, an understanding of linear functions at the level of guess-and-check is considered to be good enough.

A second example is the item Birds' Eggs in the second test form. It presents a scatter plot that shows the lengths and the widths of the eggs of some American birds (Figure 5). Part 2 of this item asks: "What does the graph show about the connection between the lengths of birds' eggs and their widths?" As data go, these clearly show a linear relationship. The Statistics and Probability Standards 8.SP 1 and 2 address this situation explicitly and state, "Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line…" One would therefore expect that nothing less than a clear-cut statement about the approximate *linear dependence* of width on the length would merit full credit. Yet, according to the rubric, the expected answer is: "Gives a correct description such as: Generally, the greater the length of the egg, the greater is its width." In other words, a cubic dependence is a possibility. Such a rubric is hardly consistent with Practice Standard 6 of CCSSM, "Attend to Precision." Furthermore, part 5 of the same item asks: "Which of the eggs A, B, C, D, and E has the greatest ratio of length to width? Explain how you decided." Now the concept of slope is introduced in grade 8 (in a way that is far more careful and detailed than in other standards) so that by the time students take such a summative test, "slope" should be second nature. If one refers to the scatter plot above, one sees that the *ratio of length to width* at a data point is the reciprocal of the slope of the line joining (0, 0) to the point in question. Therefore the question becomes "which of the lines joining (0, 0) to A, B, C, D, and E has the smallest slope?" A visual inspection then shows that E is the answer to part 5. One would therefore expect that a correct answer to part 5 will give some indication of the reasoning in terms of slope. But the rubric says: "Gives a correct explanation such as: The line joining E to the origin is the flattest of all the lines joining A, B, C, D, and E to the origin." Given the fact that linearity did not even figure in the answer to part 2, why should such a statement about "flattest of all the lines" be taken at face value? Where is the "explanation"? How can such a summative assessment inspire teachers to teach high school mathematics in a way that is consistent with the CCSSM?

To continue the discussion of the global message of these test forms, I will introduce some standard terminology. There is a component of school mathematics consisting of the concepts and skills that make sense without any reference to everyday life, for example, the standard algorithms for whole numbers, the concept of estimation to a prescribed degree of accuracy (e.g., to the nearest ten), the definition of a fraction and its arithmetic operations, the geometric concepts of parallelism and perpendicularity, etc. I will refer to this body of knowledge as **pure mathematics**. There is another component that is concerned with the applications of these concepts and skills to problems that are given a real-world context such as estimating how much water is wasted by a leaky faucet each month, or how many cells there are in 85 minutes if you start with one cell and each cell splits into 2 exactly every 5 minutes. I will refer to this component as **applied mathematics**. Naturally, the line separating the two is not always clear-cut, but on the whole this terminology will be seen to be serviceable.

When school mathematics is taught properly, there is a healthy balance between pure and applied mathematics. The concepts and skills in pure mathematics will then be presented with reasoning and coherence; in learning pure mathematics, students will also pick up the ability to reason and to carry the reasoning to its logical conclusion. When they work in applied mathematics, *they use the same*

reasoning ability to get answers to problems, regardless of the fact that the problems now have a real-world context. It is predominantly—though not completely—the latter activity that is known in the mathematics education literature as **problem solving**. Because what one does in applied mathematics is nothing but a natural extension of what one does in pure mathematics, all of school mathematics—*if* taught properly—is a continuous process of problem solving. For example, consider the following problem:

> Given a right triangle with hypotenuse (of length) c and legs a and b, is there any relationship between a, b, and c?

The solution to this problem is what has come to be called the *Pythagorean Theorem and its proof*. It is in this context that the following quote from R. C. Buck's book on advanced calculus acquires special relevance (Buck, 1956, p. vi):

> We have tried to maintain a balance between theory and application. We share the view that applied mathematics may not exist—only applied mathematicians.

