PART II

Teachers and Teaching

4

ASSESSING TEACHERS' MATHEMATICAL KNOWLEDGE

What Knowledge Matters and What Evidence Counts?

Heather C. Hill, Laurie Sleep, Jennifer M. Lewis, and Deborah Loewenberg Ball

UNIVERSITY OF MICHIGAN

For more than two hundred years, teachers—and what they know—have been objects of scrutiny. Teachers have been tested, studied, analyzed, lauded, and criticized. In short, both they and their performance have been assessed to an extent rare in other professions. Both the purposes and methods for assessing teachers have varied. With some assessments, the goal has been to evaluate individuals' qualifications for the work of teaching, while with others, the focus has been to analyze the knowledge teachers use to do that work. Most assessments have been evaluative, either explicitly or implicitly, seeking to appraise the adequacy of individual teachers' knowledge or the quality of their performance. Some assessments have contributed to building evidence about the knowledge needed for teaching, thus helping to establish criteria for professional qualifications and the methods for certifying them. What is assessed differs across approaches, and how it is assessed varies: Some are indistinguishable from a test that could be given to students, while others pose tasks special to the work of teaching. Teachers have been interviewed and observed; they have been queried, given tasks, and asked to construct portfolios representing their work. The variation in purposes, test content, and assessment methods has increased over the past 30 years, as teacher testing has become more commonplace and the number of teacher tests has multiplied.

Developing a more coherent approach to the assessment of teachers' knowledge, and in particular their knowledge of mathematics, is now both possible and necessary. Three important contemporary pressures call out for a system of assessment that is not only rigorous, but professionally relevant and broadly credible. The first is a political environment that demands that students be taught by "highly qualified" teachers. Too many students, especially those in under-resourced schools, currently are assigned to teachers who are not certified to teach mathematics (Darling-Hammond & Sykes, 2003; Ingersoll, 1999). Still, what counts as "qualified" and how this could be measured and certified remains a significant problem. A second pressure is the growing need to establish evidence on the effects of teacher education on teachers' capacity, and of teachers' knowledge and skill on their

students' learning. Many question the effectiveness of professional training and argue that it is unnecessary; by this logic, all that is needed to be qualified to teach is a mathematics major and experience. Evaluating this claim requires empirical evidence about teachers' knowledge and skill. This, in turn, necessitates a method for appraising such capacities. Third is the need to distinguish what makes teaching professional—that is, a domain of professional knowledge and skill not possessed by just any educated adult. Is the knowledge mastered by someone who majors in mathematics sufficient content knowledge for teaching? If not, then what exactly distinguishes mathematical knowledge for teaching and how could this distinction be established? Given the confluence of these three forces, the field has an opportunity to advance both the tools to assess teachers' mathematical knowledge and our collective understanding of the notion of mathematical knowledge *for teaching*. A set of agreed-upon reliable and valid methods of assessing teachers' mathematical knowledge would afford the capacity to gather and analyze the sorts of evidence needed to make progress on these issues.

Toward that end, this chapter reviews and seeks to appraise the myriad ways in which U.S. teachers' knowledge of mathematics has been assessed. It asks: What is measured on tests of teachers' mathematical knowledge? What *should* be measured? *How* should it be measured? And how has the evolution of both assessment methods and scholarly thinking about teachers' mathematical knowledge influenced these tests? Answering these questions has taken us into territories that have often remained disconnected. On one hand, the history of teacher testing consists largely of assessments designed to *certify* teachers, that is, to attest to the adequacy of teachers' knowledge to teach mathematics. On the other hand, the history of research on teaching and teacher education reveals a stream of systematic efforts to *investigate* what teachers know, and to associate that knowledge with their professional training and their instructional effectiveness. Although these two lines of work—certification and investigation—have both focused on the assessment of teachers' mathematical knowledge, they have proceeded almost independently of one another. They differ in the knowledge they seek to assess, the methods they use to do so, and the conclusions they aim to make. In this chapter, we assemble these lines of work under a common light in order to consider how teachers' knowledge might be responsibly assessed. Important to note, however, is that the assessment of teachers is hotly contested terrain. Our goal is to contribute resources that might help move the debate from one of argument and opinion to one of professional responsibility and evidence.

The chapter is organized in four parts. First we review the history of U.S. teacher certification testing, beginning in the nineteenth century, and trace its evolution and re-emergence in the 1980s. In the second section, we examine early *scholarly* research involving the measurement of teachers' mathematical knowledge. Both early certification exams and scholarly studies were constructed in the absence of any elaborated theory about the elements of mathematical knowledge for teaching, a fact that led to important limitations on the interpretations of this early work. We then show, through extended example, that in contrast to the views of teaching mathematics implicit in both early teacher certification tests and early scholarly research, teaching requires knowing more than simply how to solve the problems a student might solve. In the third section, we describe measurement methods in studies that explore the mathematical knowledge as it is *used* in teaching. This section provides a review of the tools and instruments developed to study teachers' mathematical knowledge. In the fourth section, we return once again to teacher certification exams, considering the extent to which modern exams have taken up the ideas and methods that appear in scholars' study of mathematical knowledge for teaching.

Throughout, we observe that the mathematical assessments used in each period are linked both to contemporary methods of assessment and ideas about the mathematical knowledge teachers need for success with students. We do not explore ideas about teachers' mathematical knowledge in their own right, but instead describe how these ideas have been instantiated in and influenced the development of assessments.[1] We also limit our analysis to U.S. teacher assessments, as reviewing international teacher assessments is beyond the scope of a single chapter.

TESTING TEACHERS: HISTORICAL APPROACHES TO ASSESSING TEACHERS' MATHEMATICAL KNOWLEDGE

The history of assessing teachers' mathematical knowledge begins with the history of teacher examinations for certification. For the majority of U.S. history, in fact, the main route to teaching was through

[1] For a fuller treatment of mathematical knowledge for teaching, see Ball, Lubienski, and Mewborn (2001).

taking an exam, usually offered at the local or county level, that certified those with passing scores as eligible for classroom work (Angus, 2001; Haney, Madaus, & Kreitzer, 1987). It was only at the turn of the twentieth century that the completion of professional education programs eclipsed certification exams as a pathway into teaching, though by the late twentieth century, certification exams had returned full force. In this section, we examine the earliest exams, describe their decline and subsequent rebirth in the 1980s. In doing so, we focus closely on the content of these assessments in order to examine the mathematical knowledge thought to be required for teaching.

Early teaching exams were notable for what they included—not only the three "R's", but also questions designed to assess the moral fiber, ability to manage a classroom, and, possibly, religious affiliations of those tested (Angus, 2001). Over the course of the nineteenth century, exams lengthened to include a wider array of subject matter and, by the 1850s, some pedagogical questions as well. This period was also marked by a change in exam format, from an oral exam often administered by a local district board to written exams administered by county or state officials.

The mathematics sections of these written assessments reflected closely the curriculum of the day. The state of Michigan, for instance, maintained a vigorous teacher examination system through the early twentieth century, issuing tests in physics, literature, geography, arithmetic, history, civics, botany, physiology, and the "theory and art" of teaching, among other topics. An 1895 exam[2] contained the following arithmetic problems:

1. A pole 63 feet long was broken in two unequal pieces, and 3/5 of the longer piece equaled 3/4 of the shorter. What was the length of each piece? Give a good solution.
2. Two men hire a pasture for $20. The one puts in 9 horses, and the other puts in 48 sheep. If 18 sheep eat as much as three horses, what must each man pay?
3. A boat whose rate of sailing in still water is 14 miles an hour, was accelerated 3 1/2 miles per hour in going down stream, and retarded the same distance per hour in coming up. It was five hours longer in coming up a certain distance than in going down. What was the distance?
4. (a) How do you read any decimal? (b) How do you express decimally any common fraction?

(Michigan Department of Public Instruction, 1896, pp. 297–298)

The state exam also contained the following algebra problems:

1. There is a number consisting of three digits, the first of which is to the second as the second is to the third; the number itself is to the sum of its digits as 124 to 7; and if 594 be added to the number, the digits will be reversed. What is the number?
2. Find the value of x:

$$\frac{\sqrt{3x+1}+3}{\sqrt{3x+1}-3}=\frac{\sqrt{7x+8}+4\frac{4}{5}}{\sqrt{7x+8}-4\frac{4}{5}}$$

(Michigan Department of Public Instruction, 1896, p. 297)

Candidates answering these and other problems were granted a state certificate for teaching any grade. Michigan also maintained, through this period, a separate county examination system in which graded certificates were available to candidates in rural areas. This system reflected the tradition of local control over such exams, but more important, also accommodated the realities of the rural teacher labor market. Candidates in rural areas could choose to sit for an exam leading to a general certificate, which authorized its holder to teach at any grade level in the county. These general certification exams were substantially similar in length and content to the state exam excerpted above. Records show that the vast majority of rural teacher candidates, however, chose instead to sit for exams leading to primary (K–4) certificates. These exams did not have a formal algebra section, and occasionally had less computationally intensive arithmetic problems. County education commissioners and legislators argued that requiring more difficult tests for the rural population would have led to fewer certified teachers, a chronic problem in underpopulated counties. In fact, even the primary-grade exams boasted a 50% failure rate in some counties (Michigan Department of Public Instruction, 1897, p. 152). One observer noted that, "The chief cause for failure is a want of knowledge of the subject; and this is due to poor teaching" (Michigan Department of Public Instruction, 1897, p. 153).

Beyond illustrating how mathematics was imagined to be used in daily life at the turn of the twentieth century, these exams and their attendant materials reveal the views of teacher knowledge held during this period. To start, many believed teachers should know more difficult mathematics than they were expected to teach. On the typical primary certification exam, for instance, candidates were asked to solve classic middle school arithmetic problems—rates, proportions, and percents:

[2] Throughout the chapter, the item numbering is our own, and does not necessarily reflect the item numbering on the actual tests.

1. (a) Explain as to a class the difference between the simple and local value of figures.

   (b) Write in figures fourteen million, one thousand and five hundredths; three trillion, two hundred one and one thousand seventy billionths; also write in words 7504306.040521/4.

2. I own a horse and a farm; one-fourth the value of the farm is four times the value of the horse. Both taken together are worth $1,700. Find the value of each. Write out a complete analysis.

3. A man sold a lot for $84 and by so doing gained 1/5 of what it cost. What % would he have gained if he had sold for $100? Analyze.

4. A field of 5 acres in form of a square is to be surrounded by a fence 4 1/2 feet high, to be built of boards 6 inches wide, placed horizontally. The lower board is to be four inches above the ground, and there is to be a space of 5 inches between the boards. What will be the cost of the boards required at $18 per M?

5. A merchant gets 500 barrels of flour insured for 75% of its cost, at 2 1/2%, paying $80.85 premium. For how much per barrel must he sell the flour to make 20% upon cost price?

6. A man devotes 40% of his income for household purposes, 35% for the education of his children, 16% of the remainder to charitable purposes, and saves the remainder, which is $705.60. What is his income?

7. One-half of a stack of hay will keep a cow for 20 weeks, and 3/4 of the stack will keep a horse 120 days. How many weeks will the whole stack of hay keep both the cow and the horse?

(Michigan Department of Public Instruction, 1896, p. 315)

Although it is possible that these topics were taught in the primary grades during this era, this is probably not why this content was included. In their annual reports, state superintendents of the time asserted that the "idea that a teacher should know very much more than he is expected to teach is so generally agreed upon that no discussion here is necessary" (Michigan Department of Public Instruction, 1898, p. 10). Mathematics certification exams throughout this period were also marked by the use of numerically complex examples. Whereas the examiners could have chosen mathematics problems that yielded easily to mental computation or shortcuts, few did. This implies that successful teacher candidates of this time would not only understand the general principles for solving such problems, but also have a solid foundation in facts, procedures, and patience.

According to state documents, the difficulty of these exams was in part a political move, the state bureaucracy's attempt to ratchet up the quality of rural teachers by making the test more difficult and removing the possibility of political favoritism. Defending an increase in the difficulty of test questions, the Superintendent of Public Instruction Henry R. Pattengill wrote in 1896: "The rural schools have received the benefit of a better prepared, more mature, and more broadly educated class of teachers. . . . Sympathy, politics, sect, 'pull', should play no part in the choice of teacher. It is not understood that the teacher need be examined year after year in the same studies, but every teacher should forever be a student" (p. 7). Another state report of the time presaged current political slogans: "Better qualified teachers in the country schools should be our motto" (Michigan Department of Public Instruction, 1897, p. 10).

The exams and their materials reveal other key assumptions about teacher knowledge. In Michigan, nearly all the arithmetic problems were what would today be called "real-world," ones that asked teachers to engage in the practical calculations facing businesses, farms, and banks. Problems of a purely mathematical nature—for instance, those that explored the conceptual underpinnings of arithmetic—were not included. The problems included were based on the mathematics curriculum that teachers were expected to transmit to their all-male students (Michalowicz & Howard, 2003), and suggested that test writers viewed teachers' mathematical knowledge as the knowledge they were responsible for instilling in future business owners, farmers, and bankers.

And, despite the applied focus of the exams, some Michigan state superintendents advocated what might today be called conceptual knowledge. In 1896, for instance, graders were cautioned that "in arithmetic, a knowledge of principles and general accuracy in method shall be considered not less than three times as important as obtaining a correct answer" (Michigan Department of Public Instruction, 1896, p. 313). At an 1896 conference on teacher testing, one county commissioner stated, "Our examinations should be of such a character as to demand less of the memory and more of the reason; less rote work, more sequence, more analysis, and more simplicity" (Michigan Department of Public Instruction, 1897, p. 409). Without evidence from the actual grading of exams, it is difficult to determine the extent to which this sentiment was implemented. However, the existence of these admonitions foreshadows the debates in the twentieth century over procedural fluency and conceptual understanding.

Finally, these tests also illuminate nascent political struggles about what should be known in order to teach. These struggles continue into the present; in fact, the words Pattengill wrote in 1894 could well have been written today:

> The question of teachers' examinations has always been a perplexing one. No matter how excellent the questions for such examinations may be, nor how thoroughly and honestly conducted, it will still

> be true that examinations alone are an inadequate test of a candidate's ability to teach school. The qualifications requisite for good school teaching cannot be ascertained by any set of questions whether oral or written; and yet, where the teachers must be employed for a large extent of territory and for many schools, no other plan has been suggested better than the plan of examinations combined by the supervision exercised by the country school commissioner.
>
> The first essential element in a teacher is good scholarship. No amount of tact, or method, or skill in the use of devices will make up for the deficiency in scholarship. In thus emphasizing the value of this factor, we by no means overlook the value of tact and good method. If, however, we are to do without one of these, we think we could more safely risk the teacher with scholarship than the one with method and poor scholarship. (Michigan Department of Public Instruction, 1894, pp. 1–2)

In Pattengill's view, and in the exams of the early 1890s, content knowledge was separate from knowledge of the "theory and art" of teaching. Which is more important to student learning remains both an unanswered question and a topic of hot debate.

In contrast to the business-focused mathematics on the exams of the 1890s, the period after Pattengill's tenure saw an increase in teaching-specific mathematics questions on these exams. For instance, while half the 1900 state exam contained problems similar to those found in previous years, the arithmetic section also carried the following:

1. In what arithmetical processes is drill the essential element? In what operation is memory a factor?
2. When may the algebraic equation be used to advantage in arithmetic?
   Is it advisable to teach its transformations in eighth grade arithmetic?
3. How do you present the table of Surveyors' Long Measure so that link and chain stand for ideas in the mind of a child? Write the table.
4. Outline a first lesson on the subject of ratio.

(Michigan Department of Public Instruction, 1901, p. 70)

Without information on the grading of specific answers, it is difficult to tell exactly what types of knowledge test developers desired to assess; however, it seems likely that these state assessments were shaped by Michigan's growing class of professional teacher educators, some of whom perhaps called attention to the mathematical work teachers needed to do once inside classrooms. In fact, in 1900 the State Board of Education passed control of the state-level certification exams to the president of the normal school system (Michigan Department of Public Instruction, 1901, p. 69).

These results refer to the exams in one state; without similar and detailed analyses of teacher exams in other state archives, it would be difficult to tell whether Michigan's tests were typical. However, we think it likely that they were, and that the major issues regarding test content were common to other states, as well. It is especially striking that these major issues remain unresolved today: how difficult the exams should be—and by extension, how many prospective teachers should be excluded on the basis of lack of knowledge; whether it is content knowledge or knowledge of methods that make a good teacher; whether conceptual or procedural knowledge should be assessed. The passage of 125 years has done little to bring closure to these important questions.

These exams are also revealing in how they *measured* teachers' knowledge. Nineteenth-century teacher exams were designed to draw inferences about *individuals* and their capacity to teach children. Yet it would be nearly a half century before modern psychometrics, with its concerns for reliability and validity, began to influence the development of teacher certification exams. Meanwhile these assessments were used to make high-stakes decisions about individuals' careers without actual evidence of their effectiveness in this regard.

In the interim, however, the use of teacher exams shrank rapidly. As reported in Angus (2001), the late nineteenth century saw a growing class of professional teacher educators begin to criticize these exams, primarily on the grounds that they were too easy, thus allowing incompetent teachers to serve. Histories of teacher certification also note that local control of these exams encouraged favoritism and left open the possibility that examiners would know less than the teachers they tested. In response, teacher educators successfully convinced state bureaucracies and legislatures to approve what was then an alternative form of certification: completion of professional training, usually at a normal school or teachers' college. By 1937, 28 states had abolished teacher testing (Wilson & Youngs, 2005), requiring instead the completion of a professional preparation program. These programs, according to Wilson and Youngs, came to include courses in educational history, psychology, educational foundations, teaching methods, and assessment. Although teacher educators would soon face criticism from other academic disciplines, the first half of the twentieth century saw an increasing claim to professional knowledge in education, and

the assent of state legislatures to allow teacher preparation to occur in professional schools.

The early history of teacher testing, in fact, might help explain the aversion of many teacher educators to it. Angus (2001) describes this aversion as a byproduct of professionalization: teacher educators sought to legitimize their product and accumulate resources by claiming a codified body of knowledge and basing entry to their profession on the completion of coursework designed to cover that body of knowledge. Teacher tests like Michigan's, by virtue of being "too easy" and a "back door" into teaching, threatened teacher preparation programs. These tests were also largely controlled not by educational professionals but by either bureaucrats or elected officials. These twin problems—the perceived ease of teacher tests and the lack of control over testing by educational professionals—would afflict the profession in even greater ways later in the century, after teacher certification exams regained political favor.

Enabled by the development of standardized tests and test firms, and in response to teacher surpluses and urban school officials' concerns about how to hire the "best" candidates, certification exams reemerged in the second half of the 1900s. The National Teacher Examination (NTE), established in 1940, eventually became the most widely administered—and most studied—teacher examination in the U.S. The NTE saw three major growth periods. The first two were fueled less by concerns about underqualified teachers than by Southern white attempts to block efforts to reduce racial discrimination. As Baker (2001) describes, school districts in the South typically paid comparably educated African American teachers less than whites through the 1940s. Challenged by NAACP litigation, Southern districts eliminated this practice from official policy. To continue a de facto policy of discriminatory pay, many turned to using the early NTE, on which African Americans tended to score lower than whites. Ben Wood, the major author of the NTE, aggressively promoted the test for this purpose to Southern school officials, touring the South to spread the word. As Baker recounts, "When school leaders asked about how white and black teachers might score on the NTE, [Wood] informed them that on previous administrations of the test the average score of blacks was 'at the lower fifth percentile' of whites" (p. 326). During the second period of NTE expansion in the 1960s, Arthur L. Benson, director of teacher testing at the Educational Testing Service (ETS), which assumed control of the NTE during the 1940s, repeated Wood's strategy. This time, the issue was the desegregation of schools; Benson sold the NTE to Southern states by again pointing out that African American and white teachers performed differently on the exam. By 1968, nine Southern states required the exam of at least some teacher candidates. Criticism of this practice led ETS in 1971 to issue guidelines on the proper use of the NTE and to attempt to eliminate test bias, but African Americans continued to score lower—and thus be more likely to fail the certification tests—through the 1980s.

The third wave of state adoptions of the NTE occurred in the 1980s in response to *A Nation at Risk* (National Commission on Excellence in Education, 1983). This wave of adoptions was, by all reports, driven by public suspicion about teacher quality. The most popular among these tests were basic skills assessments (Haney et al., 1987). As Wilson and Youngs (2005) report, 37 states had adopted basic skills testing as a requirement for certification by 2002. Over the last two decades of the twentieth century, the percentage of teachers taking one of these assessments prior to entry into either preservice education programs or teaching increased dramatically. As in earlier periods, professional educators and professional education associations (e.g., the National Educational Association) remained opposed to the tests, and perhaps with good reason, as pressure mounted from non-education quarters, mainly economists, political scientists, and policy-makers who questioned the value of professional education in teaching (Hess, 2002; Walsh, 2001). The solution many non-educators propose is to return to a system in which passing an exam certifies a teacher to work in classrooms, regardless of the candidate's level of formal professional training or experience. Such proposals have met with strenuous objection from some involved in teacher education (see, for example, Berliner, 2005; Darling-Hammond & Youngs, 2002).

In this debate, much hinges on the quality of the test itself—what it assesses, and whether what it assesses is reasonably related to successful classroom teaching performance. The NTE tested mathematical knowledge among elementary and secondary teachers. All teachers took mathematics "items," as mathematics problems are referred to in modern assessment terms, as part of the "Core Battery" test, which assessed communication skills, general knowledge (including mathematics), and professional knowledge. Elementary teachers also took mathematics items as part of the "Elementary School Specialty Area Test," an examination that measured subject area knowledge and pedagogical knowledge across the elementary school curriculum.

Secondary teachers intending to be certified to teach mathematics took a stand-alone mathematics test.

Several items from an elementary-level core battery test illustrate the nature of this assessment:

1. An elevator operator is not allowed to carry more than 10 passengers in the elevator at one time. If there are 35 people waiting on the ground floor, what is the minimum number of trips up that the elevator must make in order to transport these 35 people?

   a) 3 b) $3\frac{1}{2}$ c) 4 d) 5 e) $5\frac{1}{2}$

2. Statement: "The product of any two numbers is always greater than or equal to either of those numbers."

   Which of the following examples proves the statement above FALSE?

   a) $1\times1=1$ b) $3\times4=12$ c) $5\times1=5$ d) $\frac{5}{2}\times4=10$ e) $\frac{1}{2}\times4=2$

3. Which of these is NOT a correct way to find 75% of 40?

   a) 75.0×40 b) $(75\times40)\div100$ c) $\frac{75}{100}\times40$ d) $\frac{3}{4}\times40$ e) 0.75×40

4. A pattern requires $1\frac{7}{8}$ yards of a certain fabric. If five remnants of the fabric are on sale and their lengths are as follows, which of these lengths will provide enough fabric and result in the least waste?

   a) $1\frac{1}{2}$ yd b) $1\frac{3}{4}$ yd c) $1\frac{15}{16}$ yd d) 2 yd e) $2\frac{1}{16}$ yd

5. Written as a percent, 2 =

   a) 0.02% b) 0.2% c) 2% d) 20% e) 200%

(ETS, 1984, pp. 68–73)

Two observations about this assessment stand out. In contrast to the early teacher tests we uncovered, the NTE Core Battery contains a mixture of basics (e.g., item #5 above), disciplinary knowledge (e.g., item #2 above, where candidates must recognize a counter-example), and word problems (e.g., #4 above) that reflect, like the early tests, the ways mathematics was imagined to be used in contemporary society. This new mix of items reflected, perhaps, the changing nature of mathematical knowledge in society.

Second, the items are arguably easier than those on both nineteenth century tests and contemporary tests. Unlike the historical tests, for instance, few items extend beyond the formal elementary school curriculum. And few problems are, like those on the nineteenth century assessments, numerically complex; in many cases, teacher candidates likely did not have to put pencil to paper to correctly answer the item. Further, quite a number of the items are extremely basic. Item #2, for instance, might assess mathematical reasoning and proof from one perspective, but from another, it simply asks candidates to recognize a case that violates the statement made in the problem situation. One reason for the easier assessments might be the methods adopted for testing teacher candidates, described below. Another might be that these were conceived in many quarters as basic skills tests. Nevertheless, the items contained on this assessment constitute a very different theory about the substance of teacher knowledge than earlier teacher certification exams, and from the certification exams to come.

Interestingly, some items from the NTE Elementary School Specialty Area test in the 1980s seemed to tap into what today we might call "mathematical knowledge for teaching." In fact, taken as a whole, these elementary exams presage what Shulman would soon name as "pedagogical content knowledge" (Shulman, 1986): the items convey a broad sense that teachers need a specialized knowledge, something beyond what other educated adults know. The following is an example of one such item (ETS, 1987, p. 19):

521	348	863	508	108
-386	-187	-37	-43	-26
245	261	836	545	182

A pupil's work on five subtraction problems is shown above. Which of the following is the most appropriate diagnosis and suggestion about the pupil's work in arithmetic?

a) The pupil has not had enough practice in doing this type of problem. The teacher should assign the pupil at least ten such problems each day for a few weeks.

b) The pupil shows no recognizable error pattern in the problems shown; the poor performance is probably due to lack of interest. The teacher should seek some means of motivating the pupil to give more attention to class work.

c) The pupil does not understand place value as it relates to subtraction problems. The teacher should have the pupil use concrete materials to explore this concept.

d) The pupil has no understanding of the subtraction concept. The teacher should have the pupil review basic subtraction facts with the help of concrete materials.

e) The pupil probably understands the subtraction concept but has not memorized the basic subtraction facts. The teacher should use drill activities and games to help the pupil memorize those facts.

Answering correctly by choosing option (c) requires knowledge of subtraction and the standard procedure for subtracting multi-digit numbers. It also emulates the reasoning that teachers do *in practice*—sizing up student work and deciding what to do next.

Another item focuses on a common mistake that students make when learning about the multiplication of rational numbers:

> Ralph has worked with decimal representations of numbers in class. In response to a decimal exercise posed by his teacher, Ralph wrote .2 × .4 = .8 on his paper. Ralph's work is best evaluated as
>
> a) correct, and this fact can be verified by multiplication of fractions
>
> b) correct, but this fact cannot be verified by multiplication of fractions
>
> c) incorrect, and multiplication of fractions can be used to show why it is incorrect
>
> d) incorrect, and decimal addition can be used to derive the correct answer
>
> e) incorrect, but neither multiplication of fractions nor decimal addition can be used to demonstrate why it is incorrect

It takes only basic mathematical knowledge to know that answer choices (a) and (b) are incorrect, because $.2 \times .4$ is .08. It would be possible to distinguish between answer choices (c), (d), and (e) using purely mathematical knowledge as well, but most people other than teachers are unlikely to have done such work or have occasion to do so. One would need to know that the model of multiplication as repeated addition (2×4 is the same as $4 + 4$) does not apply here: $.2 \times .4$ is not the same as $.4 + .4$. That in fact would yield .8, which is incorrect. This knowledge is not the sole province of teachers, but teachers need to know the meanings of operations for teaching in a way that the common educated person does not. The answer (c) is correct, because one could show that

$$\frac{2}{10} \times \frac{4}{10} = \frac{8}{100}$$

by multiplying across numerators and denominators. Again, this is not knowledge unavailable to adults other than teachers, but teachers would have the most facility working with such ideas about meanings of operations and how they can be modeled.

The next item is similarly constructed to gauge mathematical knowledge that is used in teaching multiplication (ETS, 1987, p. 23):

> In developing concepts of multiplication, the following types of exercises should be included:
>
> (1) 2×23 (3) 4×13
>
> (2) 3×20 (4) 5×35
>
> Which of the following is the most appropriate sequencing of these exercises?
>
> a) 1, 2, 4, 3
> b) 1, 3, 4, 2
> c) 2, 1, 3, 4
> d) 2, 3, 4, 1
> e) 3, 4, 1, 2

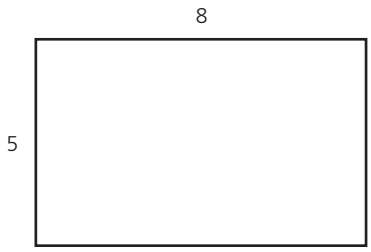
This item draws upon pure mathematical knowledge (multiplication of a one-digit by a two-digit number), but having such knowledge is not sufficient to identify the correct answer. This item also demands knowledge of how this topic is learned by children—the logical sequencing, the typical errors children make and the obstacles likely to be encountered by a child who is new to learning this skill. Thus, it encompasses basic mathematical competence but goes far beyond into mathematical territory that teachers traverse in their work with children.

However, test items that tap mathematical knowledge for teaching can be difficult to construct. As later generations of item writers discovered, particularly as scholars moved into measuring knowledge in this domain, items tend to suffer from three types of problems. One is the construction of an unambiguously correct answer. Teaching is heavily contextualized—teachers work with particular materials, and with particular children. It is also highly variable: textbooks have different approaches to teaching specific topics, and different developments of those topics over time. And students have unique learning and instructional histories. This contextualization and variability work against writing items in this domain. For instance, in the item above, arguments can be made for a sequence that starts with 3×20, as it yields easily to mental mathematics and modeling via a number line. Also, a teacher who plans to develop related number facts (3×2, 3×10, 3×20) to logically reason about the problem might choose one of the options that begins with 3×20 (b or c). However, a teacher who plans to begin with a partial product method might choose 23×2 over 20×3:

$$\begin{array}{rr} 20 & 23 \\ \times\, 3 & \times\, 2 \\ \hline 0 & 6 \\ 60 & 40 \\ \hline 60 & 46 \end{array}$$

As one can see, 20×3 yields a 0 in the first partial product, which might be confusing to children just learning this method. Items without clear correct answers were not uncommon occurrences in the tests we examined.

The following item illustrates a second problem associated with these tests—items that appear to be contextualized in teaching but that, in fact, do not require special professional knowledge (ETS, 1987, p. 28):



To find the perimeter of the rectangle above, a child can add 5 and 8 and multiply by 2 [2(5 + 8)] or multiply each number by 2 and then add [(2 × 5) + (2 × 8)]. This example illustrates an application of which of the following properties?

a) Distributive property of multiplication over addition
b) Commutative property of addition
c) Commutative property of multiplication
d) Associative property of addition
e) Associative property of multiplication

Although this is important knowledge to have, and a kind of mathematical problem teachers are likely to come across in their work with children, this item could be solved simply by knowing the mathematics (properties of addition and multiplication). The school learning context makes no further knowledge demands for determining the correct answer to the item; it is, for testing purposes, window dressing. Note this contrast with the previous two items, in which specialized knowledge of children and learning in the domain of mathematics was required to determine the correct answer.

Third, some attempts to measure the mathematical knowledge used in teaching produced items that that depended more on ideology than professional skill. For instance (ETS, 1987, p. 15):

In teaching a third-grade class regrouping with two-digit numbers for the first time, which of the following would provide the best means for a teacher to demonstrate the concept?

a) Manipulatives
b) Illustrations drawn on the chalkboard
c) A chart
d) Examples from an approrpiate mathematics textbook
e) Examples from a local newspaper

The test materials identify (a) as the correct answer. Yet we might argue that several of these possibilities are plausible, and that it is not clear on what evidence any one answer is best. For example, a diagram (b) that showed the regrouping of tens and ones and linked it to the symbolic form might help pupils to link the representation to the written procedure more clearly than with manipulatives. It is unlikely that there is any one best approach in this example, and as we pointed out above, good teaching would likely depend on the particulars of the situation. Here ideology about teaching—that concrete materials are always best—seems to override judgment.

These items demonstrate how an earlier version of teacher tests had an underlying sensibility about specialized knowledge for teaching, one that predates the formalized naming of "pedagogical content knowledge." Second, the items show the difficulty of constructing measures that tap into this knowledge domain with precision. We return to both issues below.

In both NTE tests, mathematical knowledge at the elementary level is subsumed under "general knowledge"—it is one of several areas (literature, fine arts, social science) in which the successful teacher candidate was expected to be conversant. Scores were not returned for subjects separately. In this view of teacher knowledge, general proficiency across a wide range of topics is needed for successful teaching.

In addition, the methods for measuring teacher knowledge underwent a significant shift in the years since our nineteenth-century examples. With the development of more sophisticated methods for scaling items and scoring responses, test developers could better assure candidates and education officials of at least the reliability of the measures. These new methods—and related technology, such as computer-scannable response sheets—changed the face of the tests. Multiple choice items came into vogue. More items were put on the assessments to increase the accuracy of individual-level scores. And, perhaps to increase the precision of measurement among less knowledgeable teachers, easier items were added.

One area of trouble for the NTE, however, was its predictive validity—that is, whether teachers' performance on the assessment correlated with the quality of their classroom instruction or their students' learning. As Wilson and Youngs (2005) describe, six studies addressed the relationship between NTE score and student learning; all but one failed to find a relationship. Official documents make clear that ETS never claimed predictive validity, however, instead proffering their test construction process as proof of validity. The tests were written at the request of educators, to specifications drafted and reviewed by educators, and based on an analysis identifying what teacher candidates should know in a particular area, the tests must represent the knowledge needed to teach well.

The claims to validity made by ETS and other large teacher testing firms lead back to the issue of control over these professional tests. One hallmark of a profession, in fact, is control of barriers to entry by members of the profession itself (Abbott, 1988).

Yet through the 1980s, and even today, there has been only a weak claim to professional control over teacher assessments (Haney et al., 1987). Even after control of teacher testing shifted from local and state bureaucracies to large testing firms, the representation of professionals—in this case, mathematics educators, teachers, and mathematicians—on key decision-making panels remained more nominal than real. Certainly testing firms have included teachers, school officials, disciplinary experts, and others in the test construction process. On the whole, however, the leaders in the field have not tended to play a role or to be consulted in this work. Instead, control over the tests remains to a large extent in the hands of psychometricians, other testing experts, and officials from the state departments of education that mandate the assessments.

Throughout U.S. history, then, teacher tests have been largely controlled by those outside the education profession: local officials, state legislators and bureaucrats, and testing firms. By many accounts (Berliner, 2005; Haney et al., 1987; Wilson & Youngs, 2005), these tests have tended to capture basic skills rather than more complex and job-specific competencies associated with helping students learn.

At the same time, rather than contributing expertise to the process, many educators condemn any form of standardized testing. In a 2002 issue of *English Education*, nearly all articles railed against the very idea of standardized testing for teachers:

> Virtually all of the criticisms leveled against testing in schools also apply to the quick and dirty attempt to demand accountability in testing teachers. Timed tests given to children are really evaluating speed rather than thoughtfulness, and the same is true when they're given to adults. Multiple choice tests and contrived open response items are not meaningful ways of assessing how much students understand, and neither are they particularly effective in telling us how well educators can educate. (Kohn, quoted in Appleman & Thompson, 2002, p. 96)

Others take the view that educators need to assume control over licensing exams, moving toward new forms of assessment and focusing on professionally relevant knowledge and skill. We review some of these new assessments in the last section of this chapter. First, however, we review *scholarly* approaches to *studying* teachers' mathematical knowledge—studies that measure teacher knowledge not for high-stakes decisions, but instead for the purpose of uncovering what teachers do know, understanding the relationship between teacher knowledge and student learning, and, eventually, for understanding the nature of knowledge and knowing in teaching. Insights from this body of work have opened new possibilities for the assessment of teachers' mathematical knowledge.

EARLY SCHOLARLY RESEARCH INVOLVING THE MEASUREMENT OF TEACHERS' MATHEMATICAL KNOWLEDGE

Studies measuring teachers' knowledge date to the 1960s, originating in what is now known as the "educational production function" literature. The main goal of this research program was to predict student achievement on standardized tests as a function of the resources held by students, teachers, schools, and others. Key resources were seen to include students' family background and socioeconomic status, district financial commitments to teacher salaries, teacher–pupil ratios, other material resources, and teacher and classroom characteristics (Hanushek, 1981; Greenwald, Hedges, & Laine, 1996). Published mainly in economic journals, these studies have been influential in shaping American public opinion and policy on how best to foster school improvement. In particular, educational economists' studies of teachers and their knowledge and skills have been especially influential.

One approach taken in this literature involves using data on teachers' preparation, coursework, and experience to predict student achievement. Key measures included overall teacher education level, certification status, number of post-secondary subject matter courses taken, number of teaching methods courses taken, and years of experience in classrooms. By using such measures, researchers assumed a connection between formal schooling and employment experiences and the more proximate aspects of teachers' knowledge and performance that produce student outcomes. This makes sense, given teacher educators' assertion that formal preparation provides the knowledge needed for teaching. However, reviews of educational production function studies have disputed the extent to which variables like teacher preparation and experience in fact contribute to student achievement (Begle, 1972, 1979; Greenwald et al., 1996; Hanushek, 1981, 1996), with conflicting interpretations resting on the samples of studies and methods used for conducting meta-analyses. Wayne and Youngs (2003) argue that when studies use subject-matter specific markers of teacher preparation—for instance, certification in mathematics rather than general certification—results are a bit

more positive. This increasing subject-matter specificity is one characteristic of studies conducted in the educational production function tradition.

A second approach measures teacher knowledge by looking at teachers' performance on certification exams or other tests of subject-matter competence. The first study to do so was the Coleman report, *Equality of Educational Opportunity*, completed in 1966. Coleman and colleagues measured teacher knowledge via a multiple-choice questionnaire, then used teachers' scores to predict student achievement in both reading and mathematics. They found that teacher scores did indeed predict student achievement in both subjects, and that this relationship grows stronger in the higher grades. Notably, however, none of the items on their measure focused specifically on mathematics; instead, the questionnaire asks teachers to complete a "short test of verbal facility" with items such as:

> Dick apparently had little _____ in his own ideas, for he desperately feared being laughed at.
>
> a) interest
>
> b) depth
>
> c) confidence
>
> d) difficulty
>
> e) continuity

Thus the first study to identify a positive relationship between teacher knowledge and student mathematics achievement did not actually measure teachers' mathematical knowledge at all. Instead, it measured teacher knowledge via a vocabulary test—close, some would argue, to a test of general intelligence. Similar studies have taken advantage of available data from the NTE to create composite measures of teacher knowledge to compare to student achievement scores (Strauss & Sawyer, 1986; Summers & Wolfe, 1977). And more recently, Ferguson (1991) found that the mix of reading skills and professional knowledge captured on the Texas Examination of Current Administrators and Teachers, an exam used in the 1980s, was positively related to student achievement.

By the 1990s, some studies began to focus on how teachers' *mathematical* knowledge related to student gains in mathematics achievement (Harbison & Hanushek, 1992; Mullens, Murnane, & Willett, 1996; Rowan, Chiang, & Miller, 1997). This move toward subject-matter-specific measures coincided with increasing indications that teachers' effectiveness was related to their knowledge of subject matter rather than their general or pedagogical knowledge. Rather than use proxy measures such as degrees or mathematics coursework, however, the strategy was to administer (or obtain teachers' scores on) problems they might assign students. Harbison and Hanushek (1992), for instance, administered a fourth grade student assessment to both teachers and students, using scores from the first group to predict performance among the second. Mullens et al. (1996) used teachers' scores recorded on the Belize National Selection Exam, a primary-school leaving exam administered to all students seeking access to secondary school. In both cases, the authors provided little information about the actual content of the assessment; Harbison and Hanushek (1992) listed the mathematical topics tested (number recognition, measurement, multiplication and division, rational numbers, unit measures, four operations, story problems) but did not include any actual items. Because these articles described assessments given to students, however, it seems probable that both studies measured teachers' competency with basic computation and mathematical procedures, rather than knowledge of teaching those topics.

Rowan et al. (1997) employed a somewhat different approach, one that began an increased effort to capture the mathematical knowledge used in classrooms, a trend we will take up in more detail in the next section of the paper. In this study, the authors used a one-item assessment to explore how teacher knowledge related to high-school student gains in the National Education Longitudinal Study. The item, developed by the Teacher Education and Learning to Teach study (Kennedy, Ball, & McDiarmid, 1993), reads:

> Your students have been learning how to write math statements expressing proportions. Last night you assigned the following:
>
> > A one-pound bag contains 50 percent more tan M&Ms than green ones. Write a mathematical statement that represents the relationship between the tan(t) and green(g) M&Ms, using t and g to stand for the number of tan and green M&Ms.
>
> Here are some responses you get from students:
>
> Kelly: 1.5t = g
> Lee: .50t = g
> Pat: .5g = t
> Sandy: g = 1/2g = t
>
> Which of the students has represented the relationship best? (Mark ONE)
>
> ❑ All of them
> ❑ Kelly
> ❑ Lee
> ❑ Pat
> ❑ Sandy
> ❑ None of them. It should be _____
> ❑ Don't know.

When stripped of the teaching context, this might easily be a problem that would appear on an end-of-chap-

ter test or a student exam. The problem is different in that it is situated in a common task of teaching. However, like items on the NTE test, the teaching situation can be considered window-dressing.

Studies that directly measure teachers' verbal or mathematical knowledge typically show a positive relationship to student mathematics achievement (e.g., Boardman, Davis, & Sanday, 1977; Ferguson, 1991; Hanushek, 1972; Harbison & Hanushek, 1992; Mullens et al., 1996; Rowan et al., 1997; Strauss & Sawyer, 1986; Tatto, Nielsen, Cummings, Kularatna, & Dharmadasa, 1993; for an exception, see Summers & Wolfe, 1977; for reviews, see Greenwald et al., 1996; Hanushek, 1996; Wayne & Youngs, 2003). By using such measures, these studies assume a relationship between teacher content knowledge as measured by such assessments and the kinds of teaching performances that produce improved student achievement. Yet there is little evidence in this literature for what is actually measured by these assessments, and how what is measured relates to classroom performance. These tests were not developed, for the most part, by specifying domains of teacher knowledge and developing an assessment from this map. Instead, the claim to validity is that these assessments covered material in the K–12 curriculum, and which should be known by the students themselves at the end of schooling. This is true—but may miss important elements of the knowledge that makes teachers successful. And using these theoretically impoverished tools also limits the conclusions that could be drawn from these studies. Is it that teachers who perform better on these student-level tests simply make fewer mistakes in classroom teaching? Or, do teachers who perform better offer a qualitatively different kind of mathematics instruction? Measuring quality *teachers* through performance on tests of verbal or mathematics ability may overlook key elements in what produces quality *teaching*.