Unfortunately, what has dominated the school mathematics classrooms for the past several decades is what may be called the *bipolar approach to mathematics instruction:* Teach pure mathematics to students more or less by rote,[5] but make sure that when it comes to applied mathematics, we teach them how to solve problems.[6] One consequence of such instruction is that students may be able to use the tools they learn in pure mathematics in a superficial way, but their ability to reason with them is fragile because, not having been exposed to the reasoning that underlies these tools, they do not have a good model to follow when they try to reason on their own. Worse, many students do not even learn how to use the tools because they get used to learning them by rote and therefore reasoning is not part of their mathematical DNA. To make the situation more perilous, some people who are not familiar with the school classroom begin to emphasize the importance of "problem solving" in the sense of *problem solving in applied mathematics*. This emphasis leaves the rote-teaching in pure mathematics intact, thereby making it extremely difficult, if not impossible, to improve students' ability to solve problems in both pure and applied mathematics.

In order to prepare students better to learn how to solve problems in applied mathematics, it is quite plain that we must begin by teaching problem solving in pure mathematics better. At least, we must explain to students why the tools they are going to use are true. In mathematical parlance, we must explain to them the proofs of the basic theorems. Beyond this utilitarian motive for improving the teaching of pure mathematics, there is a cultural motive too. Pure mathematics is, after all, the crystallization of four millennia of the effort to understand, by use of reason alone, the world around us: not just the physical world but also the abstract one. It is also the foundation of science and technology. If a summative assessment is to assess what students have learned about mathematics, it should rightfully assess whether they understand a little bit of the basic structure of pure mathematics as well as some of its highlights. For example, what is the reasoning behind each of the following?

- The Pythagorean Theorem and its converse.
- The sum of (the lengths of) two sides of a triangle is bigger than the third.
- The zeros of a quadratic function are given by the Quadratic Formula.
- Triangles with equal base angles are isosceles, and conversely.
- The graph of $ax + by = c$ is a straight line, and conversely.
- The sum of the angles of (the degrees of) a triangle is 180.
- The summation formula for a finite geometric series.
- If a number c is the zero of a polynomial $p(x)$, then $(x - c)$ divides $p(x)$.
- A tangent to a circle is perpendicular to the radius at its point of tangency.

If we are going to create ten constructed response items for assessment, then one or two can afford to be about the proofs of (possibly mild variations of) basic theorems like these. In addition to such obvious items in high school mathematics, many more readily come to mind, for instance: Why does the same number π appear in both the area of a circle of radius r (i.e., πr^2) and its circumference (i.e., $2\pi r$)? Why is $(-1)(-1)$ equal to 1? Why does the long division algorithm yield the quotient and remainder of a division?[7] Why does the long division of the numerator of a fraction by the denominator yield a decimal equal to the fraction?[8]

It is time that we return to the discussion of the possible impact the MAP test forms could have in the school classroom. It was mentioned from the beginning that these test forms do not assess students' knowledge of known theorems in pure mathematics. Rather, they assess whether students can apply these theorems to solve problems. The global message of these test forms is therefore to endorse *bipolar instruction*. Teachers will be

[5]To make a long story short, low-quality textbooks are mainly responsible for this crime.

[6]Only those who hold fast onto the Wishful Thinking Syndrome would believe that this kind of learning is possible.

[7]See Wu, 2011c, Chapter 7.

[8]See Wu, 2011c, Chapter 42.

encouraged to continue teaching *how to use the known tools* but not so much *why these tools are true*.

Perhaps a bit more flexibility in the construction of the test forms would lead to a more balanced summative test. Consider, for example, the item called Square in the first test form that asks for the verification that four given points (with integer coordinates) in the coordinate plane are the vertices of a square. This assesses, among other things, whether students know how to make use of the fact that two lines are perpendicular if the product of their slopes is −1. Would it not make sense to assess instead whether they know *why* two lines are perpendicular if the product of their slopes is −1? Making students and teachers know that they are accountable for the reasoning in the pure mathematics of the curriculum would be making a good first step in breaking the stranglehold of bipolar mathematics instruction on the nation's classrooms.