The conjecture that teachers may need to know subject matter differently than their students or nonteachers has become the subject of intensive research over the last two decades. In his presidential address to the American Educational Research Association, Shulman (1986) argued for the centrality of subject matter in teaching, drawing attention to the particular ways that teachers must know and use content knowledge in their work. He introduced the term "pedagogical content knowledge," as a special kind of teacher knowledge that intertwines content and pedagogy. According to Shulman (1986, 1987) and colleagues (Wilson, Shulman, & Richert, 1987), pedagogical content knowledge includes understanding which topics students find interesting or difficult, the common misconceptions that students have, and what forms of representation are useful for teaching particular topics. More recently, scholars who focus on studies of mathematics teachers and teaching have hypothesized that in addition to topic-specific knowledge of students and teaching, teachers might in fact use subject-matter knowledge in ways unique to teaching. Our research group, for instance, has developed the notion that teachers have mathematical knowledge that is "specialized" for the work of teaching. This specialized knowledge of the content can be seen in the demands of providing explanations for mathematical procedures or ideas, representing mathematical phenomena, and working flexibly with non-standard solution methods (Ball & Bass, 2003; Ball, Thames, & Phelps, 2005; Hill, Schilling, & Ball, 2004). In her study of Chinese teachers, Ma (1999) points to the depth and coherence of these teachers' knowledge both over the range of mathematics topics and over time, a kind of knowledge that she labeled "profound understanding of fundamental mathematics." And Ferrini-Mundy, Floden, McCrory, Burrill, and Sandow (2005) argue that teaching algebra requires teaching-specific mathematical practices, among them "bridging" between what students know and disciplinary knowledge, and "trimming" disciplinary knowledge to levels that are understandable by students, but that are also still mathematically accurate.

To illustrate what it means to know mathematics *for teaching*, not just to know mathematics, we turn to an example from a topic in the upper elementary curriculum: division of fractions. By considering the ways teachers might need to know the mathematics behind this topic, we offer conceptual proof that there is more to be known than has been captured to date in most of the educational production function studies described above. This conceptual proof also serves as an outline of the mathematics-specific teaching knowledge used by teachers. Following the example, we then turn in the next section to measurement methods used in the literature that uncovered this knowledge.

We begin with the following problem:

$$\frac{5}{6} \div \frac{1}{3}$$

A common way to calculate the answer is using the infamous "invert-and-multiply" algorithm:

$$\frac{5}{6} \div \frac{1}{3} = \frac{5}{6} \times \frac{3}{1} = \frac{15}{6} = \frac{5}{2} = 2\frac{1}{2}$$

To teach division of fractions to students, teachers must of course be able to carry out the procedure correctly themselves. Thus one aspect of knowing mathematics for teaching is being able to do the mathematics that one is teaching one's students. Yet being able to divide fractions oneself is far from sufficient. After all, teaching mathematics is not simply "knowing" in front of students. Teaching requires making the content accessible, interpreting students' questions and productions, and being able to explain or represent ideas and procedures in multiple ways. For example, a student might ask a teacher why it works to invert and multiply, or why the second fraction is "flipped over" rather than the first. Or, a student might be perplexed because the answer, 2½, is larger than both 5/6 and 1/3—isn't dividing supposed to make the answer smaller? To respond to these questions, teachers must understand and be able to explain why the algorithm works, and be able to explain why and in what cases the quotient is larger than both the dividend and the divisor.

Or, instead of questioning why the standard procedure works, a student might claim that she doesn't even need to learn how to invert and multiply because she has found an easier way to divide fractions—namely, to just divide the numerators and divide the denominators:

$$\frac{5}{6} \div \frac{1}{3} = \frac{5 \div 1}{6 \div 3} = \frac{5}{2} = 2\tfrac{1}{2}$$

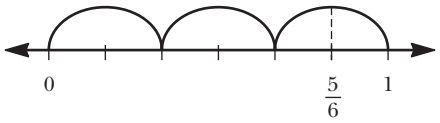
This situation is not uncommon; students often develop their own methods of computation—some valid and some not—and teachers must be able to size up these alternative approaches. In this case, the teacher would have to be able to judge whether the student's method for dividing fractions is mathematically valid or whether the correct answer was simply a coincidence. And, if the method is valid for this particular problem, would it work in general to divide any two fractions, or does it only work in special cases? Generating and testing examples and counterexamples in teaching involves mathematical reasoning. In fact, even the disposition to consider a method's generalizability and efficiency is in itself an example of the type of mathematical knowledge used in teaching.

What else arises in teaching division of fractions? Students, of course, do not always solve problems correctly. For example, suppose a teacher sees the following calculation on a student's paper:

$$\frac{5}{6} \div \frac{1}{3} = \frac{6}{15} = \frac{3}{5}$$

Spotting this as an incorrect answer is insufficient. A teacher must be able to figure out what steps the student might have taken to produce this error, as well as the reasons it might have been made, and then, in light of this mathematical analysis, determine an appropriate response.

After analyzing the above error, for example, a teacher might decide to pose a word problem that corresponds to the calculation to encourage the student to think about whether 3/5 is a reasonable answer. Or, the teacher might decide to use a number line to represent the calculation:



Choosing representations such as story problems or diagrams involves mathematical reasoning and skill. For example, to use the number line representation above, a teacher needs to be able to map each element of the computation to the representation: Where is the 5/6, the 1/3, and the 2 ½? Why does this representation model division? In this case, the representation is using a measurement interpretation of division (i.e., How many 1/3s are in 5/6?). Five-sixths is marked on the number line and each loop represents "measuring off" 1/3. Together, the loops represent the number of 1/3s that "go into" 5/6: Two full loops and a half a loop, or two and a half 1/3s. In addition to making these correspondences, explaining this representation requires an understanding of central mathematical ideas such as the meaning of division and attention to the unit. The teacher must also recognize that, although this representation models division with fractions and shows why the quotient is greater than the dividend and the divisor, it does not readily explain the invert-and-multiply algorithm.

For this purpose, a teacher might rely on the inverse relationship between multiplication and division, noting that dividing by a number is equivalent to multiplying by its reciprocal. Alternatively, a teacher could appeal to students' knowledge that multiplying by 1 does not change the value of a number. In this case, the complex fraction

$$\frac{\frac{5}{6}}{\frac{1}{3}}$$

is multiplied by a convenient form of 1,

$$\frac{\frac{3}{1}}{\frac{3}{1}},$$

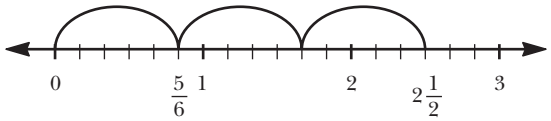
to show that $\frac{5}{6} \div \frac{1}{3} = \frac{5}{6} \times \frac{3}{1}$:

$$\frac{5}{6} \div \frac{1}{3} = \frac{\frac{5}{6}}{\frac{1}{3}} = \frac{\frac{5}{6}}{\frac{1}{3}} \times 1 = \frac{\frac{5}{6}}{\frac{1}{3}} \times \frac{\frac{3}{1}}{\frac{3}{1}} = \frac{\frac{5}{6} \times \frac{3}{1}}{\frac{1}{3} \times \frac{3}{1}} = \frac{\frac{5}{6} \times \frac{3}{1}}{1} = \frac{5}{6} \times \frac{3}{1}$$

Another means for explaining the invert-and-multiply algorithm again turns to the meaning of division, but this time using a partitive interpretation. In the case of

$$\frac{5}{6} \div \frac{1}{3},$$

the partitive interpretation asks: What number is 5/6 one-third *of*? Like the measurement interpretation, the partitive interpretation can also be represented on a number line. If 5/6 is one-third of the length of the whole, then the length of whole can be found by putting together three lengths of 5/6, in other words, multiplying 5/6 by 3, which is equivalent to $5/6 \times 3/1$, which is precisely the invert-and-multiply step:



Understanding these subtleties requires a robust understanding of the different meanings of operations and the ways they can be represented. And, as this example shows, teachers not only need to be able to map between a representation and the problem or concept being modeled, they must also determine which representation or explanation would be most appropriate for a particular instructional goal. Thus, and significantly for the assessment of teachers' mathematical knowledge, an item on a test for teachers that asks for the correct answer to

$$\frac{5}{6} \div \frac{1}{3}$$

might reveal knowledge of basic mathematics, but not *all* the possible types of knowledge needed for teaching this topic.

Clearly, global and verbal assessments do not capture the sort of mathematical knowledge and reasoning required to navigate the situations of teaching sketched in the example above. At the time when most of the education production function studies were conducted, however, few alternatives existed; scholars needed to make inferences about teachers' mathematical knowledge based on proxies and indirect indicators. Aligning teachers' and students' scores on tests was a good bet. Yet these studies can only make inferences about the value of the specific knowledge domains that appear on the assessments: teachers' knowledge of the content they are directly responsible for teaching. The same can be said of the teacher certification exams reviewed to this point: although teaching-specific mathematics questions crept in and out of the tests of the nineteenth and twentieth centuries, they never captured much real estate. Why is this important? Without broader conceptualizations of teacher knowledge and without assessments that reflect this broader conceptualization, we lack precision about the kinds of teacher knowledge that matter. Is it, as some have concluded from analyzing the educational production function literature, simple competency in the topics teachers teach students? Or do teachers need to know "advanced" knowledge, a view common in the 1800s and again today (Education Trust, 1999)? Or, do teachers need specialized knowledge of the topics they teach, illustrated by the teaching tasks in our example? If we had assessments that reliably and validly measured these different conceptions of mathematical knowledge for teaching, and could construct models of student achievement using teachers' scores on such assessments, we could specify more precisely the nature of the knowledge that makes a difference for instructional quality and student learning. Without this precise specification, we cannot design preservice and in-service learning experiences that equip teachers with this knowledge.

Through the 1980s and 1990s, researchers began to approach these questions from another perspective: studying teachers' knowledge by studying teaching and teachers. Through this work, researchers began to hypothesize about the discipline-specific knowledge of content, students, and instruction required for the work of teaching. The studies they conducted focused on teachers' knowledge in practice situations; nearly all involved measuring or making inferences about teacher knowledge in some way. We turn next to a discussion of the strategies used by researchers engaged in this line of research.

METHODS FOR MEASURING *PROFESSIONALLY SITUATED* MATHEMATICAL KNOWLEDGE

Sparked in part by Shulman and colleagues' conceptualization of pedagogical content knowledge, and by a perceived need to assess the content knowledge needed for teaching as a basis for certification and licensure decisions, researchers in the 1980s

and 1990s sought to identify what teachers know (or should know) in teaching classroom mathematics. From the start, these studies suggested that teaching requires more than the ability to do the mathematics in the school curriculum. Teaching mathematics, in this view, is not the same as standing at the board and doing mathematics in front of students; it entails additional mathematical knowledge, competencies, and skills—what we call "mathematical knowledge for teaching" (Ball & Bass, 2003). This knowledge, as shown in the division of fractions example above, is multifaceted, including not only teachers' ability to solve the problems their students are expected to solve, but also to understand the content in the particular ways needed for teaching it, to understand what students are likely to make of the content, and to craft instruction that takes into account both students and the mathematics.

In this section of the chapter, we discuss the methods used to study—and later measure—mathematical knowledge for teaching. To do so, we shift from making inferences about conceptions of teachers' mathematical knowledge based on analyses of assessments, as we did with historical teacher exams and the NTE, to exploring the tools used to develop a theory of mathematical knowledge for teaching. We do not provide results from these studies, or chronicle the growth of this concept, because others have done so (Ball et al., 2001). Instead, we focus on the methods used, in part to provide readers with a sense for the array of methods and instruments that can measure mathematical knowledge for teaching. We also discuss the inferences that can be made from each type of research and consider each method's advantages and disadvantages.

We have organized this section of the chapter both by method and chronology. We first discuss observations of teaching practice, beginning with efforts in the 1980s to uncover aspects of mathematical knowledge for teaching and later, post-2000, as they began to be used as a method for quantifying the mathematical characteristics of classroom instruction. We next discuss written tasks and interviews, linked methods used in the 1980s and beyond to assess the extent to which teachers held mathematical knowledge in forms useful to instruction. As concern over the level of teachers' mathematical knowledge grew, so did programs designed to improve such knowledge—and measures to evaluate the results of these programs. We then describe some of the results of efforts to develop such evaluation tools—multiple-choice and other methods designed to assess teacher learning.

The assessment methods used in (and lessons learned from) the studies described below are related only distantly to the methods used in early teacher certification exams and to the assessments given to teachers in the educational production function studies. While they are related more closely to the newer teacher certification exams described in the next section, development of the methods and content for both types of assessment have proceeded, to some degree, on separate tracks. One result is that lessons from one field (e.g., large-scale certification testing) do not easily penetrate another field (e.g., close studies of teacher knowledge). We observe here and in the next section the extent to which the separation of these paths has inhibited the development of stronger measures.

Uncovering Mathematical Knowledge for Teaching: Observations of Teaching Practice

Perhaps the earliest and most widely adopted technique for measuring teachers' mathematical knowledge is one that, on its face, does not appear to be a measurement strategy at all: direct or videotaped observations of teachers' mathematical instruction. However, because scholars followed these observations by careful analysis and explication of the quality and characteristics of the mathematics delivered to students, and because scholars often made inferences about teachers' knowledge from such evidence, we argue that this strategy verged on measurement. We describe some of these early efforts below.

Early observational research focused on teachers' mathematical knowledge had some common elements. Researchers typically collected tens, if not hundreds, of hours of observations or videotapes; published research, though, typically focused on a tiny fraction of the data, often even just a few minutes. Analysis that appeared in print was primarily qualitative, with researchers using methods and coding systems tailored specifically to the mathematical topics and questions at hand. Scholars typically combined observational records with other sources of data to gain insight into mathematical knowledge for teaching.

One classic example of a study in this tradition is Leinhardt and Smith's (1985) study of expertise in mathematics instruction. To explore the relationship between teacher behavior and subject matter knowledge, the authors studied eight teachers intensively, collecting three months of observational field notes from their mathematics lessons, ten hours of videotaped lessons, interviews about the videotaped lessons and other topics, and responses to a card sort task. They used the classroom observations of instruction to construct a measure, ranking teachers' knowledge as high, medium, or low based on "in-class discussions

over 3 years and by considering their presentations and explanations as well as their errors" (p. 251). This strategy—sorting teachers by their actual in-class mathematical performance—is rare in the literature, and, unfortunately, not well explicated in their published work. The authors then examined teachers' knowledge in light of performance on interview tasks and examined three teachers' teaching of fractions in much more depth by intensive description of single lessons on simplifying fractions. This method—thick description (Geertz, 1973)—would prove a mainstay in probing the mathematical knowledge needed for teaching.

Borko et al. (1992) provide another example of the observational method for measuring teachers' knowledge of mathematics. Their article focuses on a few minutes of an hour-long review lesson, moments in which a student teacher was asked by a student to explain the division of fractions algorithm. Audiotapes of the lesson were augmented by fieldnotes taken by live observers, and allowed the construction of a dependent variable of sorts: an assessment of this teacher's capacity to provide a conceptually-based justification for the standard algorithm. The authors showed that this teacher's capability in this area was poor—she used a concrete model that actually represented multiplication rather than division. They then used data from interviews, her performance on open-ended mathematics problems, and records from this teacher's teacher education program to explain her in-class performance.

These early observational studies began to record elements of mathematical knowledge for teaching: presentations, representations, explanations, linking between key mathematical elements in instruction, flexibility, and other uses of content knowledge that could be seen in observations of teachers' work. As a result, many of these studies were largely descriptive; rather than inferring any *individual's* overall level of knowledge, classroom practice, or learning, these studies served as explorations of the *territory* of mathematical knowledge for teaching. Borko et al., for instance, did not make global generalizations about their teacher's level of mathematical knowledge for teaching or mathematics instruction per se, or attempt to use their observations to inform any high-stakes (e.g., licensure) decision. This is entirely appropriate, given the limitations of the method.

More recently, several standardized protocols for observing the mathematical quality of classroom instruction (or videotaped records) have emerged in response to the need to study teacher performance, learning, and change. These protocols are designed to make more generalizable inferences about particular teachers' classroom mathematical work, and thus differ from the early observational studies in the sense that they are intended to evaluate, with some degree of confidence, how well a teacher or group of teachers can teach mathematics. Leaving aside instruments that focus solely on the pedagogical aspects of teaching mathematics—e.g., the degree to which students work in groups, work on extended investigations, or answer questions—a number of instruments combine descriptions of the nature of classroom work with estimates of teachers' skill and knowledge in teaching: the *Reformed Teaching Observation Protocol* (RTOP) (Sawada & Pilburn, 2000), *Inside the Classroom Observation and Analytic Protocol* (Horizon Research, 2000), and the *Learning Mathematics for Teaching: Quality of Mathematics in Instruction (LMT-QMI)* instrument (LMT, 2006a).[3] These instruments vary in the extent to which they infer the *quality of the mathematics in instruction*, as opposed to the *quality of mathematics instruction*. All three instruments, for instance, ask for ratings of the extent to which content is presented accurately. All three ask whether the content presented to students is mathematically worthwhile. And all three ask about some elements of what some consider "rich" instruction—the use of representations, explanations, and abstractions, for instance. As such, all three contain elements that some might use to infer the *quality of the mathematics in instruction*, or the accuracy of content, richness of representation and explanation, and connectedness of classroom tasks to mathematical principles. We argue that while not a measure of teacher knowledge per se, the quality of the mathematics in instruction is a product of, and thus closely related to, teachers' mathematical knowledge for teaching.

RTOP and Horizon's instruments, however, both embed the ratings of teacher knowledge in larger scales intended to measure the extent to which classroom instruction aligns with the National Council for Teachers of Mathematics standards. For instance, RTOP has the following prompts:

- In this lesson, student exploration preceded formal presentation.
- This lesson encouraged students to seek and value alternative modes of investigation or of problem-solving.
- Connections with other content disciplines and/or real world phenomena were explored and valued.

[3] We also surveyed the TIMSS-R Video Math Coding Manual (2003). The coding effort here appears more focused on the quality of the problems presented and enacted in classrooms, rather than making inferences about teacher knowledge.

- Students were actively engaged in thought-provoking activity that often involved the critical assessment of procedures.

The issue here is not whether these are desirable lesson characteristics, but instead what is being measured. These two instruments are designed, according to their materials, to measure the quality of mathematics instruction, where quality is defined as both the richness and correctness of mathematical content *and* the way material is conveyed to students. In these instruments, no direct measure of teacher knowledge is available. Instead, teacher knowledge is estimated as a component of how mathematical material is presented to students. The RTOP and Horizon instruments are also designed for rating both science and mathematics lessons, which limits the specificity with which they can ask about particular mathematical practices.

The LMT-QMI instrument was intended to serve as a means to validate multiple-choice measures of teachers' knowledge for teaching mathematics. As such, it is catholic with respect to teaching style, but places heavy emphasis on ways in which teachers' mathematical knowledge might appear in instruction. It includes many of the elements common to the three instruments, but at a finer level of specification. The accuracy of mathematical content, for instance, is assessed over a broad domain of prompts, including use of mathematical language, computational errors, explanations, and notation. The rubric also asks about the presence and accuracy of mathematically rich elements of instruction, including representations, links between multiple representations, explanations, justifications, and the explicit development of mathematical practices. As such, the LMT-QMI measures the quality of the mathematics in instruction.

Observations of teaching have high validity, if the goal is measuring mathematical knowledge for teaching, because classrooms are where mathematical knowledge for teaching is expressed—in teachers' classroom moves, explanations, representations, and computations. However, all three instruments suffer from two main problems. The first involves language and interpretation. Put simply, prompts such as "the mathematics/science content was significant and worthwhile" are just a collection of words; their interpretation will be shaped by the coders' knowledge and views of mathematics itself, and will vary significantly across individuals. In our own work using LMT-QMI, for instance, we have found that the rating of an explanation as accurate or inaccurate—or even as an explanation at all—is shaped by observers' knowledge of the specific mathematical topic, knowledge of what constitutes explanation within mathematics, knowledge of mathematics education, and interpretation of the instrument. The three instruments attempt to eliminate this problem in various ways: Horizon researchers, for instance, spent two days in training and were given an annotated coding manual prior to conducting actual observations. LMT coders spent nearly two years coming to agreement on the coding scheme itself, and on how particular video clips should be rated. And RTOP offers online training for those interested in using their instrument.

The second problem relates to generalizability and the logistics of measurement in the context of research projects: one must observe teachers for multiple lessons before reaching any conclusions about their level of mathematical knowledge for teaching (Rowan, Harrison, & Hayes, 2004; Stein, Baxter, & Leinhardt, 1990). On any particular day, a teacher might be working on a topic with which she is unfamiliar, making a generalization from that day's performance to her overall level of knowledge invalid; there is also natural variation in teachers' knowledge across content areas (number, operations, geometry) that may bias results from a small sample of lessons. Clearly, making many observations over a large sample of teachers participating in preservice or in-service coursework would be a significant burden on any study; the hundreds of teachers required to model statistically student achievement effectively prohibits this strategy. Thus while classroom and videotape observations can be generative for answering fine-grained questions about particular teachers working with students around particular topics, this technique is not amenable to studies that formally test the effects of preservice or teacher development programs, or that link different conceptualizations of teacher knowledge to student achievement.

Exploring Teacher Knowledge: Mathematical Interviews and Tasks

A second method used widely for investigating mathematical knowledge for teaching is mathematical tasks and interviews. These mathematical tasks are different from mathematics assessments, which we consider below, by virtue of the fact that they are not designed to yield generalizable inferences about individual participants' knowledge, but to help scholars understand the nature and extent of teachers' knowledge. In other words, the focus of inference is a group of teachers' performance on the task itself, often to extend our understanding of the task or of teacher knowledge very generally, rather than the performance of individual teachers (e.g., drawing inferences

about mathematical knowledge as one component of suitability for teaching). Tasks can be paper-and-pencil problems or included as components of interviews exploring teachers' mathematical knowledge (Borko et al., 1992; Ma, 1999; Simon, 1993; Tirosh & Graeber, 1990) and have typically been designed to study the nature of teachers' mathematical knowledge in specific content areas such as multiplication, rational numbers, division, geometry, and functions (Ball et al., 2001). Some of the tasks that have been used to study teachers' knowledge, such as those derived from elementary textbooks or from studies of children's thinking, could also be used to assess students' mathematical knowledge. Other tasks have been explicitly designed to investigate mathematical knowledge beyond what students would be expected to be able to do, such as explaining what misconception led to a student's alternative method or generating a representation that could be used to teach a particular concept. These tasks are often based on situations that arise in teaching such as analyzing a student's error or answering a student's question. Providing a detailed analysis of all of the mathematical tasks and interview probes that have been used to study teachers' mathematical knowledge for teaching is beyond the scope of this chapter; however, we describe below some of the tasks that have been used in one content area, division, as a representative sample of the types of tasks that have been used.

Some tasks used on written assessments and interviews were, literally, the same tasks given to students. For instance, Graeber, Tirosh, and Glover (1989) investigated the extent to which prospective elementary teachers held misconceptions about multiplication and division similar to those commonly held by children—namely, that multiplication always yields larger numbers and division always yields smaller numbers. They used a written test constructed by slightly modifying 26 problems that had been administered by Fischbein, Deri, Nello, and Marino (1985) in their study of adolescents' misconceptions about operations used to solve multiplication and division word problems. The test presented various word problems and asked respondents not to perform the calculation, but "to write an expression in the form of 'a number, an operation, and a number' that would lead to the solution of the problem" (Graeber et al., 1989, p. 96). It was administered to 129 prospective elementary teachers; 33 follow-up interviews were conducted. Interviewees were given a problem similar to one that had been answered incorrectly on their written form. If they still answered incorrectly, they were asked to explain their answer and show how they would check their solution. Upon realizing (or being told) that the original expression was incorrect, interviewees were asked to explain what they thought might have caused them to write an incorrect expression. This example shows the ways tasks and interviews might be combined. It also demonstrates one major strength of this method: questions can be tailored to the answers given by specific respondents, often on the spot, to probe the reasons for misconceptions or the support for understanding. It also illustrates a central finding of the studies that used this method: some prospective teachers had serious gaps in their understanding of the mathematics they would be expected to teach.

In another study, Tirosh and Graeber (1989) investigated whether prospective teachers' misconceptions about multiplication and division were explicitly held or just implicitly influenced their calculations. A written instrument was administered to 136 prospective elementary teachers, and then 71 were interviewed. The test asked respondents to label each of the following statements as "true" or "false" and to provide a justification for their response:

A. In a multiplication problem, the product is greater than either factor.

B. The product of .45 x 90 is less than 90.

C. In a division problem, the quotient must be less than the dividend.

D. In a division problem, the divisor must be a whole number.

E. The quotient for the problem 60/.65 is greater than 60.

F. The quotient for the problem 70 ÷ ½ is less than 70.

There were also questions that asked respondents to write expressions for word problems, as well as perform calculations that served as counterexamples to the statements above. This example demonstrates a class of problems slightly removed from actual classroom teaching itself, but that is not knowledge used exclusively in teaching. Few teachers would give the above examples A–F directly to students; yet a firm foundation in operations with rational numbers, and in mathematical proof and reasoning, is necessary for teaching children successfully. Building on this work, Tirosh and Graeber (1990) then explored whether cognitive conflict could be introduced to prompt prospective teachers to confront their misconceptions, in particular, the misconception that in division the quotient must be less than the dividend. Twenty-one respondents who correctly computed 3.75 ÷ .75 but agreed that "the quotient must be less than the dividend," were selected for interviews. The interview protocol was designed to point out the conflict between these two responses and to consider the source of this inconsistency. This study reinforced the predominant

findings of other research—i.e., that some teacher candidates do not adequately know the mathematics they will teach—but also used tasks and interviews as a means of studying teacher learning.

With the work of Ball (1990), the field turned more directly toward examining knowledge of mathematics that is specialized to teaching. Ball studied elementary and secondary prospective teachers' knowledge of division in three contexts: division with fractions, division by zero, and division with algebraic equations. The following example tasks required respondents to generate and explain representations, and interview probes were designed to gather information about their notions of what counts as a mathematical explanation:

1. People have different approaches to solving problems involving division with fractions. How would *you* solve this one:

   $1\frac{3}{4} \div \frac{1}{2}$

2. Sometimes teachers try to come up with real-world situations or story problems to show the meaning or application of some particular piece of content. This can be pretty challenging to do. What would you say would be a good situation or story for $1\frac{3}{4} \div \frac{1}{2}$—something real for which $1\frac{3}{4} \div \frac{1}{2}$ is the appropriate mathematical formulation?
3. Suppose that a student asks you what 7 divided by 0 is. How would you respond? Why is that what you'd want to say?
4. Suppose that one of your students asks you for help with the following:

   $$\text{If } \frac{x}{0.2} = 5, \text{then } x =$$

   How would you respond? Why is that what you'd do?

These four items illustrate different aspects of what Ball and her colleagues (Ball & Bass, 2003; Ball, Thames, et al., 2005) call specialized mathematical knowledge. Using fractions to calculate the answer to real-life problems is not uncommon among mathematically literate adults; however, constructing real-life problems to illustrate fractional computation is likely limited to teachers. Responding to students' questions about division by zero requires more than the knowledge that "this cannot be done"—it requires mathematical knowledge of the reason this is undefined. And the fourth item requires insight into how to make the procedures of algebraic manipulations comprehensible to learners, a different kind of understanding from simply being able to do the procedure oneself.

These items, written by Ball and others working on the Teacher Education and Learning to Teach project (TELT) (Kennedy et al., 1993), have been widely used in the field. Ma (1999), for instance, compared U.S. teachers in the TELT sample with Chinese teachers, analyzing responses to the items. She found that many Chinese teachers had a depth of understanding, for example how mathematical topics are related/connected, not present in U.S. teachers.

Another example of an instrument that taps specialized knowledge for teaching is described by Simon (1993), who investigated prospective teachers' knowledge of division using written open-response items and individual interviews. The questions were designed to assess the connectedness of prospective teachers' knowledge and their understanding of units:

1. Write three different story problems that would be solved by dividing 51 by 4 and for which the answers would be, respectively:

   a) $12\frac{3}{4}$ b) 13 c) 12

2. Write a story problem for which $\frac{3}{4}$ divided by $\frac{1}{4}$ would represent the operation used to solve the problem.
3. In the long division carried out as in the example below [the actual item shows the problem 715 divided by 12 being calculated with the standard long division algorithm], the sequence *divide*, *multiply*, *subtract*, *bring down* is repeated. Explain what information the *multiply* step and the *subtract* step provide and how they contribute to arriving at the answer.

Like Ball's items, these ask respondents to perform tasks that are unique to teaching: writing story problems that take into account the different interpretations of remainders, and explaining how portions of the standard long division algorithm work. The entire written instrument was administered to 33 prospective elementary teachers, and another eight were interviewed about a subset of the problems. Interviewees were asked to "think aloud as they solved the problem" and their comments were probed by the interviewer. Simon's results echoed previous studies, finding that, in general, prospective teachers' knowledge was fragmented and procedural.

The division tasks described above are examples of the types of written questions and interview prompts that have been used to better understand teachers' knowledge of mathematics. In developing the tasks, researchers had to grapple with questions about what mathematical knowledge is needed for teaching and how it might be assessed. It makes sense that the source of many such tasks were problems students encounter in K–12 schools, for teachers certainly must be able to solve the problems they give to their students. But the tasks also try to tap into mathematical knowledge needed for teaching beyond what students need to know, for example, knowing not only how to perform an algorithm, but also why it works and how it can be represented to students. Thus, designing these tasks

required researchers to reconceptualize what it means to know elementary mathematics in ways that are central to teaching and enabled scholars to develop a more refined conception of the mathematical knowledge needed for teaching.

The use of these tasks helped explicate what knowledge for teaching might look like—in essence, making the case for specialized, professional knowledge. These studies have also brought attention to problems with the quality of mathematical instruction delivered in the U.S. and can be used to inform the content and methods of teacher education courses. And, from a measurement standpoint, there are certain benefits to using these types of measures. Once tasks are designed, they are fairly low cost and easy to implement. There is also high face validity, at least as compared to standard mathematics tests and/or multiple-choice exams.

However, their use is also constrained in ways similar to direct observations. Analyzing or grading teachers' responses is, at scale, a lengthy and expensive process. Further, these tasks and interviews cannot be used to make inferences about particular individuals' (or even groups of individuals, we argue) level of knowledge, for they have seldom been studied to provide information regarding their validity, reliability, and generalizability. Contrary to some beliefs in mathematics education research, using an open-ended task or interview format does not free researchers from an obligation to examine the measurement qualities of the assessments used. Instruments with only a few mathematics problems, or with a very narrowly defined band of mathematics content, encounter the same generalizability problem that using only a handful of observations engenders, that of making a broad claim about knowledge from only a few samples of teachers' work. Open-ended tasks and interview probes are not always reliable, either; teachers' performance on one open-ended task designed by Ball—asking teachers to provide a story that represents 1¾ divided by ½—has been used as evidence that U.S. teachers lag in mathematical knowledge for teaching. A pilot of the multiple-choice version of this measure, however, suggests that part of the item fails to discriminate reliably between highly knowledgeable and less knowledgeable teachers.

Moreover, although these tasks are often situated in teaching contexts, they do not necessarily indicate how a teacher uses mathematical knowledge in practice. A teacher may perform well on the tasks in a clinical or instructional setting but not be able to access this knowledge during a lesson (Borko et al., 1992). Observations of teaching have revealed differences in teachers' mathematical knowledge that were not evident in interview tasks, such as their skill in selecting and using representations and their emphasis of key concepts during lessons (Leinhardt & Smith, 1985).

Assessing Teacher and Student Learning: A New Generation of Teacher Assessments

As researchers uncovered elements of mathematics knowledge unique to teaching, scholars and evaluators faced new measurement needs. One concerned the development of tools to test the relationship between mathematical knowledge for teaching and student achievement; the other involved the development of efficient tools to track change in teachers' mathematical knowledge over time. In both cases, researchers required tools that could be used to make valid and reliable inferences about individuals' or groups' mathematical knowledge for teaching. And researchers needed these tools at scale. Although they are not used for teacher certification, they are often used to deliver information about dozens if not hundreds of teachers, often at multiple time points.

These issues spurred the development of what many consider to be an unusual choice given the history of antipathy between professional educators and teacher assessments: pencil-and-paper tests. In some cases, such as the Study of Instructional Improvement/Learning Mathematics for Teaching (SII/LMT) measures (LMT, 2006b) and the SimCalc rate and proportionality teaching survey (Shechtman et al., 2006), these tests come in multiple-choice format. In other cases, such as the Knowledge for Algebra Teaching (KAT) measures (Knowing Mathematics for Teaching Algebra Project, 2006) and the Diagnostic Teacher Assessments in Mathematics and Science (DTAMS) measures (Bush, Ronau, Brown, & Myers, 2006) the instruments contain both multiple-choice and short open responses—often simply the answer to a computation problem, but sometimes short explanations for problems or procedures. All of the above tests, however, emphasize the ability to measure teachers' knowledge at scale, and with known reliability and validity. Teachers' answers to the items are marked as correct or incorrect, and a score is calculated, typically in standard deviation units. Thus these assessments are easily scored, with the ease of scoring a major factor in the selection of this format.

One reason for this selection was the scope of the research projects and intended uses of these measures. Having new insight into mathematical knowledge for teaching, scholars now turned to modeling its growth and contributions to the educational production function. For instance, the SII/LMT measures grew out of

a study designed to examine and compare the performance of elementary schools, teachers, and students participating in one of three whole-school reforms. Teacher learning was one potential target of these reforms, and also a mediator of student outcomes (Hill, Rowan, & Ball, 2005). Later, the instrument's authors received continued funding from the National Science Foundation to pilot additional items and extend the measures upwards to middle school. The SimCalc project began as a curriculum development project, one aimed at introducing the mathematics of change and variation in the early grades, typically with some technological component. The SimCalc instrument grew out of interest in teachers' learning as a result of engaging with their professional development and curriculum, and the desire to model student achievement as a function of teacher learning (Shechtman et al., 2006). KAT began as a project to explore the nature of knowledge used in teaching algebra and progressed to instrument development. Instrument developers hope to use these measures in models of student achievement. DTAMS is solely a measures development project, but has designed measures to be used not only to diagnose the status of middle school teachers' mathematical knowledge, but to track teacher learning and student achievement as a function of teacher knowledge.

This new generation of pencil-and-paper tests shares some common attributes. All, for instance, contain items intended to represent the mathematics problems encountered in teaching, rather than only "common" content knowledge, or knowledge of mathematics that is common across professions and available in the public domain. As such, items focus on teachers' grasp of representations of content, unusual student work or mathematical methods, student errors, mathematical explanations, teaching moves, and other raw materials of classroom mathematics. This sets these instruments apart from historical teacher tests, which focused only on asking teachers to solve the mathematics problems they would be teaching to students, or in some cases, mathematics problems somewhat more advanced than the curriculum they would be teaching. The new instruments also focus on relatively narrow mathematical domains, rather than returning an overall score across all mathematical content. And, critically in our view, all these tests have been developed by experts in mathematics, in the teaching of mathematics, and in psychometrics. This collaboration has led to explicit interest in the multidimensionality of the measures. Rather than attempting to write measures that represent "pure" mathematical or pedagogical knowledge, as many test development firms do for technical reasons, these measures embrace the idea that teachers' knowledge has many facets that must be included in any instrument. Project psychometricians, for their part, have been willing to work with this specification and, in turn, lent generous expertise in gauging the reliability and validity of these instruments. Reliability and validity analyses, which we shall discuss in more detail below, is relatively rare anywhere in the educational assessment enterprise (Messick, 1989).

One major difference among these pencil-and-paper assessments is in instrument goals and, by extension, the interpretation of scores. Two projects, KAT and DTAMS, criterion referenced their exams—DTAMS to nine policy documents describing both middle school student *and* teacher knowledge and KAT to a rubric that outlines the categories of teacher knowledge and the content involved in teaching algebra. By choosing to criterion reference and assuming defensible cutoffs for different performance levels, scores on this measure can be concretely interpreted vis-à-vis this standard—e.g., teacher Y knew X% of the mathematics needed according to standards. A third project, SII/LMT, initially designed its measures to discriminate accurately among teachers, in essence ordering them as correctly as possible relative to one another and to the underlying trait being assessed, mathematical knowledge for teaching. Although this design allows users to measure change over time as teachers learn, raw frequencies have no meaning relative to any criterion, such as "the mathematical knowledge teachers need to know." Finally, the fourth project, SimCalc, designed its measures to gauge curriculum-specific mastery of content—whether teachers grasp key principles in rate and proportionality, principles needed to teach using the SimCalc curriculum.

It is in the arena of underlying theory, however, that these instruments differ most. Despite claiming to cover roughly the same terrain, these projects have strikingly different approaches to specifying domains for measurement—in essence, different approaches to organizing what is "in" mathematical knowledge for teaching. SII/LMT has four knowledge domains; SRI has two; DTAMS has four; and KAT's conceptual framework crosses three aspects of knowledge for teaching and four mathematical domains. None of these conceptual maps match one another exactly, and none match Shulman's original formulation of pedagogical content knowledge. From one view, these differences might reflect ways the content itself (algebra, number) influences the knowledge needed for teaching. But at least to those interested in building theoretical coherence around mathematical knowledge for teaching, the variety of approaches is distressing.

Some examples from the actual assessments help elaborate these theoretical differences. Three-quarters of DTAMS' items measure important types of mathematical knowledge that teachers are charged with developing in students:

1. *Memorized facts and skills:* Which of the following numbers is the least common denominator of 5/9 and 7/12? (a: 108; b: 36; c: 72; d: 3)
2. *Conceptual understanding:* Name two numbers between 1.35 and 1.36. (a: $1.3\overline{5}$ and $1.\overline{35}$; b: 0.351 and 0.352; c: 1.345 and 1.354; d: There are no numbers between them.)
3. *High-order thinking:* Mr. Short is three paper clips in height. Mr. Tall is five buttons in height. Two buttons laid end-to-end are the same height as three paper clips laid end-to-end. Find the height of Mr. Tall in paper clips. Justify your solution.

These are relatively common problems in the "public domain," so to speak, of mathematical knowledge. Other than the fact that teachers are more likely than the general population to be solving such problems, none represent knowledge specific to teaching. One quarter of the items measure mathematically situated pedagogical knowledge:

1. A student said, "Whenever you add two non-zero integers the sum is always between the two integers." Explain why this is incorrect. Explain how you would use a diagram or model to help the student understand she is incorrect.
2. A student claims that all squares are congruent to each other because they all have four right angles. Why is this claim incorrect? Explain how you would help the student understand the error in her thinking.

These pedagogical items, which like the higher-order thinking items are in short-answer format and hand-scored, mimic the mathematical judgments teachers must make during classroom work—understanding common mathematical errors and designing instruction to correct them.

The SII/LMT items, by contrast, are in multiple-choice format and cover three domains of teacher knowledge, with one domain sub-divided into two constituent parts:

Content Knowledge:

Common content knowledge, or the mathematical knowledge teachers are responsible for developing in students:

Ms. Dominguez was working with a new textbook and she noticed that it gave more attention to the number 0 than her old book. She came across a page that asked students to determine if a few statements about 0 were true or false. Intrigued, she showed them to her sister who is also a teacher, and asked her what she thought.

Which statement(s) should the sisters select as being true? (Mark YES, NO, or I'M NOT SURE for each item below.)

	Yes	No	I'm not sure
0 is an even number	1	2	3
0 is not really a number. It is a placeholder in writing big numbers.	1	2	3
The number 8 can be written as 008.	1	2	3

Items in this category are similar to DTAMS' memorized facts and skills and conceptual understanding. They tap knowledge that is in the public domain, including that used in other professions (e.g., writing 8 as 008 as in computer science).

Specialized content knowledge, or mathematical knowledge that is used in teaching, but not directly taught to students:

Imagine that you are working with your class on multiplying large numbers. Among your students' papers, you notice that some have displayed their work in the following ways:

Student A	Student B	Student C
35 ×25 125 +75 875	35 ×25 175 +700 875	35 ×25 25 150 100 +600 875

Which of these students would you judge to be using a method that could be used to multiply any two whole numbers?

	Method would work for all whole numbers	Method would *not* work for all whole numbers	I'm not sure
Method A	1	2	3
Method B	1	2	3
Method C	1	2	3

Here, the teacher must inspect the three student methods to determine what, in fact, is occurring. Student A, for instance, has multiplied 5×25 and then 30×25; student B is using a version of the standard U.S. algorithm; and student C has used a partial product method. If the teacher understands these explanations for the students' methods, she must then make a determination about whether each method generalizes to the multiplication

of other whole numbers, perhaps by referencing the commutative or distributive properties of multiplication. This work is entirely mathematical, but it is not mathematical work done by many non-teaching adults.

Knowledge of Content and Students

Knowledge of content and students, or the amalgamated knowledge that teachers possess about how students learn content:

Takeem's teacher asks him to make a drawing to compare $\frac{3}{4}$ and $\frac{5}{6}$. He draws the following:

and claims that $\frac{3}{4}$ and $\frac{5}{6}$ are the same amount.

What is the most likely explanation for Takeem's answer? (Mark ONE answer.)

a) Takeem is noticing that each figure leaves one square unshaded.

b) Takeem has not yet learned the procedure for finding common denominators.

c) Takeem is adding 2 to both the numerator and denominator of $\frac{3}{4}$, and he sees that that equals $\frac{5}{6}$.

d) All of the above are equally likely.

Here, a teacher must look beyond the standard method for comparing fractions (finding common denominators) to see that Takeem has focused on the amount that is missing from each whole. This knowledge is likely held by teachers who have worked with children learning to compare fractions; it also corresponds to Shulman's "common student misconceptions" portion of pedagogical content knowledge.

Knowledge of Content and Teaching

Knowledge of content and teaching, or mathematical knowledge of the design of instruction, includes how to choose examples and representations, and how to guide student discussions toward accurate mathematical ideas.

While planning an introductory lesson on primes and composites, Mr. Rubenstein is considering what numbers to use as initial examples. He is concerned because he knows that choosing poor examples can mislead students about these important ideas. Of the choices below, which set of numbers would be <u>best</u> for introducing primes as composites? (Mark one answer.)