I am aware of the concerns about how students would be encouraged to memorize proofs; undoubtedly some will do just that, but many will achieve some understanding through the process of trying to repeat a proof. The net result will be a worthwhile trade-off if the alternative is that they know *no* proof at all.

We have already taken note of the fact that the two test forms MAP-PST1 and MAP-PST2 stand in stark contrast to almost any other standardized summative test by virtue of not having any multiple choice items. Most standardized summative tests used to be entirely in the multiple choice format, and the infusion of a few constructed response items is a phenomenon of recent vintage. The move toward a better assessment of mathematics learning is unmistakable, but one should be realistic about how far this move can go. In order to sharpen our focus, I will concentrate on high-stakes tests in the remainder of this article.

There should be no mystery as to why the multiple choice format—or something similar such as the short-answer-with-no-partial-credit format in the Short Tasks of the two MAP test forms—is favored in high-stakes standardized tests. A dream high-stakes test is one that is easy and inexpensive to grade and yet can accurately assess student achievement. The multiple choice format satisfies at least the first two conditions. In particular, it guarantees a speedy feedback. Administrators need speedy feedback to make timely decisions about students and, in some cases, even schools. Should a student be required to take remedial courses in summer school? Grade repetition is not helpful and leads to dropping out. Do the aggregate scores of a school remain flat three years in a row, triggering some sort of intervention? The decision ought to be made long before the new school year begins. And so on. But there is an inherent conflict between the need for speedy response and the ability to assess student achievement accurately: the nature of mathematics stands in the way. On the one hand, excellence in mathematics at any level is predicated on the ability to sustain sequential thinking in order to carry a logical argument to its conclusion.[9] In plain language, can a student negotiate the inevitable twists and turns that one sometimes must make, without any prompts, in order to go from hypothesis to conclusion? This is what problem solving is all about, but test items that assess this ability will not be easy or inexpensive to grade. On the other hand, a speedy response requires that high-stakes tests be easily gradable, which rules out test items designed to assess sustained sequential thinking. Because the need for speedy response is paramount, we have to accept the fact that high-stakes standardized tests are fundamentally too crude to measure different levels of excellence in mathematics achievement. In general, policy makers, educators, administrators, and the general public seem unaware of this built-in limitation of high-stakes tests, and thus do not realize that they should interpret such test scores with caution.

Well-chosen constructed response items can assess sequential thinking, of course, but their presence on high-stakes tests is severely limited by the need to make the tests easily and cheaply gradable. Consequently, such constructed response items are constrained to be not too hard, and they are also usually broken down into smaller pieces so that each piece becomes easier to grade. This defeats the very purpose of assessing sequential thinking. But even the best constructed response items are further compromised by the grading process itself. Because thousands of people will be involved in grading, the responsibility for enforcing a uniform grading standard falls on the instructions in the scoring rubric. In case anyone believes that there is no hardship in following a scoring rubric, let me relate my personal experience as a calculus instructor for some forty years.

I have taught at least forty calculus courses in my career at Berkeley. They are part of a two-year calculus sequence for freshmen and sophomores. These are always large lectures, with between 200 and 300 students in my days, and three times a semester I would be grading exam papers with my six to ten teaching assistants. My exams were on the whole similar to the MAP test forms, in that they were always three-hour tests and, typically, about 30% of the test items were short-answer-with-no-partial-credit ones, and there would be five or six constructed response items. Some of the latter would consist of a single question, i.e., not broken up into smaller steps, so that I could directly assess whether students could think through a problem without prompts. In other words, were they capable of sequential thinking? In the grading sessions, I gave out scoring rubrics. Since my teaching assistants could ask me about any ambiguities in the rubric, there could be no misunderstanding about my intentions.

[9]One can argue that originality is even more important, but in K–12, it would be improper to make originality part of the regular assessment.