	Primes	Composites
a)	3, 5, 11	6, 30, 44
b)	2, 5, 17	8, 14, 32
c)	3, 7, 11	4, 16, 25
d)	2, 7, 13	9, 24, 40

e) All of these would work equally well to introduce prime and composite numbers.

In this item, the teacher must first know that students are likely to think that all prime numbers are odd; including 2 is important given this misconception. Students are also likely to think that odd numbers cannot be composite—making choice (d), which tests both misconceptions, the best of the four options. This SII/LMT category is the furthest from pure mathematical knowledge, as it involves reasoning about both students and teaching. It corresponds with another component of Shulman's pedagogical content knowledge (choosing the "best representation") and to DTAMS' pedagogical knowledge category.

The schematic for item development used by the Knowledge of Algebra for Teaching is more detailed, consisting of 24 separate cells that are the product of two algebraic content areas, three types of algebra knowledge for teaching, and four domains of mathematical knowledge. The first side of this three-dimensional matrix names two central algebraic topics: expressions, equations, and inequalities; and functions and their properties. The second side names algebra knowledge for teaching, which in this view consists of knowledge of school algebra, or the mathematics in the intended curriculum that *students* should learn; advanced knowledge, or college-level mathematics that gives teachers perspectives on the trajectory of mathematics beyond middle or high school; and teaching knowledge, or mathematical knowledge beyond that directly taught to students, but which is useful in instruction. This last category includes the mathematics involved with student misconceptions, knowledge of materials and texts, and ways to develop mathematics within and across lessons with particular goals in mind. Finally, the third side of the matrix encompasses four domains of mathematical knowledge, including "core" content knowledge, or declarative or substantive knowledge; representations of mathematical content such as number lines, tables, graphs, and area models as well as the materials that can be used to represent content during instruction (e.g., algebra tiles); applications and contexts, such as situations that can be modeled by linear functions; and reasoning and proof, or knowledge of how truth is established in the discipline.

Four items illustrate aspects of this domain specification. The first item comes from the category of school algebra, or the topics that teachers would directly teach students. It illustrates an "application" problem from KAT's four domains of mathematical knowledge, modeling the ways in which exponential functions might be used in real-life situations:

Which of the following situations can be modeled using an exponential function?

i. The height *h* of a ball *t* seconds after it is thrown into the air.

ii. The population *P* of a community after *t* years with an increase of *n* people annually.

iii. The value *V* of a car after *t* years if it depreciates *d*% per year.

A. i only

B. ii only

C. iii only

D. i and ii only

E. ii and iii only

This problem is similar, perhaps, to those found in high school or college textbooks. This item would fall in the SII/LMT common content knowledge category, and in DTAMS' conceptual knowledge category.

The next KAT item is from the category of advanced knowledge or knowledge beyond the actual middle or high school curriculum, and focuses on functions and their properties:

For which of the following sets S is the following statement true?
For all a and b in S, if ab = 0, then either a = 0 or b = 0.

i. the set of real numbers

ii. the set of complex numbers

iii. the set of integers mod 6

iv. the set of integers mod 5

v. the set of 2 × 2 matrices with real number entries

A. i only

B. i and ii only

C. i, ii, and iv only

D. i, ii, iii, and iv only

E. i, ii, iii, iv, and v

Answering this item properly requires an important type of mathematical reasoning. Teachers must be sensitive to the fact that familiar properties from algebra are not necessarily true in general, but, rather, depend on the set being referenced. In the item above, the statement *if ab = 0, then either a = 0 or b = 0* is true when working with all the choices save for the set of real 2 × 2 matrices and integers mod 6.[4] So although integers mod 6 is not in the K–12 curriculum, teachers must not lead high school students to believe that the original statement is universally true. This category is unique to the KAT measures and represents one very traditional line of thinking about the mathematical knowledge teachers need—that a teacher should know "very much more than he is expected to teach" (Michigan Department of Public Instruction, 1898, p. 10; see also Education Trust, 1999, p. 4).

The third item is open-ended, tapping "teaching knowledge" of student errors, and is also classified as "core content knowledge" by the KAT project:

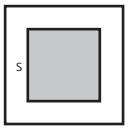
A student solved the equation 3(n – 7) = 4 – n and obtained the solution n = 2.75.

What might the student have done wrong?

To answer correctly, teachers must know how to correctly solve this equation for themselves, and they must also know a common student error, namely that students will forget to distribute the 3 over the 7 as well as the n. This corresponds to the SII/LMT category named knowledge of students and content.

The last item is also open-ended and draws from the "teaching knowledge" category, focusing on representations of expressions:

Hot tubs and swimming pools are sometimes surrounded by borders of tiles. The drawing at the right shows a square hot tub with sides of length s feet. This tub is surrounded by a border of 1 foot by 1 foot square tiles.



How many 1-foot square tiles will be needed for the border of this pool?

a. Paul wrote the following expression:

$$2s + 2(s + 2)$$

Explain how Paul might have come up with his expression.

b. Bill found the following expression:

$$(s + 2)^2 - s^2$$

Explain how Bill might have found his expression.

c. How would you convince the students in your class that the two expressions above are equivalent?

Here, teachers must "see" how the same representation might yield very different expressions, mapping closely between the representation and the construction of each expression. Teachers must also make a judgment about how best to prove these two expressions are equivalent, a question that grapples with the complex intersection between content, students, and teaching. This item would fall, in the SII/LMT scheme, into the "specialized" and "knowledge of content and teaching" categories.

[4] For example, when working with the set of integers mod 6, 4 times 3 equals 12, which is 0 mod 6, but neither 4 nor 3 is equal to 0 mod 6.

Finally, the SimCalc materials have no similar domain map, instead using a domain map for *student* learning of their "mathematics of change" curriculum materials. Nevertheless, their items do reveal a theory of teacher knowledge. The item below, for instance, asks teachers to analyze an unusual solution method—and in so doing, to draw on underlying knowledge of proportional relationships as represented in such a table:

Students are working on creating tables of proportional relationships. During the lesson, Mr. Lewis has the following table up on the board.

x	y
3	12
2	8
4	16
8	32

He asks the students to come up with another (*x,y*) pair that could fill in the blank cells. One student says that you can create a new (*x,y*) pair by taking the averages of the last two rows in the table. That is, the new *x* is

$$\frac{4+8}{2}=6,$$

and the new *y* is

$$\frac{16+32}{2}=24.$$

Mr. Lewis had hoped that the students would pick a value for *x*, then multiply by 4 to get a value of *y*. He is intrigued by this student's solution to the problem, however, and wonders whether it will always work.

This method will produce a valid *x,y* pair: (Mark ALL that apply.)

a) Only when *x* and *y* are always whole numbers.

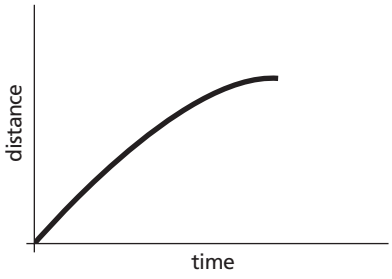
b) Only when the last two rows have even numbers.

c) Only when the last row is exactly double the next-to-last row.

d) Always, as long as the table represents a proportional relationship.

e) Always, as long as the table represents a linear function.

In addition, because this novel curriculum asks students to explain, interpret, and represent linear functions, proportional relationships, and rates, teachers are often asked to do the same:



Ms. Chopra asks her students to explain why the graph above is NOT proportional. Her students' explanations are given below.

Which of the following student statements explains why the graph is NOT proportional? (Mark ALL that apply.)

a) You can't have a curve in a graph.

b) The relationship between the distance and time is not always constant.

c) The graph is not labeled correctly.

d) The ratio of rise over run is different at different points on the curve.

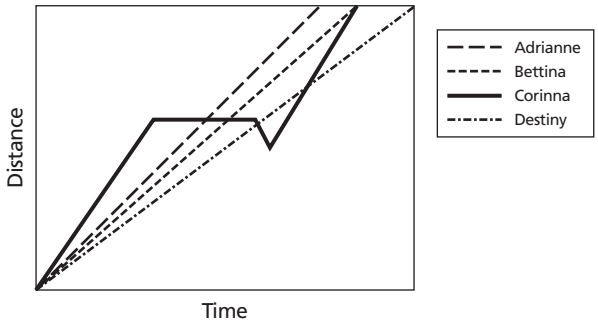
e) At every point on the curve, the d/t is not always the same ratio.

f) None of the above.

This item blurs the distinction between common and specialized knowledge set out by the SII/LMT instrument developers; here, both students and teachers are asked to know mathematics in ways that probe the underlying meanings and to provide mathematical explanations to counter a common misconception.

However, many more of the SimCalc items ask teachers to solve the same types of problems they would be asking students to solve:

Mr. Aneesh presents the following graph to his class. The graph represents the entire race run by Adrianne, Bettina, Corinna, and Destiny.



Mr. Aneesh asks the students to describe what is happening during the race.

Which of the following student statements are FALSE? (Mark ALL that apply.)

a) Corinna and Bettina finished at the same time.

b) All runners finished the race at the same time.

c) Adrianne had the highest average speed.

d) Adrianne led the race from start to finish.

These four projects have progressed the furthest in instrument development, in the sense that all have taken care to specify domain maps, completed pilot work with large groups of teachers, conducted psychometric analyses, and begun work on validation of the measures. However, many other pencil-and-paper tests exist. Most faculty responsible for teaching mathematics content and methods courses to preservice teachers, for instance, write end-of-unit or end-of-course assessments. Research projects interested in studying

the development of teacher knowledge design their own instruments, tailored to their own program or purpose. Schweingruber and Nease (2000) describe a content knowledge assessment used to gauge learning in a teacher professional development program at Rice University. Phillip et al. (in press) describe and make available a test that combines an assessment of preservice elementary teachers' content knowledge with their knowledge of students. In it, preservice teachers are presented with student work, asked whether that work is correct and, if not, to explain the student error. Some of the correct student work presents unusual mathematical solutions or representations, such as non-standard algorithms or unusual representations of number and content. Many of the student errors are drawn from the literature on student learning and error analysis.

As this review shows, the field has responded strongly to the need for pencil-and-paper assessments that can be used to study teacher and student learning. Many build on the types of items and mathematics described in the initial work on teachers' knowledge identified in the beginning of this section, and thus come closer to tests of professional knowledge than have past assessments, particularly those used in the educational production function literature. Although development is not complete on any of these assessments, and such assessments are far from perfect vehicles for this work, significant progress has been made. One reason for this is the interest on the part of funders, particularly the federal government, in developing such measures; three of the projects described in detail above were funded by the National Science Foundation, and one by the U.S. Department of Education. Measures development is expensive, and a commitment on the part of major funding agencies is critical. Another reason might be a climate of increased accountability, not only for districts, schools and students, but also for the professional development providers who serve these educators.

Like any form of assessment, the instruments reviewed here have significant drawbacks. To start, few would agree that instruments that use multiple choice or short-answer responses can represent completely the mathematical knowledge teachers use in their work. First, the construct of mathematical knowledge for teaching is still emergent. Without better theoretical mapping of this domain, no instrument can hope to fully capture the knowledge and reasoning skills teachers possess. Second, teaching any subject matter involves complex professional judgments, and such judgments are not always amenable to testing in a format with single correct answers. While there may be some regularities in student errors, for instance, the errors of any particular student are shaped by his or her prior learning experiences and the task at hand; either may lead to idiosyncrasies. And while base ten materials may generally work well to represent multiplication of decimals, whether they will work well in a particular situation depends in part on the teacher's skill in using them, students' exposure to the use of base ten materials to model whole number operations, and so forth. Teachers' reasoning about these situations is not well reflected in right/wrong formats. Problems that draw only on mathematics avoid some of these complications, but also fail to gauge other aspects of knowledge and reasoning that are key to classroom teaching.

A third problem with multiple-choice/short-answer assessments involves teachers' reactions to such instruments. Several of the projects described here note teacher resistance to assessments and, in particular, to multiple-choice forms of assessments. Teachers, whose work is about helping others learn, are well aware of the problems inherent in assessing complex learning outcomes. They know that many important achievements are difficult to measure, and that multiple-choice formats are often inadequate in capturing learning. Using multiple-choice formats to assess teachers' own knowledge suggests a limited view of the complexity of professional knowledge and thus conflicts with professionals' reasonable experience with and distrust of test questions framed in this form.[5]

Fourth, validity work on these tests has generally lagged behind assessment development itself. Test developers are most likely to have conducted standard psychometric analyses, such as construct identification via factor analyses and reliability analyses. Reliabilities for these assessments meet or exceed industry standards for the specific test goals listed, such as diagnosing teacher knowledge or tracking groups of teachers over time. However, for the most part, validation work has been limited. The DTAMS and the SII/LMT projects both conducted content validity checks, ensuring that the test reflected key elements of the mathematics curriculum for students, and elements of teacher knowledge suggested by the literature described above. KAT is currently conducting a discriminant validity study, giving their assessment to college mathematics majors, for instance, to see whether this group scores high on advanced mathematics knowledge but low, as expected, on teaching mathematics

[5] That most multiple-choice tests are disappointing in their capacity does not rule out the possibility of designing new types of items that take advantage of the format rather than distorting the nature of the knowledge being assessed.

knowledge. Construct identification, however, is only one form of validity. Another important form is predictive—whether teachers' performance on the assessment is related to the mathematical quality of instruction, and to student learning. Currently, only the SII/LMT test developers have performed this work, and even there, more work remains to be done.

Finally, using multiple-choice/short answer assessments as part of an evaluation or study requires at least basic statistical expertise—a t-test or ANOVA to examine teacher gains, for instance. In some cases, such as the SII/LMT measures, scores are returned not in raw frequencies but in standard deviation units. Users also have the option of creating their own forms from the item pool. This contrasts with the analytic demands of many of the other methods described here. Eventually, many of these pencil-and-paper assessments may be put online, a condition under which these assessments will likely be scored automatically by computer.

Other Methods

New technologies may also allow improvements on pencil-and-paper assessments, at least for small-to-moderate scale research projects. Kersting (2004) and associates at LessonLab have developed an online measure that gauges teacher knowledge by examining responses to ten short video clips of teaching. Teachers' responses are credited according to the amount they discuss features of the mathematics, student thinking and understanding, teaching strategies, and according to the amount that the discussion of the clips embeds these observations in a cause/effect framework. Like interviews or open-ended tasks, this method may be cumbersome to score. However, the technical properties of the instrument appear promising, and the assessment format itself holds three appealing features. First, it complements current professional development methods, particularly the use of videotape for teacher analysis and reflection. Second, an analysis of mathematics teaching may be less threatening to teachers than a pencil-and-paper assessment. Finally, the assessment can be embedded in the flow of teachers' work in a video-based professional development setting.

Other embedded assessment techniques also exist. For instance, Sowder, Phillip, Armstrong, and Shappelle (1998) effectively used discourse analysis—or studying what teachers say and how they talk about mathematics—to trace growth in teachers' knowledge of division of fractions. One benefit of this method is that it requires no statistical expertise. Discourse analysis, however, does require close and specialized training. It is a promising approach because teachers' use of mathematical language is both an indicator and target of growth in knowledge and skill; teachers need, in classrooms, to be able to use mathematical terms accurately and precisely. But more, they also need to be able to "translate" between children's home language, informal mathematical language, and disciplinary language (Ball, Masters-Goffney, & Bass, 2005; Pimm, 1987). Our own observations suggest that this linguistic capacity may be held variably in today's teaching population.

The drawbacks to discourse analysis include a lack of generalizability past the particular professional development (or other) setting that participants are in and difficulties separating individuals' skill with language from group skill with language. Further, such measurements do not scale easily, and they require finely calibrated frameworks for interpreting and analyzing discourse patterns. That said, this remains an intriguing possibility, one that should be investigated more formally.

Finally, some key studies have used self-reports of improvements in knowledge and skill as proxy measures for actual growth (e.g., Garet, Porter, Desimone, Birman, & Yoon, 2001; Horizon Research, 2002; NCES, 1999). However, the degree to which these are related to actual teacher knowledge growth is largely unknown.

In sum, scholars have developed a variety of methods for studying mathematical knowledge for teaching, from those intended to help understand the construct itself to those designed to provide evidence of teachers' learning in this domain. Our review of these methods has taught us several things. First, developers in this field report that there is in fact considerable interplay between assessment and theoretical development. Writing test items, for instance, helps illuminate aspects of mathematical knowledge specific to teaching; conducting a discourse analysis can help illuminate the understandings teachers have of students and mathematics. Second, every method has both advantages and drawbacks; there is no one perfect method, although there may be better and worse methods for particular research questions. In fact, more diversity in assessments and assessment formats should be encouraged, to allow better matching between the subjects of inquiry and instruments used.

Third, these assessments should be widely available, and widely used. This requires broader support than exists now in the mathematics education community. For instance, there is currently no clearinghouse through which scholars "shopping" for a measure can collect and compare information about each measure's content, design, and technical features.

Also, there is no systematic method for disseminating measures or for training individuals in their use. Instead, scholars and evaluators sit at the mercy of the research projects and agencies developing these measures, whose capacity to disseminate varies according to funding and the availability of personnel. But broader use of these measures in the study of teacher knowledge and learning would surely benefit the field, as results from different research projects would be more directly comparable and, we hope, more rigorously evaluated.

Finally, each of the efforts described here have made an attempt to bring the measurement of teachers' mathematical knowledge closer to the actual practice of teaching itself. In many cases, the developers of these assessments explicitly state their desire to measure the mathematical knowledge used in teaching, although the frameworks for what those measures might be vary. One would expect that this improves the validity of the assessments, although as we noted above, most projects do little predictive validity work on their measures. A similar effort to move the assessment of mathematical knowledge closer to actual teaching has been occurring in the field of teacher certification testing, a topic to which we now turn.

CONTEMPORARY APPROACHES TO TESTING TEACHERS FOR CERTIFICATION

In the last decades of the twentieth and into the twenty-first centuries, three elements converged to spur the development of new forms of teacher certification tests. First, policy-makers and the public made increasing demands for accountability on the part of teachers and teacher educators; as we have seen, the amount of required teacher testing grew markedly over the period 1980–2000. Second, mathematics educators began to understand the nature of mathematical knowledge for teaching, investing heavily in the idea that beyond knowing content, teachers must also have profession-specific mathematical knowledge about content, students, and teaching. Finally, this era saw emergent interest on the part of professional educators in testing teachers; although opposition to teacher testing still runs high in some quarters, at least two national professional organizations have been established to promote new forms of teacher assessment. In this way, the historic antipathy between teaching and teacher educator organizations and testing agencies has begun to crumble. Whether the professional organizations can gain control over widespread teacher testing, and whether the tests produced are satisfactory, remains to be seen.

In this section, we describe in detail the mathematical elements of the three most prominent of these efforts. Two of the assessments are for beginning teachers: the Praxis Series (the successor to the NTE), and the Interstate New Teacher Assessment and Support Consortium's (INTASC) portfolio assessment system. Although INTASC is not now formally in use for certification in any state, Connecticut's second-stage licensure system uses a portfolio assessment based on INTASC principles, and INTASC is looked to as a leader in the field. The third assessment is an endorsement for experienced practitioners: the National Board for Professional Teaching Standards (NBPTS) certification. We examine each of these assessments along four dimensions: its history and development; a description of the test and how it is used; an analysis of sample mathematics items; and, based on our analysis of available items, some conjectures about the test's implicit views of teaching, of mathematics, and of the mathematics needed for teaching. We also briefly review other tests used in the U.S., including the California Subject Examinations for Teachers (CSET), and the tests developed by the American Board for Certification of Teacher Excellence (ABCTE). Each assessment plays a unique role in the current debates over how to certify teachers.

Before beginning, it is important to note what these modern certification tests are designed to do. Unlike the mathematical assessments common in the scholarly literature, the tests discussed below are designed to allow the user to draw inferences about a candidate's suitability for teaching. As this implies, there are often great stakes attached to teachers' performance on these exams: pay raises, at a minimum, and more often the ability to compete for jobs in one's chosen profession. The high-stakes nature of many of these assessments helps explain why there has been so much debate over them and so much effort devoted to ensuring their accuracy. We begin with Praxis, the most commonly used teacher certification assessment in the U.S.

The Praxis Series

The Praxis Series began to replace the National Teacher Examination (NTE) in 1993. Currently, 43 states use at least one portion of the series for beginning teacher certification (ETS, 2006a). In the wake of increasing calls for teacher testing, and building on

new advances in psychology and research on teaching, Praxis was to take a more robust view of teaching and the knowledge needed to teach. Praxis takes a three-stage approach to teacher certification that includes tests of prerequisite content knowledge, subject-specific content knowledge and pedagogical knowledge, and interactive teaching skills. Each construct in Praxis would be measured separately, and at different stages in a new teacher's career.

In many states, teachers begin by taking Praxis I, a multiple-choice test of general academic knowledge, before admission to teacher education programs. This test is similar to the SAT and in some states the tests are used interchangeably. If the study guide questions are representative, the mathematics questions in Praxis I are similar to the SII/LMT "common content knowledge" category, meaning mathematics that is common across professions and, in many cases, in the public domain. This makes sense, given that this assessment is administered prior to the start of formal teacher education programs.

Praxis II is a range of multiple-choice and constructed-answer assessments that test general academic and pedagogical knowledge as well as subject-specific knowledge for teaching. Praxis II is generally taken at the completion of a teacher education program. Elementary teachers might see mathematics questions on one of three separate tests: the content knowledge test for elementary teaching, a widely used assessment, containing roughly 30 multiple-choice items on mathematics; the curriculum, instruction, and assessment test, where roughly 20% of 110 multiple choice items ask about mathematics; and the content area exercises, a test that includes as one of its four prompts a problem that asks teachers to design mathematics instruction or respond to a student error. At the secondary level, ETS maintains several mathematics-specific Praxis II assessments, including content knowledge; proofs, models, and problems; general mathematics; and pedagogy. The mathematics pedagogy assessment is similar to the elementary content area exercise test, in that it asks teachers to respond to prompts about how best to design or implement mathematics instruction.

Praxis III is an inventory of teaching skills that is conducted by observing the beginning teacher in her classroom. Only two states are currently using Praxis III, and none require it for initial licensure. Praxis III is used instead in some districts and states for professional development.

Each state that uses Praxis requires its own combination of tests, and each state determines its own passing rates for licensure. So, for example, Idaho requires that its teacher candidates take a single subject-area Praxis II test for licensure; Hawaii requires the Praxis I test and two Praxis II tests, the Principles of Learning and Teaching as well as a subject-matter area test. Because of the prevalence of Praxis II, and its claim to measure subject-matter-specific disciplinary knowledge, we focus on it for much of the discussion below.

Praxis II, like other modern teacher certification tests, stakes its claim to validity on the idea that it captures the knowledge needed to successfully teach. Two aspects of assessment design buttress this claim. First, the Praxis II tests were developed through a process designed to include key stakeholders in the education and certification of teachers. In mathematics, this included subject-matter specialists (mathematicians, mathematics educators), teachers, mathematics coordinators in school districts, psychometricians, and psychologists. Second, the items on the assessment were developed following what the test developers call a "job analysis," or a review of the mathematical knowledge teachers need in their jobs. This "job analysis" is not an ethnographic study of a mathematics classroom with special focus paid to the teacher and the knowledge employed in teaching. Rather, by "job analysis," the developers mean that subject-matter specialists examined earlier tests such as the NTE, textbooks, the NCTM Standards documents, curricular documents, then drew up lists of topics that teachers would need to know for teaching mathematics. This list was then shared with teachers and other school personnel who made suggestions and changes and, when approved, translated into assessment items.

Because ETS does not release active Praxis items, it is difficult to characterize the results from this process with any degree of certainty. If Praxis study materials are any guide, however, the problems prospective teachers face on the content area tests are largely dependent on the Praxis II assessment their state requires. For both elementary and secondary content knowledge tests, among the most commonly used, the items resemble those that have comprised teacher assessments for generations. The following items are taken from study guides and test preparation materials (ETS, 2005b, p. 11; ETS, 2006c, p. 4):

1. $15(4 + 3) = 15 \times 4 + 15 \times 3$

The equation above demonstrates which of the following?

a) The distributive property of multiplication over addition
b) The commutative property of multiplication
c) The associative property of multiplication
d) Additive inverse and additive identity

2. Riding on a school bus are 20 students in 9th grade, 10 in 10th grade, 9 in 11th grade, and 7 in 12th grade. Approximately what percent of the students on the bus are in 9th grade?

a) 23% b) 43% c) 46% d 76%

3. Which of the following is equal to 8^4?

a) 4,032 b) 4,064 c) 4,096 d) 4,128

4. For which of the following values of k does the equation $x^4 - 4x^2 + x + k = 0$ have four distinct real roots?

I. –2
II. 1
III. 3

a) II only b) III only c) II and III only d) I, II, and III

Items 1–3 are from elementary materials; item 4 is from secondary test preparation materials. These items ask teachers to solve problems that contain mathematical content common to the public and other professions, including mathematics as a discipline. And they are not far from the types of items found in most teacher certification exams for the past two centuries, with changes in mathematical content and item design to allow the tests to stay current with the content taught in elementary and secondary schools.

By contrast, the small number of mathematics items on the elementary curriculum, instruction, and assessment test appear aimed at assessing prospective teachers' ability to work with the mathematics in instruction (ETS, 2005a, pp. 5–6):

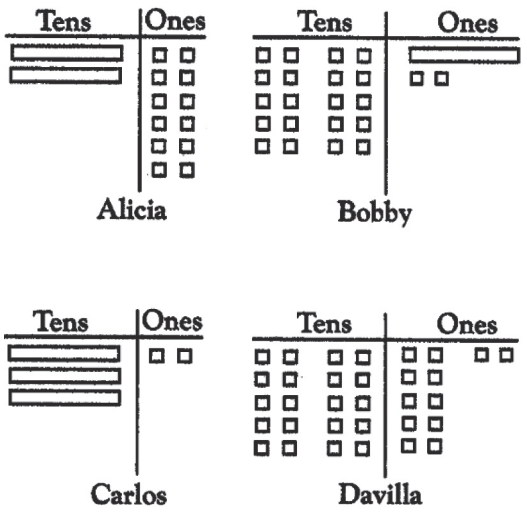
1.

$$\frac{4}{16} - \frac{1}{8} = \frac{3}{8} \qquad \frac{5}{9} - \frac{1}{2} = \frac{4}{7} \qquad \frac{7}{16} - \frac{1}{5} = \frac{6}{11}$$

The examples above are representative of a student's work. If the error pattern indicated in these examples continues, the student's answer to the problem 9/11 – 1/7 will most likely be

a) $\frac{10}{4}$ b) $\frac{8}{7}$ c) $\frac{8}{4}$ d) $\frac{9}{18}$

2.



The illustrations above show how four students, Alicia, Bobby, Carlos, and Davilla, used base ten blocks to represent the number 32. Which of the students used the blocks to represent the number 32 in a way that reveals an understanding of the underlying concepts of a numeration system based on powers of ten?

a) Alicia b) Bobby c) Carlos d) Davilla

The first item asks teachers to recognize or reason through students' errors; the second asks about how materials represent concepts. Taken together, these items seem to draw from the kinds of thinking described in Shulman's "pedagogical content knowledge" and the finer subdivisions of this construct by modern academic test developers (e.g., KAT's "teaching knowledge" or SII/LMT's "knowledge of content and students" and "knowledge of content and teaching"). These items are also the modern version of those on the NTE's elementary school specialist area test.

Note that the second item again illustrates the difficulty of writing defensible multiple-choice items in this arena. According to test materials, Carlos' representation is best. However, Alicia might also have a solid understanding of base ten numeration, as evidenced by the fact that she showed 2 tens and 12 ones—a representation of 32 commonly used for subtraction with regrouping. In fact, many teachers use base ten materials to explicitly teach the equivalence of Carlos' and Alicia's representations of 32, and we hazard that some educators might go as far as to argue that Alicia has a better understanding than Carlos for particular purposes (e.g., subtraction with regrouping). Although the presence of this item on Praxis released materials might reflect a simple mistake, it may also indicate a significant flaw in the construction of this assessment, in that we cannot imagine many mathematics educators would approve of such

an item. If mathematics educators do not have a voice in the construction of this assessment, it again calls into question the degree of professional control over the assessment. It also harkens to the precursor to the Praxis II, the National Teacher Examination, which carried other such ambiguous items.

Moving further toward practice, two Praxis II tests (elementary content area exercises and secondary mathematics pedagogy assessment) ask teachers to answer essay questions involving mathematical scenarios like the following:

> You are teaching a unit on solving quadratic equations. You have already taught the students how to solve quadratics by taking square roots and by factoring. In your next lesson, you plan to teach the students how to solve quadratic equations by completing the square.
>
> Design a homework assignment for your students to complete after the lesson on solving quadratic equations by completing the square. The homework assignment should consist of 5 problems that review previously taught skills and concepts while also providing practice in the newly introduced material.
>
> Briefly explain your rationale for including the skills and concepts that the problems illustrate.
>
> (ETS, 2003, p. 106)

Here, and in every other sample secondary-level prompt available to us, teachers must engage in lesson design. Other prompts include "Describe an investigation you would have your students make to discover the difference between 'regular' and 'equilateral' [polygons]. Your investigation may involve the use of any type of manipulative or any software package" (ETS, 2003, p. 114). And "Describe a strategy, using pictures or manipulatives, that you could use to help foster the students' conceptual understanding of equivalent fractions" (p. 110). On their face, these tasks do appear to tap mathematical knowledge as it is used in teaching; responses are scored based on not only demonstrated understanding of the mathematics, but also on the basis that the instruction described therein is "very likely to achieve the desired goals" (ETS, 2005, p. 2). However, it does so in a particular way, assuming that good teaching equates with constructing lessons from scratch. Left out in this type of assessment is the kind of mathematical knowledge used when adapting curriculum materials, or using these materials well.

Finally, like some of the professional organizations we describe below, the Praxis has moved into the realm of measuring teaching in the classroom. The Praxis III assessment, now used in two states, evaluates teachers during their first and second years in the classroom via observations and interviews by trained assessors. It assesses knowledge in 19 criteria in four interrelated domains: organizing content knowledge for student learning; creating an environment for student learning; teaching for student learning; and teacher professionalism. Although not specific to mathematics instruction, the prominence of content knowledge is noteworthy. The description of this domain begins with the following: "Knowledge of the content to be taught underlies all aspects of good instruction. Domain A focuses on how teachers use their understanding of students and subject matter to decide on learning goals, to design or select appropriate activities and instructional materials, to sequence instruction in ways that will help students to meet short- and long-term curricular goals, and to design or select informative evaluation strategies" (Dwyer, 1998, p. 182). Nonetheless, the specifications for content knowledge observed in Praxis III are generic, so how and whether teachers' mathematical knowledge for teaching is observed and appraised remains an open question. Whether mathematical knowledge is closely evaluated also depends in large part upon whether the observer is knowledgeable in this area.

Interstate New Teacher Assessment and Support Consortium (INTASC) Portfolio Assessments

Another effort to reformulate teacher certification has taken place under the auspices of the Interstate New Teacher Assessment and Support Consortium (INTASC), a consortium of state education agencies and national educational organizations sponsored by the Council of Chief State School Officers. INTASC was created in 1987 to provide a forum for states to learn collaboratively about and develop new accountability requirements for teacher preparation programs, new assessments for teacher licensing and evaluation, and new professional development programs. Following the publication of its *Model Standards in Mathematics for Beginning Teacher Licensing & Development: A Resource for State Dialogue* (1995), INTASC began to develop standards-based licensing tests for teachers. On the view that no single test could adequately assess a prospective teacher's ability to teach, INTASC recommended three types of licensing tests: 1) a test of content knowledge; 2) a test of teaching knowledge (pedagogy); and 3) an assessment of actual teaching. The first two types of tests—content knowledge and teaching knowledge—are intended for the conclusion of teacher preparation programs as a requirement for a provisional licensure.

To qualify for a permanent license, teachers would complete a third, performance-based test in the initial years of teaching. INTASC used a portfolio format for its assessment, "a collection of documents that tell the story of a candidate's teaching as it develops over a period of time" (INTASC, 2006). Portfolios had been growing in use as K–12 assessments through the 1980s, and their extension to teacher assessments made sense: they can provide evidence of the development of a teacher's practice and include instructional materials, student work samples, video records of teaching, and written commentaries. Like their model standards, INTASC's portfolio assessments are intended to be a resource that can be adopted or adapted, for example, by schools of education as a requirement for program completion, or by states as a requirement for certification or completion during the induction years. We focus on INTASC's portfolio assessment both because of its novel format and its view of teacher knowledge.

The INTASC mathematics teaching portfolio documents approximately 8 to 12 hours of a beginning teacher's practice, organized around a central mathematical topic. Portfolio entries include a description of the teaching context; a sequence of lesson plans highlighting central features of the lessons such as the mathematics content, tasks, and accommodations for three focal students; two featured lessons each documented with 20–30 minutes of video, assessments, and student work samples; a cumulative student assessment and scoring criteria, accompanied by student work marked with feedback; and an analysis of teaching and professional growth. Along with these records of practice, each entry also requires a written commentary with a rationale for and evaluation of the teacher's instructional choices. The following excerpt describes part of the commentary required for the series of lessons:

> **The Mathematics:** Identify your goals and expectations for student learning across this series of lessons. Begin your commentary with a statement of these goals. Continue with a description of the broad mathematical concept(s) or idea(s) that unifies the lessons. Why is this concept or idea important to the study of mathematics? How do the lessons you have designed build a cohesive set of plans to address this concept or idea? Explain the mathematical connections across the lessons. Provide one or two examples of how the series of lessons builds on mathematics the students have already learned. Give one or two examples of how the series of lessons serves as a foundation for the mathematics students will learn later.
>
> Provide a general description of your goals and expectations for students to reason mathematically, solve problems, communicate mathematically, and see connections within mathematics, connections to other disciplines, and connections to the real world across these lessons. Provide specific examples of how your lessons will promote reasoning, problem solving, communication, and connections through the mathematical concept or idea that is the focus.
>
> (INTASC, 1996, p. 25)

Completed portfolios are evaluated based on five interpretive categories—tasks, discourse, learning environment, analysis of learning, and analysis of teaching. Guiding questions for each category illustrate aspects of teachers' performance that could be considered in the collection and analysis of evidence.

The INTASC portfolio assessment is thus very different from prior teacher certification exams in that teacher knowledge of mathematics is assessed through the tasks, discourse, and analysis of student learning presented in the portfolio. Three of the evaluation framework's guiding questions are identified as specifically capturing facets of mathematical knowledge used in teaching: the appropriateness of the tasks for the instructional goals; the teacher's use of notation, language, and representation; and the accuracy of the teacher's interpretation and evaluation of information about students. Other areas of the portfolio might also demonstrate a teacher's use of mathematical knowledge in teaching, for example, in making decisions about when to probe students or in the ability to maintain the cognitive demand of a task during its implementation. Thus, the INTASC assessment has the potential for capturing the type of professionally situated knowledge identified by researchers—mathematical knowledge as it is deployed in the work of teaching. However, recognizing a teacher's skillful task selection in mathematics or imprecise use of mathematical language requires portfolio readers with significant mathematical knowledge. Thus, although the INTASC portfolio has the possibility of capturing mathematical knowledge as it is used, or misused, in teaching, its ability to do so would depend on the reader.

It is also worth noting that INTASC materials embrace the particular vision of mathematics instruction that is articulated in the INTASC standards and NCTM's *Professional Standards for the Teaching of Mathematics* (1991). Although the assessment handbook says that "there is no singular right way to teach mathematics" (INTASC, 1996, p. 36), teachers are required to demonstrate evidence of particular instructional formats and types of mathematics tasks. For example, one of the featured videotaped lessons must show students engaged in "mathematical problem solving and reasoning" where problem solving is defined as "students exploring new ideas and trying to pull prior

learning together to address non-routine problems, not simply applying an algorithm to a new context," and reasoning is defined as "students making and testing conjectures, constructing arguments, judging the validity of arguments, formulating counterexamples, etc." (INTASC, 1996, p. 27). In addition to problem solving and reasoning, the featured lessons must, between them, capture other aspects of instruction such as both whole group and non-whole group formats, student-to-student discourse, and the introduction or development of a mathematical concept. Whereas historical teacher tests kept knowledge of pedagogical techniques and knowledge of content separate, they are here intertwined; teachers must exhibit specific instructional skills. While this is part of a larger story about changes in "pedagogical" knowledge over the last two centuries, it warrants noting here because of its implications for the orientations and skills novice teachers must possess.

National Board for Professional Teaching Standards (NBPTS) Certification

In contrast to both the Praxis and INTASC assessments, the National Board for Professional Teaching Standards (NBPTS) certification is a voluntary endorsement for teachers who already possess the minimum state requirements for teacher certification and who have at least three years of experience. Certification is currently available in 24 areas (NBPTS, 2006b)—including Early Childhood/Generalist, Middle Childhood/Generalist, and Early Adolescence/Mathematics—and is valid for ten years, after which teachers can apply for renewal. There are specific standards for each certificate that describe "what an accomplished teacher should know and be able to do" (NBPTS, 2006a). In fact, the National Board Standards informed the development of INTASC standards, and, like INTASC, the NBPTS is endorsed both by governance (e.g., National Governors' Association; National School Boards Association) and teacher education associations (e.g., NCATE); NBPTS is also backed by the National Education Association.

NBPTS certification has two components: assessment center exercises and portfolio entries. These two types of performance-based assessments are designed to measure not only teachers' content knowledge, but also the skills and judgment they must deploy routinely in practice. Although the specifics and standards vary, the process is the same across different types of certificates.

Assessment center exercises focus primarily on content knowledge and are designed to measure what accomplished teachers should *know*, rather than questions that can be studied for. Teachers complete six open-ended response exercises at designated testing centers, located across the nation. Candidates are given up to 30 minutes to respond to each exercise. Organized around "challenging teaching issues," exercises require teachers to use their content knowledge in pedagogical situations, for example, to identify student misconceptions and plan an appropriate instructional strategy, or to describe a learning experience based on a foundational concept or in response to a student's inquiry. NBPTS contracted with ETS, the same organization that publishes Praxis, to develop its scoring systems.

All six assessment center exercises for the Early Adolescence/Mathematics certificate are focused on mathematics topics: algebra and functions, connections, data analysis, geometry, number and operation sense, and technology and manipulatives. For both the Early and Middle Childhood/Generalist certificates, however, only one of the six exercises explicitly involves mathematics. The Early Childhood/Generalist scoring guide describes such an exercise:

> Candidates will demonstrate their ability to identify mathematical misconceptions or difficulty in a student's work, to state the fundamental prerequisites needed by this student in order to learn this particular mathematical concept, and to plan an instructional strategy based on real-world applications. They will also be asked to choose the materials and to provide a rationale for their choice of these materials that will be used to teach these prerequisites. (NBPTS, 2005b, p. 9)

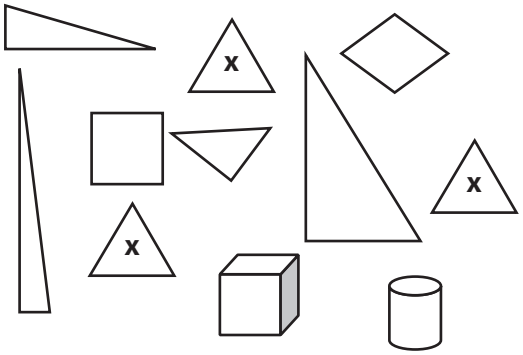
The scoring guide provides the example in Figure 4.1.

In this example, candidates must identify and explain a common student misunderstanding—failing to recognize non-equilateral triangles as triangles—then design instruction to remedy this problem. This exemplar can be characterized as "pedagogical content knowledge," for it asks teachers to solve mathematics-specific problems involving student misunderstandings and teaching strategies. However, because there is only one mathematics-specific task on the generalist certification assessments, it is unlikely that *mathematical* pedagogical content knowledge is the construct that is contained in the overall score teachers receive. Instead, teachers' scores on all of the assessment center exercises and portfolio entries, across content areas, are weighted and combined to make a final assessment about their level of expertise.

The NBPTS portfolio allows for in-depth examination of practice. It requires four entries, three of which document specific aspects of a teacher's practice

Stimulus

A kindergarten student is having difficulty with a mathematical concept:



The student was asked to mark an X on all of the triangles.

Prompts

1. Identify the mathematical misconception/difficulty of this student work sample.
2. What fundamental concepts are prerequisites for this student at this grade level in order to learn this skill?
3. Based on real-world applications, state your goal for an instructional strategy or learning experience to help this student. Plan a learning experience or instructional strategy based on this goal that would further the student's understanding of this mathematical concept.
4. What materials would you use to teach this mathematical concept to this child? Provide a rationale for your choice of materials.

Figure 4.1 Example in the Early Childhood/Generalist scoring guide. (Reprinted with permission from page 84 of the National Board for Professional Teaching Standards, Early Childhood/Generalist Scoring Guide, 2005. All rights reserved.)

using records such as videotapes, student work samples, and related instructional tasks and classroom assessments.[6] In addition to creating these records, teachers write reflective commentaries that describe, analyze, and evaluate their practice. Similar to the assessment center exercises, the Early and Middle Childhood/Generalist certificates require only one of the four portfolio entries to focus on mathematics. And in fact, this entry requires teachers to document their integration of mathematics and science, as described in the 308-page portfolio instructions document for the Middle Childhood/Generalist certificate:

> In this entry you will demonstrate how you help students better understand a "big idea" in science using relevant science and mathematics knowledge. You will engage students in the discovery, exploration, and implementation of these science and mathematics concepts, procedures, and processes by integrating these two disciplinary areas. This entry is designed for you to provide evidence of your ability to plan, describe, illustrate, assess, and reflect on your teaching practice.
>
> For this entry, you must submit the following:
>
> - Written Commentary (14 pages maximum) that contextualizes, analyzes, and evaluates the teaching and integration of the math and science instruction.
> - One Video Recording (15 minutes maximum) that demonstrates how you engage two small groups of students, with at least three students per group, in a science lesson that integrates mathematics.
> - Three Instructional Materials (no more than a total of 6 pages) related to the lesson featured on the video recording that will help assessors understand what occurred during the lesson. (NBPTS, 2006d, p. 169)

Each of the portfolio entries is evaluated independently according to a detailed rubric addressing selected NBPTS standards. For example, the portfolio entry described above is scored for standards such as knowledge of students, knowledge of content and curriculum, meaningful applications of knowledge, and reflection. For the two mathematics-specific certificates (Early Adolescence/Mathematics and Adolescence and Young Adulthood/Mathematics), the four portfolio entries each focus on a different aspect of mathematics teaching: developing and assessing mathematical thinking and reasoning, whole class mathematical discourse, small group mathematical collaborations, and contributions to student learning.