Nevertheless, those sessions were always teeth-grinding occasions because it was impossible to anticipate all the erratic writing styles of the students. Many times, we had to make guesses as to what a disconnected collection of symbols was trying to tell us. How many points did they deserve? Other times, students' thinking, as reflected on the exam papers, was definitely not sequential (i.e., neither logical nor coherent) even if the answer was correct. We had to agree on an *ad hoc* rubric each time it happened. All things considered, was I confident that the collective judgment on partial credits—made on the fly, as it were—was always accurate? I would hesitate to give an affirmative answer.

Keep in mind that my students were college students taking some form of calculus in their first two years[10] and would be therefore at least in the first quartile of high school graduates in terms of mathematics achievement. If grading their papers was often a gut-wrenching experience, what does it say about grading constructed response items of standardized tests, *statewide*, by people who may not be as mathematically well-informed as the teaching assistants at Berkeley, and who most likely would not have direct access to the author of the scoring rubric? On this basis, we have to conclude that making a refined judgment about student achievement is simply not within the capability of such a blunt instrument.

The same statement can be made about the MAP test forms, except that when almost the whole test consists of constructed response items, then some additional comments are necessary. When the number of constructed response items increases, the uncertainties mentioned in the preceding two paragraphs multiply and the connection between students' test scores and their actual achievement becomes even more tenuous. In this light, the perceived laxity in the scoring rubrics mentioned in connection with the Best Buy Tickets and Birds' Eggs items may very well be the product of a conscious decision to minimize this uncertainty by (artificially) expanding the meaning of a "correct answer." But if this is the case, then one can only conclude that there is a price to pay either way for having many constructed response items on a large scale standardized test: at the end, one is no longer certain whether the score can faithfully reflect the achievement. In addition, the fact that only two out of the twenty items in the MAP test forms are not broken down into smaller steps means that even these test forms are limited in their ability to assess sequential thinking. There is a third concern: By devoting over 90% of the test to the *applications* of tools in pure mathematics to solve new (to them) problems, these test forms fail to assess students' knowledge of the tools themselves and, as such, do not constitute an even-handed assessment of what students really know. A colleague of mine, James Sethian, once told me about his attitude toward exams: "An exam is not the time to find out how clever my students are. I only want to find out if they learned anything." I completely agree with him. Solving 10 new problems in something like 160 minutes requires ingenuity-on-demand (unless the problems are uniformly easy, and those in the MAP test forms are not). For most of us, ingenuity comes and goes in a time of stress. Therefore a high-stakes test that overemphasizes ingenuity may not be an accurate assessment of what students have learned. That was why I tried earlier to make the case that there should be some items on the proofs of major (known) theorems because these are at least things they *should* know, and are expected to know.

In summary, these many imponderables prevent the scores from the MAP test forms from being true indicators of different shades of achievement.

To my knowledge, lawyers do not advertise their Bar Exam scores to show they are good lawyers, and nor do physicians advertise Medical Licensing Exam scores. To enter these professions, the important thing is to make sure that certain basic requirements are met rather than to worry about *how well* they are met. One can speculate that they know only too well that one exam cannot define excellence and therefore these professional licensure exams concentrate entirely on setting the bar high enough to ensure basic competence. They stay within the limitations of an exam.

Having discussed at length why high-stakes standardized tests in K–12 mathematics, on theoretical grounds, must fail to measure achievement accurately, now I want to reverse course and advocate that we adopt the legal and medical models and concentrate on making high-stakes tests a reliable measurement of basic competence in mathematics. It should be quite clear that assessing basic competence is much easier than assessing excellence. For one thing, we no longer need to be preoccupied with *sustained* sequential logical thinking; some indication of the ability to do sequential thinking would already serve the purpose. One can even try to find out whether or not a completely multiple choice format, consisting of many multi-step problems using the sophisticated model of the Japanese University Entrance Examinations (L.-E. E. T. Wu, 1993), would already suffice for this purpose. In fact, I predict that, with the active participation of knowledgeable mathematicians, we can very quickly achieve a broad consensus on *the minimum basic skills and concepts* that a high school graduate must know. Once that is done, we will be able to make explicit the new goal of high-stakes tests: like the legal and medical tests, high-stakes tests will henceforth measure basic competence and nothing more. Let us see how this approach to assessment can bring clarity to the discussion of high-stakes tests in mathematics education.