Our inspection of the NBPTS documents suggests several stances toward the mathematical knowledge needed for teaching, and toward the measurement of this construct. First, the NBPTS assessments do offer the opportunity to capture how teachers use their knowledge of mathematics in practice. For example, the mathematics-related assessment center exercise for the Middle Childhood/Generalist certificate describes a high scoring response as follows:

> The response offers clear, consistent, and convincing evidence of the ability to demonstrate pedagogical and content knowledge of math by accurately identifying the math misconception/error, providing an instructional strategy to assist student understanding of concepts/skills needed to accurately solve the mathematical problem, and providing a rationale that supports the instructional strategy. (NBPTS, 2005c, p. 53)

Like the INTASC standards, the NBPTS assessment presumes that teachers need intertwined knowledge

[6] The fourth focuses on accomplishments outside the classroom.

of content, students, and instructional design—and fluency with that knowledge, as evidenced by the short span of time allotted for teachers to answer the assessment center exercises. In this way, both the assessment center exercises and the portfolio entries reflect two aspects of teachers' mathematical work in classrooms: using integrated knowledge, and using it in the moment, rather than in a context more removed from actual classroom work. Thus, the NBPTS assessments better measure professionally situated mathematical knowledge than did past teacher certification exams, but, like the INTASC portfolios, are dependent on the samples of practice the teacher submits, as well as the evaluator's ability to recognize a teacher's use of mathematical knowledge in practice.

Second, there are a number of measurement-related concerns, some more pressing than others. While both INTASC and NBPTS have provided evidence for the validity of their measures, evidence on their reliability and other technical properties is more difficult to come by. Such evidence should be broadly available. Also, the small number of mathematics/science questions within K–8 exercises and prompts raises concerns about the degree to which the sampled tasks generalize to teachers' overall level of expertise in teaching. This is complicated by the fact that teachers can choose lessons for portfolio entries, and those lessons may not represent their everyday practice (Porter, Youngs & Odden, 2001). And this is even further complicated in the NBPTS generalist certificates by the combination of mathematics and science in the same portfolio task. Because NBPTS is interested in measuring accomplished general practice, they chose not to design separate measures of proficiency in teaching these two content areas. But accomplished practice might be discipline-specific. This leaves under-assessed the specific mathematical qualities of a teacher's instructional practice.

Third, both the NBPTS and INTASC portfolios seem as much designed to capture teachers' *reflection* on practice as the practice itself. For instance, the written commentary guidelines for the NBPTS include prompts such as: *What evidence of inquiry, intellectual engagement, discussion, and content is demonstrated in your video recording? How did you further students' knowledge and skills and engage them intellectually? How does the discussion and/or activity featured on the video recording reveal students' reasoning and understanding? Describe a specific example from this lesson as seen on the video recording that shows how you ensure fairness, equity, and access for students in your class* (NBPTS, 2006d, p. 173–174). This is a view of teaching that is heavily tailored toward ideas about teachers as "reflective practitioners" (Schön, 1995). It is worth considering, however, whether accomplished reflection on teaching is required for accomplished teaching; if it is not, then NBPTS might be missing a class of proficient educators. Moreover, the relationship of articulate reflection to effective instruction is not clearly established; yet the NBPTS makes this a central demand of its assessment strategy. Are teachers who can describe, explain, and reflect on their work better teachers than those who are less able to articulate their designs, purposes, or analyses? Possibly they are, for it may be that the ability to articulate one's practice is an indicator of deliberateness, and that the ability to write cogent reflections an indicator of analytic capacity, both of which may predict student achievement. But the nature of the evidence for this remains unexplained.

Finally, like INTASC, with its very specific view of mathematics instruction, and Praxis, with its instructional design tasks, the view of instruction presented in the NBPTS precludes some types of teachers from being recognized as accomplished—for instance, those who follow scripted curriculum materials, or those who do not teach mathematics using the inquiry/discussion/engagement methods suggested above. In the scoring guide, for instance, teachers whose portfolio is rated poorly are urged to consider whether it "provide(s) evidence that you have created a classroom environment that promotes active learning by all your students" and "provide(s) evidence that you were able to engage students in an effective classroom discussion or inquiry appropriate to the goals of your teaching" (NBPTS, 2005a, p. 15). Thus, there are examples of competent teaching that might not be recognized by NBPTS. Should a teacher who controls most classroom talk, but who also gives clear explanations of mathematical procedures or careful demonstrations of proofs be recognized as accomplished? A read of NBPTS documents suggests not: the organization bases its scoring system on what we have called the *quality of mathematics instruction*, or teachers' mathematical knowledge *and* pedagogical choices, considered together. However, other stances are also possible.

Other Certification Tests

Several other certification tests affect a large number of prospective teachers.

American Board for Certification of Teacher Excellence (ABCTE)

The certification program designed by the ABCTE is meant to attract career-changing adults into teaching, allowing these individuals to circumvent teacher

education programs and other state requirements for entry into the profession. As part of ABCTE certification, teachers must possess a bachelor's degree, pass "rigorous examinations of subject area and professional teaching knowledge" (ABCTE, 2006a), and submit materials for a federal background check. This process brings to mind certification routes from the eighteenth and nineteenth centuries, in which teacher candidates possessed some minimum level of schooling, passed an exam, and were assessed for moral fiber. As we shall see below, many of the debates evoked by the ABCTE program echo those from the late nineteenth century.

In mathematics, as in most other ABCTE subjects, the exam is entirely multiple-choice and administered via computer. Test content is based on ABCTE standards, which were developed in turn by reviewing state and professional standards. Standards for elementary teachers suggest that test developers expect teachers to know not only the content in the K–6 curriculum, but also content their students will encounter in middle or high school. For instance, in the area of algebra, teachers certified via the elementary exam must know the symbolic language of mathematics, "including applying proportional reasoning, defining the inverse function and performing arithmetic operations on functions, solving systems of linear and quadratic equations and inequalities, and extending and recognizing linear patterns" (ABCTE, 2006b). Many of these topics are typically taught in middle or high school. The middle/high school standards extend into content often taught in post-secondary education, such as least squares regression, matrix algebra, and calculus topics such as limits, derivatives, integrals, and convergence of sums. Secondary test items reflect this emphasis:

1. Twenty gallons of solution A, which is 35% acid, is mixed with 30 gallons of solution B. The resulting mixture contains 20% acid.

   What percent of solution B is acid?

   a) 5% b) 10% c) 12.5% d) 15%

2. The following list of ordered pairs represents a functional relationship:

   (1,1) (2,4) (3,9) (4,16)

   What principle of functional relationships supports this statement?

   a) The x values are all different.
   b) The relationship between each pair of numbers is reciprocal.
   c) The pairs are listed in consecutive order.
   d) The relationship between each pair of numbers is the same.

3. What is $\lim_{x \to 2} \frac{x^3 - 8}{x - 2}$?

   a) 0 b) 4 c) 8 d) 12

The first item is perhaps a modern-day equivalent of the types of problems that appeared on the 1890s Michigan teacher exams; the second taps more conceptual knowledge, investigating teachers' understanding of the definition of functions; the third item taps a topic typically not taught until college. Overall, this is a mathematically demanding assessment, with items that are computationally intensive, designed to push on candidates' misconceptions, and to push into territory beyond that typically taught in high school.

To date, the ABCTE certification process has not been widely used. Although the program is funded by two U.S. Department of Education grants totaling $40 million, it certified only 135 individuals between its founding in 2001 and April of 2006 (Schimmel, personal communication, April 19, 2006). One reason for the slow start is that the ABCTE certification must be recognized by states before teachers can gain employment; in mid-2006, only five states recognized the ABCTE certification. Another reason might be the proliferation of state and district-based alternative certification programs, which reduce incentives for teachers to seek the ABCTE stamp of approval, and also for states to recognize the organization's certification process.

The tepid response by states and teachers has not dulled the uproar over the ABCTE certification process. For the most part, critiques center around the assumption that teachers can be certified via a multiple-choice exam, absent professional training and field experiences. As Berliner (2005) writes of the ABCTE and similar tests, "it is a near impossibility to adequately assess quality teaching through pencil-and-paper tests of professional knowledge. . . . These tests fail in part because of the complexity of classroom environments and the near impossibility of capturing that reality in pencil-and-paper formats" (p. 211). Educators also assert that in other professions, such as law, medicine and architecture, the certification process includes not only an exam but also extended professional learning and apprenticeship.

Thus the ABCTE has reopened the debate over who is qualified to teach, and how best to assess individuals' skills and capacity for this work. Teacher educators, many of whom oppose any form of alternative certification, have singled out this test-based certification method for particular opprobrium, much as educators at the turn of the twentieth century railed against the teacher tests administered by local school boards and officials. Many issues are the same. Teacher educators claim there is a professional body of knowledge learned primarily through preservice preparation and apprenticeship, and that tests provide a "back door" for underqualified individuals to enter teaching. ABCTE support-

ers argue that teachers' subject matter knowledge is of the foremost importance, and that there needs to be a method to identify knowledgeable individuals to help alleviate teacher shortages, shortages caused in part by the bureaucratic restrictions most states have placed on the entry to teaching. ABCTE supporters also take the view that teachers need to know more advanced content than they are expected to teach students, much as the Michigan superintendent (and test promoter) Henry R. Pattengill did over 100 years earlier. Teacher educators, for their part, tend to focus more on promoting deep knowledge of the specific curriculum teachers will teach students.

California Subject Examinations for Teachers

The California Subject Examinations for Teachers (CSET) are reviewed here briefly as an example of both a state-specific test, and as an example of a customized exam developed by National Evaluation Systems (NES, 2006) specifically for one state's use. NES is the second-largest supplier of teacher certification assessments, providing tests to not only California but also Texas and a number of smaller states. In California, NES customized tests to align with California curriculum frameworks and student standards.

Two sample items from the elementary-level (multiple subjects) exam provide some insight into the composition of the test:

1. If the number 360 is written as a product of its prime factors in the form a^3b^2c, what is the numerical value of $a + b + c$?

   a) 10 b) 16 c) 17 d) 22

2. The problem below shows steps in finding the product of two two-digit numbers using this standard multiplication algorithm. The missing digits in the problem are represented by the symbol □.

$$\begin{array}{r} \square 9 \\ \times\ 3\square \\ \hline 2 9 \square \\ +\ \square 4 \square\square \\ \hline \square\square\square\square \end{array}$$

   What is the hundreds digit in the product of the two numbers?

   a) 1 b) 4 c) 6 d) 7

These and other items assess upper-elementary content (e.g., slope; proportionality). They also appear designed to draw on both candidates' basic skills and on their ability to reason through novel mathematics problems. On this test, candidates' mathematics and science ability is reported as a separate subtest score, allowing test users to gauge individual competency in both areas combined. Praxis II, in contrast, only reports overall scores to employers.

Candidates for secondary mathematics positions must take two mathematics examinations, one focusing on number theory and algebra and the other focusing on geometry and probability and statistics. To teach advanced placement classes or classes in analysis and calculus, a candidate must take a third mathematics examination covering the history of mathematics and calculus. The examinations were designed by NES to follow the mathematics content standards for California Public Schools, and these are expressed as lists of topics to know for teaching. This parallels the development of Praxis, where items were developed to reflect mathematical knowledge represented by lists of topics. The example below shows the kind of background CSET provides to explain the undergirding of its mathematics items:

1.4 Linear Algebra

a. Understand and apply the geometric interpretation and basic operations of vectors in two and three dimensions, including their scalar multiples and scalar (dot) and cross products.

b. Prove the basic properties of vectors (e.g., perpendicular vectors have zero dot product)

c. Understand and apply the basic properties and operations of matrices and determinants (e.g., to determine the solvability of linear systems of equations)

(*Mathematics Content Standards for California Public Schools*, Algebra I: 9.0; Algebra II: 2.0; Mathematical Analysis: 1.0; Linear Algebra: 1.0–12.0)

One can see the direct connection between this policy statement and the item below:

If vectors $\vec{a} = (a_1, a_2)$ and $\vec{b} = (b_1, b_2)$ are perpendicular,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

and if we identify vectors with column matrices in the usual manner, then show that the vectors $A\vec{a}$ and $A\vec{b}$ are perpendicular for all values of θ.

Like the elementary exam, the secondary exam appears to emphasize reasoning and proof, at least in its constructed response items. The exam also appears quite challenging; the following items draw from post-secondary content:

1. Which of the following statements refutes the claim that $GL_R(3)$, the set of 3×3 invertible matrices over the real numbers, is a field?

   a) There exist elements A and B of $GL_R(3)$ such that $AB \neq BA$.

   b) There exist elements A and B of $GL_R(3)$ such that $\det(AB) = \det(A)\det(B)$.

   c) If A is an element of $GL_R(3)$, then there exists a matrix A^{-1} such that $A^{-1}A = I$.

   d) If A is an element of $GL_R(3)$, then there exists a matrix A such that $\det(A) \neq 1$.

2. Show that the subset of complex numbers of the form $a + bi$ with a and b rational numbers satisfies the axioms of a field under the operations of addition and multiplication of complex numbers.

In all, however, many of the items on the CSET, ABCTE, and secondary Praxis II look quite similar. This overlap is one characteristic of current secondary teacher certification tests: aside from different content emphases and more or less difficult items, there is less variation than one would think, given the proliferation of current tests.

The Future of Certifying Teachers of Mathematics

At no time in this nation's history have there been so many routes to becoming a certified teacher; teachers can do everything from taking multiple-choice content exams to constructing elaborate records of their teaching practice. Where does this leave those interested in moving teacher certification toward an assessment of more job-relevant, discipline-specific professional skills? We begin by discussing the benefits and drawbacks of current certification exams, then in our conclusion consider some of the issues that might lead to the next generation of certification tests.

Clearly, most new forms of teacher licensure testing have significant benefits as compared to past forms and, in some cases, even as compared to scholarly methods for assessing mathematical knowledge for teaching. To start, unlike their historical counterparts, most of the current tests attend carefully to issues of reliability and, increasingly, validity. For instance, a number of studies comparing NBPTS teacher candidates who have received certification and those who have not have revealed measurable differences between NBPTS-certified and regularly certified teachers and their impact on student learning and growth (Bond, Smith, Baker, & Hattie, 2000; Cavalluzzo, 2004; Goldhaber, Perry & Anthony, 2004; Vandevoort, Amrein-Beardsley, & Berliner, 2004; for a counterexample, see Sanders, Ashton, & Wright, 2005). This is a key question centering on predictive validity: Does the assessment provide information about the candidate's actual effectiveness in the classroom? If not, the assessment is of little value.

Second, most current certification tests explicitly intend to assess professionally specific knowledge of mathematics, such as pedagogical content knowledge. And, in the case of performance-based assessments—such as observations, portfolios, the Praxis content exercises, or the NBPTS assessment exercises—not only does the content of the assessments more accurately reflect ways that teachers use mathematics in practice, the assessments themselves emerge from or are closely tied to the actual work teachers do. Given that sociologists typically define professions by the presence of a codified body of knowledge with professionally controlled barriers to entry for new practitioners, this is a significant advance. In fact, efforts such as INTASC and NBPTS actually contribute to this development, for their standards attempt to articulate a coherent and comprehensive vision of the knowledge, skills, and dispositions needed for teaching.

Another benefit of performance-based assessments is they may actually stimulate teacher reflection and learning, thus providing an opportunity for professional development. For instance, individuals who seek ABCTE certification take a self-assessment and, if necessary, work with a learning advisor to refresh and/or learn content areas prior to taking the actual exam. Teachers have reported that the NBPTS certification process inspires growth and self-evaluation (Rotberg, Futrell, & Lieberman, 1998) and encourages teachers to become more reflective about their practice (Haynes, 1995; Letofsky, 1999; Sumner, 1997; Tracz et al., 1995).

However, new forms of teacher certification also have drawbacks. To start, the closer to practice the assessment, the less likely it is to be used. The secondary content knowledge Praxis II, for instance, was taken by six times more examinees than the secondary mathematics pedagogy exam (ETS, 2006b, p. 11). Only 2 of the 43 states that use Praxis I and/or II for certification require Praxis III. And only small numbers of teachers seek NBPTS certification; one recent survey found that only 2% of middle grades teachers and 3% of high school teachers had applied for certification in the past three years (Horizon Research, 2002).

A second issue centers on equity: Research routinely suggests certification testing has adverse impacts on teachers from minority groups, who for instance have historically scored lower on both the NTE and Praxis.[7] Minority candidates may successfully complete teacher education coursework and internships but not meet their states' requirements for a passing score on Praxis II (Albers, 2002). This raises questions about the construction of the Praxis tests themselves. Some suggest the content and the format privilege the knowledge base of certain groups: for example, that Praxis tests in English favor white literature over minority literatures, and therefore disadvantage minority teacher education students. It has also been suggested that the curricula and style of learning predominant

[7] See Haney et al. (1987) for a discussion of the NTE. For Praxis, Wakefield (2003) examined statistics released from the Georgia Professional Standards Commission. These statistics revealed that in 2001, European Americans had a 93% pass rate on Praxis II, while African Americans had a 71% pass rate (p. 383).

in historically black universities and colleges is not reflected in standardized tests in general and in Praxis in particular (Albers, 2002).

There also have been equity concerns about NBPTS certification. Some believe that the process creates an unnecessary hierarchy within the field of teaching. Even though a growing array of grants, scholarships, and no-cost loans are now available to cover the cost of participation in the certification process, some argue that the $2,500 fee may still limit who can participate in the process (King, 1994; Marshall, 1996). Others critique the lack of inclusion of "culturally sensitive pedagogies" in the framework (Hamsa, 1998; Irvine & Fraser, 1998). Furthermore, while the number of minority applicants is proportionate to their representation in the teaching work force (17%), their achievement of certification is lower (only 11% of NBPTS-certified teachers are minority). NBPTS and other researchers are critically examining the role of cultural bias in these results as well as support mechanisms that might address this situation (Kraft, 2001; SRI, 2004; Serafini, 2002).

A third concern involves reliability. How accurately can the assessments determine how much mathematical knowledge for teaching an individual has? Standardized assessments, such as Praxis, return indices of reliability that can be used to assess the accuracy of the test. For more novel formats, such as portfolios and constructed response tasks, the determination of the degree of accuracy is more complex—dependent on the selection of specific tasks, scoring rubrics, and also on the ability of the person scoring the assessment. Reliability between scorers must be reached, but there must also be some assurance, in our view, that if one took another sample of a teacher's mathematical practice, the same overall judgment regarding certification would be reached. With any stakes attached to the assessment, this is a critical issue.

Yet another issue is the lack of evidence on the validity of many of these assessments, or the degree to which these certification tests predict what teachers can do in classrooms to help students learn. Aside from the NBPTS and some of the scholarly measures, few assessments have been validated by anything more than a review of the literature and signatures from the practicing teachers and mathematics educators who serve on standards development and assessment review panels. The need for more evidence on predictive validity in particular—that is, whether teachers who pass specific certification tests produce better-educated students than those who do not—is dire.

Fourth, it is interesting to note how the mathematical topics to be tested have changed over time. Today there is an increasing emphasis on algebra in some exams: fully 40% of the CSET items for mathematics certification in California are algebra problems, and the Praxis II examination for secondary mathematics teachers has 16–18% algebra items, up from 12–14% on the NTE. Interest in geometry has grown as well: in the days of the NTE, geometry items represented 16–18% of the entire pool of mathematics items, and the Praxis II exams, which replaced the NTE in most states, now have 22–26% geometry out of all math items. What counts as mathematically important knowledge for teaching seems to vary over time and by test.

So where does this leave us? Clearly, there is currently no ideal certification assessment, at least in mathematics. However, we do stand in a better position than just twenty years ago. Basic research into mathematical knowledge for teaching has supplied a vision for what might be measured on these assessments, one that contrasts with the "teachers should know a little more than their students" view taken during most of U.S. history. And advances in assessment theory—both psychometrics and test theory in general—have encouraged the adoption of better formats and improved test characteristics. Yet much remains to be done. Despite the alternative vision of what should be measured, we do not know whether this is the mathematics that actually matters to student learning. Pattengill's old question—whether teachers need to know content, or teaching methods, or both—survives today, although in slightly different form. And the items on this list are complicated by the fluid nature of professional knowledge, and by debates over mathematics teaching more broadly. Some of what counted as profession-specific mathematical knowledge on the 1900 certification exams does not count as knowledge today; what counts as knowledge today might not count as knowledge tomorrow.

CONCLUSION

Teacher examinations have a complicated history and pose a number of serious contemporary dilemmas. Even 150 years after the first written teacher tests emerged in the U.S., there is little agreement on what, whom, and how to measure, and for what purpose. Pressing questions—such as the balance of knowledge of content *and* knowledge of pedagogy, the nature of content knowledge useful for teaching, and the "content" of pedagogical knowledge — have not been answered. Meanwhile, the number of teacher exams continues to increase. And pressures for more accountability, both among the programs

that prepare teachers and among teachers themselves, continue to mount.