First, it is now imperative to eliminate from all high-stakes tests the kind of mathematical flaws in summative assessments that were discussed earlier in this article. If

[10]Most skipped the first course of the two-year calculus sequence and took the second one when they came in.

there is any doubt about the urgency of this action, imagine that the Bar Exam or the Medical Licensing Exam were infested with similar flaws. What would the reaction be?

Second, we will be able to explain to policy makers and the general public that a high-stakes test is no longer about different levels of achievement. Only basic competence. They will learn to think of it as the driver's license test in math education: it pass it and forget it.

The analogy with driver's license tests bears a closer scrutiny. We all know that passing both parts of the driver's license test, the written test and the performance test, is very far from making all drivers safe drivers. The performance test does not include driving on the freeway, for example. But given the scale of the test, perhaps its modest scope is all that is possible. So we legalize the licensing system without pretending that a licensed driver is a good driver. Because this message seems to be clearly understood, nobody is under any illusion that getting 100 on both tests immediately makes that person an excellent driver, and no city is known to brag about its traffic safety because its people have the highest average grade in the driver's license test. Now suppose we can get across this message about the new high-stakes tests to schools and school districts. Then they know that it is part of their obligation to help all students pass this test for basic competence; they also know that this is an achievable goal in that they do not need to worry about teaching fancy tricks or complex problem-solving skills, only the most basic concepts and skills. Finally, they *can* teach to the test and not have to feel guilty about it. At the same time, they also know how to look at this test: pass it and forget it. In order to achieve excellence in teaching, they will have to get it done in their own classrooms and not by coaching their students to get high scores on those tests.[11] Furthermore, this realignment of high-stakes tests will help clarify one aspect of the use of value-added models to rate teacher quality. Briefly, some people have begun to rate teachers' effectiveness by measuring the growth (or decline) over time of the high-stakes test scores of all the students taught by a particular teacher (see Ewing, 2011, for example). The Los Angeles Times, infamously, published a value-added rating of all the elementary school teachers in the Los Angeles Unified School District, first in 2010 and then with an update in 2011 (Song & Felch, 2011). This rating makes many false assumptions, not the least of which is that high-stakes test scores accurately measure different levels of excellence in students' achievement, a misconception that we have spent considerable time to combat. If all policy makers could grasp the nature of high-stakes tests—to the effect that they are incapable of finely measuring achievement—then maybe students' scores would not be used for this purpose in the future.

Last but not least, we can now stand up to the critics who want to abolish all high-stakes tests altogether. We need a reasonable way to maintain quality control in schools, in the same way that we need a driver's license test to ensure a certain amount of traffic safety no matter how inadequate that test may be. Far too many high schools still graduate students who can barely read or do arithmetic fluently. The original intent of the No Child Left Behind legislation was precisely to eradicate such abuse, but the legislation was too complex and tried to accomplish too much. With a basic competency test, we can at last accurately identify those students and schools most in need of assistance. This is a genuine equity issue in education that any responsible person should support. In the context of mathematics education, there is at present a real discontent about how the high-stakes tests mandated by No Child Left Behind drive out genuine student learning, and this discontent is what provides fuel to those who want to abolish all high-stakes tests. The shift to basic competence should put to rest this radical advocacy.

It is human nature to try to oversimplify everything to a number. Think of IQ as a measure of intelligence, for example, and we can understand the love affair of the education establishment with high-stakes test scores. It took some time for people to realize that maybe IQ wasn't quite what it was cracked up to be. There is a society, *Mensa International*, whose members all have IQ's at or above the 98th percentile, but (fortunately) not a single Nobel laureate has been a member of Mensa. It is possible that now is the time for people in education to acquire the proper perspective on high-stakes test scores and not exaggerate their importance out of proportion. Assessment will only fulfill its promise of being an aid to education when it is but one component in a complete system, including sound curriculum, rigorous standards (which the CCSSM are), excellent textbooks, and high-quality teacher education. Let us hope that, with this recognition and redesign, high-stakes tests can contribute positively to mathematics education.

Acknowledgements

The author is grateful to David Collins for pointing out some typos and Larry Francis for a long list of corrections. He wants especially to thank Lisa Hansel for her many perceptive and profound suggestions.

Notes

Correspondence concerning this article should be addressed to Hung-Hsi Wu, Department of Mathematics #3840, University of California, Berkeley, CA 94720-3840. E-mail: wu@math.berkeley.edu.

[11] I am under no illusion about how difficult, and even unrealistic it is to achieve excellence within the classroom. Think about professional development, for starters. But a goal without pretension is always better than one that is full of it.

References

Buck, R. C. (1956). *Advanced Calculus.* New York, Toronto, London. McGraw-Hill Book Company.

Common Core State Standards for Mathematics. (2010). http://www.corestandards.org/the-standards/mathematics

Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007). *Validity Study of the NAEP mathematics Assessment: Grade 4 and 8.* Washington, DC: National Center for Education Statistics and American Institute of Research.

Ewing, J. (2011). Mathematical intimidation: Driven by data. Notices Amer. Math. Soc. 58, 667–673.

Kaplan, A., Debold, C., Rogalski, S., & Bourdreau, P. (1998). *Math on Call*. Wilmington, MA: Great Source Education Group.

Mathematics Assessment Project (Undated). http://map.mathshell.org/materials/background.php

MAP-PST1. Prototype Summative Tests: CCR-C1, Form and Rubric. http://map.mathshell.org/materials/tests.php

MAP-PST2. Prototype Summative Tests: CCR-C2, Form and Rubric. http://map.mathshell.org/materials/tests.php

Merseth, K. K. (2003). *Windows on Teaching Math.* New York, NY: Teachers College Press.

Milgram, R. J. (Undated) The discussion of well-posed problems. ftp://math.stanford.edu/pub/papers/milgram/The-discussion-of-well-posed-problems.pdf

National Assessment of Educational Progress (NAEP). NAEP Questions Tool: Mathematics. http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=mathematics

National Center for Education Evaluation (2011). Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation. Washington, DC: US Department of Education.

The Mathematics Advisory Panel (2008). *Foundations for Success: Reports of the Task Groups and Sub-Committees*. US Department of Education, Washington, DC. http://www.ed.gov/about/bdscomm/list/mathpanel/reports.html

Partnership for Assessment of Readiness for College and Careers (Undated). http://www.parcconline.org/parcc-assessment

SMARTER Balanced Assessment Consortium (Undated). http://www.k12.wa.us/smarter/

Song, J. & Felch, J. (2011). Times updates and expands value-added ratings for Los Angeles elementary school teachers. *Los Angeles Times*. May 7, 2011. http://www.latimes.com/news/local/la-me-value-added-20110508,0,930050.story

Wu, L.-E. E. T. (1993). *Japanese University Entrance Examination. Problems in Mathematics.* Washington, DC: Mathematical Association of America.

Wu, H. (2004). "Order of operations" and other oddities in school mathematics. http://math.berkeley.edu/~wu/order5.pdf

Wu, H. (2011a). Bringing the Common Core State Mathematics Standards to Life. *American Educator*, Fall 2011, Vol. 35, No. 3, pp. 3–13. http://www.aft.org/pdfs/americaneducator/fall2011/Wu.pdf

Wu, H. (2011b). Professional development and *Textbook* school mathematics. http://math.berkeley.edu/~wu/AMS_COE_2011.pdf

Wu, H. (2011c). Understanding Numbers in Elementary School Mathematics. Providence, RI: American Mathematical Society.

Wu, H. (2011d). Teaching Fractions According to the Common Core Standards. http://math.berkeley.edu/~wu/CCSS-Fractions.pdf

Wu, H. (2012). Teaching Geometry According to the Common Core Standards. http://math.berkeley.edu/~wu/Progressions_Geometry.pdf