Given these pressures, it is not likely that teacher testing will subside. The challenge, perhaps, is to create the best tests possible. Measuring teacher knowledge, even using standardized modes of assessment, *can* be done in ways that honor and define the work of teaching, ratify teachers' expertise, and help to ensure that every child has a qualified teacher. Doing so requires carefully constructed instruments that take seriously the work of teaching and that can be used at scale. In this concluding section, we turn to consider possible ways to contribute to the improvement of teacher assessment, both standardized and other forms. Our remarks take the form of proposed directions and foci, based on our review of practice in this domain.

Measure Mathematical Knowledge for Teaching

It is no surprise that we argue for measuring the mathematical knowledge used in teaching. Valid teacher assessments should not be remote from what teachers are asked to do in classrooms, with real students, materials, and content. The extended example of division of fractions in the middle of this chapter illustrates some of our own bets about what kind of knowledge matters; others have different bets, and there will no doubt be some evolution toward the "correct" mix of problem types, tasks, and items that should be contained on teacher assessments. With effort, there can also be an evolution toward appropriate formats for the assessment, through in-person observations, pencil-and-paper assessments, portfolios, or other formats. These directions can bridge more effectively the scholarly work on teacher knowledge and the policy-related work on assessing teacher quality.

However, we are worried about two trends. One is this field's tendency to conflate teachers' knowledge of mathematics for teaching with other types of knowledge or skill. For instance, many of the assessments we studied conflated knowledge of instructional approaches (e.g., using manipulatives; involving students in meaning-making; teaching in line with the NCTM standards) with knowledge of mathematics. This is reminiscent of our distinction between the *quality of mathematics instruction*, which we defined as inclusive of instructional approaches, and the *quality of the mathematics in instruction*, which we defined as focused specifically on the actual mathematics deployed in the course of a lesson. Other assessments of mathematical knowledge for teaching required teachers to design lessons, a possibly separate skill. And still other assessments appear to be scored based on a teacher's ability to reflect about her instruction. Such assessments are not necessarily invalid; however, these other elements of teacher knowledge should be recognized as a crucial component of what is being measured, and test scores should be interpreted appropriately.

Mathematical knowledge for teaching should also not be represented as only one element in a composite score. Mathematics teaching is important in its own right, and students deserve teachers who have qualified for their jobs based on knowing mathematics in the ways it should be known for teaching—not demonstrating preparedness in many areas, including mathematics. This is especially true, in our analysis, for elementary-school certification exams, where mathematics is likely to be combined with other areas (e.g., science, in the elementary CSET and NBPTS) or with all subjects (e.g., elementary content area assessment of the Praxis II).

Measure with Care

No test is perfect; as we have shown, every format, whether multiple choice, observational, or performance based, has advantages and disadvantages. Recognizing this, the task facing educators and policy-makers is to carefully and responsibly link the research/policy question being investigated to the most appropriate form of assessment.

Two other concerns are worthy of mention. Assessments can send unintended messages about their purpose. The format of the assessment may serve, inadvertently, as a model of assessment that teachers believe should then be used in assessing their students. Multiple-choice exams are vulnerable to many criticisms, yet their use for the testing of teachers may suggest to teachers that this is a preferred assessment method. Even when the content of the multiple choice questions is focused on robust expressions of what teachers know to teach, the format may speak louder and have greater implications than intended by the developers. Similarly, mathematics educators have worked hard to counter the misconception that mathematical competence is demonstrated by quick solutions to routine mathematical problems. Certain testing formats, multiple-choice in particular, may solidify this misconception among teachers.

This field also needs to recognize how the technical requirements for standardized testing affect test content. In his scathing critique of standardized testing in history, Wineburg (2004) notes that the vagaries of modern psychometrics conspire to hide what students know about history rather than measure what they *do* know. Wineburg describes how analytic techniques

such as biserial correlation disadvantage those whose knowledge base is different than the mainstream, and that only items that produce a good "discrimination index" (that is, that differentiate between high achievers and low achievers) are included in multiple-choice exams, regardless of the intrinsic value of their content. He writes, "If ETS statisticians determined during pilot testing that most students could identify George Washington, 'The Star-Spangled Banner,' Rosa Parks, the dropping of the bomb on Hiroshima, slavery as a main cause of the Civil War, the purpose of Auschwitz, Babe Ruth, Harriet Tubman, the civil rights movement, the 'I Have a Dream' speech, all those items would be eliminated from the test, for such questions fail to discriminate among students" (2004, p. 1409). Measures of mathematical knowledge for teaching are vulnerable to the same dilemma.

Use Multiple Approaches

One way to effect well-considered improvements in this enterprise is to avoid repeating the mistakes of the past. One of the problems with assessing professional knowledge has been reliance on traditional test formats. In this chapter, we have examined both standard assessments formats and more novel forms of assessment, including discourse analysis, video analysis, and portfolio assessment. The most widely used teacher test today is the Praxis series, which offers multiple-choice and open-ended items in the paper-and-pencil tests that measure inert knowledge for teaching, as well as real-time observation of teaching that is designed to assess knowledge in action. Only two states are currently using the observation-based Praxis III for certification, but we think this combination of testing formats is promising. New uses of multimedia for documenting teaching offer other avenues for comprehensive appraisals of teaching work, as discussed earlier in the chapter.

Meet Professional Standards of Rigor in Assessment

Another important step is to confront the challenge of rigor, and what "rigor" might appropriately mean. New assessments should be held to more rigorous standards than old assessments, which were amazingly often validated only with reference to the content they covered, despite their use in making inferences about practice and its quality. New assessments should be rooted in theory, and scores should be validated using a number of methods. In our own work, paper-and-pencil measures of the mathematical knowledge for teaching are validated by close analyses of the mathematical work of teachers that we videotaped. Is the mathematical knowledge that teachers demonstrate on a paper-and-pencil test expressed in their teaching? Can we show that students of highly qualified teachers, as defined by this paper-and-pencil test, learn more mathematics? If all mathematics assessments for teachers answered such questions, tests could bring real value to our efforts to improve educational quality. Rigorous standards for assessment already exist in the form of American Educational Research Association/American Psychological Association guidelines (1999) and the field can readily use them.

Learn from Other Measurement Methods

In the introduction to this chapter, we noted that seldom do lessons from one line of work measuring teachers' mathematical knowledge "travel" to other lines of work. Qualitative researchers have much to learn from large-scale test developers; large-scale test developers need lessons learned in qualitative research to succeed. More crossover needs to occur between these separate enterprises.

Attend to Issues of Equity

Attention to equity is crucial. As we saw, exams have a long and disturbing history of excluding people of color from the teaching profession. That these assessments regularly discriminated among candidates on the basis of characteristics unrelated to the knowledge and skill needed for teaching was a serious problem, and a threat to the validity of the exams themselves. Doing well on the tests must be related to success in teaching, not demographic characteristics. When patterns of performance are predicted by race, examinations must be scrutinized to uncover what it is about the test questions that may be producing this result. Tests that are invalid across groups in such ways must be challenged publicly; moreover, their use also serves to inhibit the development of a diverse teaching force.

Investigate the Relationship among Mathematical Knowledge for Teaching, Other Domains of Teaching Knowledge, and Student Learning

Much work remains to be done to specify what we are calling "mathematical knowledge for teaching." Important to understand is how this domain intersects with other domains of knowledge for teaching, and how these together affect student

learning. We know, for example, that mathematical knowledge for teaching intersects knowledge for equitable teaching. Being explicit about mathematical practices makes mathematics learning accessible for wider populations. Mathematical knowledge for teaching may also act upon (or be acted upon by) teachers' ability to motivate students to learn, to organize classrooms for productive instruction, and to produce learning gains.

Increase Professional Role and Control

The history of U.S. teacher testing in mathematics has been one in which the tests have largely been developed by those outside the profession: local officials, bureaucrats, and later, professional test developers. In the coming years, professionals in education and in mathematics must develop the earned authority to control the identification of professional knowledge and the means of certifying it reliably. The judgments involved require specialized insight and perspective, rooted in evidence that is tied to instructional effectiveness—that is, instructional practices that reliably produce student achievement. Improvement in assessing teachers depends on close involvement of experts in mathematics instruction.

Where does the tour on which we have been lead? We have seen that teacher assessment was originally narrow, closely bound to the content of the curriculum. Teacher qualification was judged by teachers' ability to solve difficult word problems of the sort they might teach their students. Over the last 100 years, approaches to assessment have multiplied. What is assessed has broadened dramatically. How it is assessed has similarly expanded. Even within the same testing firm, different assessments are developed. The field has moved from a narrow and underspecified conception of the knowledge necessary for teaching to a broadband approach which lacks the precision and common agreement about requirements that one would expect of a profession. The agenda for assessing teacher knowledge is clear: The need to prepare and certify knowledgeable and skilled teachers who can be reliably effective with a wide range of students has perhaps never been greater. With the expansion of interest in and approaches to assessment, the field needs to move toward clearer precision, specification and agreement of measures, and broader skilled use of reliable and valid tools for appraising teachers' knowledge, skill, and performance. Research on teaching and teaching effectiveness is ready for this step; the current state of policy and practice demands it.

REFERENCES

Abbott, A. D. (1988). *The system of the professions: An essay on the division of expert labor*. Chicago: University of Chicago Press.

Albers, P. (2002). Praxis II and African American teacher candidates (or, is everything black bad?). *English Education 34*(2), 105–125.

American Board for Certification of Teacher Excellence. (2006a). *About passport to teaching*. Retrieved June 21, 2006, from http://www.abcte.org/passport

American Board for Certification of Teacher Excellence. (2006b). *Multiple subject exam*. Retrieved June 21, 2006, from http://www.abcte.org/standards/mse

American Education Research Association/American Psychological Association. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Angus, D. L. (2001). *Professionalism and the public good: A brief history of teacher certification*. Washington, DC: Thomas Fordham Foundation.

Appleman, D., & Thompson, M. J. (2002). "Fighting the toxic status quo": Alfie Kohn on standardized tests and teacher education. *English Education 34*(2), 95–103.

Baker, R. S. (2001). The paradoxes of desegregation: Race, class, and education, 1935–1975. *American Journal of Education, 109*(3), 320–343.

Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *Elementary School Journal, 90*, 449–466.

Ball, D. L., & Bass, H. (2003). Making mathematics reasonable in school. In G. Martin (Ed.), *Research compendium for the Principles and Standards for School Mathematics* (pp. 27–44). Reston, VA: National Council of Teachers of Mathematics.

Ball, D. L., Lubienski, S., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). Washington, DC: American Educational Research Association.

Ball, D. L., Masters-Goffney, I., & Bass, H. (2005). The role of mathematics instruction in building a socially just and diverse democracy. *The Mathematics Educator, 15*(1), 2–6.

Ball, D. L., Thames, M. H., & Phelps, G. (2005, April). *Articulating domains of mathematical knowledge for teaching*. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Quebec.

Begle, E. G. (1972). *Teacher knowledge and student achievement in algebra* (SMSG Rep. No. 9). Palo Alto, CA: Stanford University.

Begle, E. G. (1979). *Critical variables in mathematics education: Findings from a survey of the empirical literature*. Washington, DC: Mathematical Association of America and National Council of Teachers of Mathematics.

Berliner, D. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*(3), 205–213.

Boardman, A. E., Davis, O. A., & Sanday, P. R. (1977). A simultaneous equations model of the educational process. *Journal of Public Economics, 7*, 23–49.

Borko, H., Eisenhart, M., Brown, C. A., Underhill, R. G., Jones, D., & Agard, P. C. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education, 23*, 194–222.

Bond, L., Smith, T., Baker, W., & Hattie, J. (2000). *The certification system of the National Board for Professional Teaching Standards:*

A construct and consequential validity study. Greensboro: University of North Carolina, Greensboro, Center for Educational Research and Evaluation.

Bush, W. S., Ronau, R., Brown, T. E., & Myers, M. H. (2006, April). *Reliability and validity of diagnostic mathematics assessments for middle school teachers*. Paper presented at the American Educational Association Annual Meeting, San Francisco.

Cavalluzzo, L. (2004, November). *Is National Board certification an effective signal of teacher quality?* Alexandria, VA: The CNA Corporation.

Coleman, J. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Darling-Hammond, L., & Sykes, G. (2003, September 17). Wanted: A national teacher supply policy for education: The right way to meet the "highly qualified teacher" challenge. *Education Policy Analysis Archives, 11*(33). Retrieved June 21, 2006, from http://epaa.asu.edu/epaa/v11n33/.

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teacher": What does "scientifically-based research" actually tell us? *Educational Researcher, 31*(9), 13–25.

Dwyer, C. A. (1998). Psychometrics of Praxis III: Classroom performance assessments. *Journal of Personnel Evaluation in Education, 12*(2), 163–187.

Education Trust. (1999). Not good enough: A content analysis of teacher licensing exams. *Thinking K–16, 3*(1).

Educational Testing Service. (1984). *A guide to the NTE core battery tests: Communication skills, general knowledge, professional knowledge*. Princeton, NJ: Author.

Educational Testing Service (1987). *NTE: A guide to the Education in the Elementary School Specialty Area Test*. Princeton, NJ: Author.

Educational Testing Service. (2003). *Praxis study guide for the mathematics tests*. Princeton, NJ: Author.

Educational Testing Service. (2005a). *Elementary education: Curriculum, instruction, and assessment (0011)*. Princeton, NJ: Author.

Educational Testing Service. (2005b). *Mathematics: Content knowledge (0061)*. Princeton, NJ: Author.

Educational Testing Service. (2005c). *Mathematics: Pedagogy test at a glance*. Princeton, NJ: Author.

Educational Testing Service (2006a). *State requirements*. Princeton, NJ: Author. Retrieved June 27, 2006, from http://www.ets.org/portal/site/ets/menuitem.22f30af61d34e9c39a77b13bc3921509/vgnextoid=d378197a484f4010VgnVCM10000022f95190RCRD.

Educational Testing Service (2006b). *Understanding your Praxis scores, 2005–2006*. Princeton, NJ: Author. Retrieved June 15, 2006, from http://www.ets.org/Media/Tests/PRAXIS/pdf/09706PRAXIS.pdf

Educational Testing Service. (2006c). *Elementary education: Content knowledge (0014)*. Princeton, NJ: Author.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*, 458–498.

Ferrini-Mundy, J., Floden, R., McCrory, R., Burrill, G., & Sandow, D. (2005). *A conceptual framework for knowledge for teaching school algebra*. East Lansing, MI: Authors.

Fischbein, E., Deri, M., Nello, M. S., & Marino, M. S. (1985). The role of implicit models in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education, 16*(1), 3–17.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K .S. (2001). What makes professional development effective? Lessons from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945.

Geertz, C. (1973). *The interpretation of cultures: Selected essays*. New York: Basic Books.

Goldhaber, D., Perry, D., & Anthony, E. (2004). *National Board certification: Who applies and what factors are associated with success?* (Working Paper). Washington, DC: The Urban Institute, Education Policy Center.

Graeber, A. O., Tirosh, D., & Glover, R. (1989). Preservice teachers' misconceptions in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education, 20*(1), 95–102.

Greenwald, R., Hedges, L.V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 6*, 361–396.

Hamsa, I. S. (1998). The role of the National Board for Professional Teaching Standards. *Education, 118*(3), 452–458.

Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education. *Review of Research in Education, 14*, 169–238.

Hanushek, E. A. (1972). *Education and race: An analysis of the educational production process*. Lexington, MA: D. C. Heath.

Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management, 1*, 19–41.

Hanushek, E. A. (1996). A more complete picture of school resource policies. *Review of Educational Research, 66*, 397–409.

Harbison, R. W., & Hanushek, E. A. (1992). *Educational performance for the poor: Lessons from rural northeast Brazil*. Oxford, England: Oxford University Press.

Haynes, D. (1995). One teacher's experience with National Board assessment. *Educational Leadership, 52,* 58–60.

Hess, F. (2002). Break the link. *Education Next, 2*(1), 22–28.

Hill, H.C., Rowan, B., & Ball, D.L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement**.** *American Educational Research Journal, 42*, 371–406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*, 11–30.

Horizon Research. (2000). *Inside the classroom observation and analytic protocol*. Chapel Hill, NC: Horizon Research.

Horizon Research. (2002). *The 2000 national survey of science and mathematics education: Compendium of tables*. Chapel Hill, NC: Horizon Research.

Ingersoll, R. (1999). The problem of underqualified teachers in American secondary schools. *Educational Researcher, 28*(2), 26–37.

Interstate New Teacher Assessment and Support Consortium. (1995). *Model standards in mathematics for beginning teacher licensing and development: A resource for state dialog*. Washington, DC: Author.

Interstate New Teacher Assessment and Support Consortium. (1996). *INTASC mathematics teacher performance assessment handbook*. Washington, DC: Author.

Interstate New Teacher Assessment and Support Consortium. (2006). *INTASC portfolio development*. Retrieved June 21, 2006, from http://www.ccsso.org/projects/Interstate_New_Teacher_Assessment_and_Support_Consortium/Projects/Portfolio_Development/

Irvine, J. J., & Fraser, J. W. (1998). Warm demanders: Culturally responsive pedagogy of African American teachers. *Education Week, 17*(35), 42.

Kennedy, M. M., Ball, D. L., & McDiarmid, G. W. (1993). *A study package for examining and tracking changes in teachers' knowledge* (Technical Series 93–1). East Lansing, MI: The National Center for Research on Teacher Education.

Kersting, N. (2004, April). *Assessing what teachers learn from professional development programs centered around classroom videos and the analysis of teaching: The importance of reliable and valid measures to understand program effectiveness.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

King, M. B. (1994). Locking ourselves in: National standards for the teaching profession. *Teaching and Teacher Education, 10*(1), 95–108.

Knowing Mathematics for Teaching Algebra Project. (2006). *Survey of knowledge for teaching algebra.* East Lansing: Michigan State University. Retrieved June 21, 2006, from http://www.msu.edu/~kat/

Kraft, N. P. (2001). *Standards in teacher education: A critical analysis of NCATE, INTASC, and NBPTS (A conceptual paper/review of the research).* (ERIC Document No. ED462378), 1–29.

Learning Mathematics for Teaching Project. (2006a). *A coding rubric for measuring the quality of mathematics in instruction.* Ann Arbor, MI: Author.

Learning Mathematics for Teaching Project. (2006b). Measures of teachers' knowledge for teaching mathematics. Ann Arbor, MI: Author. Retreived June 15, 2006, from www.sitemaker.umich.edu/lmt.

Leinhardt, G., & Smith, D. A. (1985). Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology, 77*, 247–271.

Letofsky, J. (1999). National Board certification. *Center-space, 13*(2), 1–5.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.* Mahwah, NJ: Erlbaum.

Marshall, M. (1996). Familiar stories: Public discourse, National Board standards and professionalizing teaching. *English Education, 28*, 39–67.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement,* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Michalowicz, K. D., & Howard, A. C. (2003). Pedagogy in text: An analysis of mathematics texts from the nineteenth century. In G. M. A. Stanic & J. Kilpatrick (Eds.), *A history of school mathematics* (Vol. 1, pp. 77–109). Reston, VA: National Council of Teachers of Mathematics.

Michigan Department of Public Instruction. (1894). *Fifty-seventh annual report of the superintendent of public instruction, Michigan.* Lansing: State Printers. Bentley Historical Library, University of Michigan.

Michigan Department of Public Instruction. (1896). *Fifty-ninth annual report of the superintendent of public instruction, Michigan.* Lansing: State Printers. Bentley Historical Library, University of Michigan.

Michigan Department of Public Instruction. (1897). *Sixtieth annual report of the superintendent of public instruction, Michigan.* Lansing: State Printers. Bentley Historical Library, University of Michigan.

Michigan Department of Public Instruction. (1898). *Sixty-first annual report of the superintendent of public instruction, Michigan.* Lansing: State Printers. Bentley Historical Library, University of Michigan.

Michigan Department of Public Instruction. (1901). *Sixty-fourth annual report of the superintendent of public instruction, Michigan.* Lansing: State Printers. Bentley Historical Library, University of Michigan.

Mullens, J. E., Murnane, R. J., & Willett, J. B. (1996). The contribution of training and subject matter knowledge to teaching effectiveness: A multilevel analysis of longitudinal evidence from Belize. *Comparative Education Review, 40*, 139–157.

National Board for Professional Teaching Standards. (2005a). *Handbook on National Board Certification.* Retrieved June 29, 2006, from http://www.nbpts.org/UserFiles/File/scoringhandbook.pdf

National Board for Professional Teaching Standards. (2005b). *NBPTS early childhood generalist scoring guide.* Retrieved July 20, 2006, from http://www.nbpts.org/for_candidates/certificate_areas?ID=17&x=49&y=6

National Board for Professional Teaching Standards. (2005c). *NBPTS middle childhood generalist scoring guide.* Retrieved July 20, 2006, from http://www.nbpts.org/for_candidates/certificate_areas?ID=27&x=48&y=6

National Board for Professional Teaching Standards. (2006a). Retrieved July 20, 2006, from http://www.nbpts.org/.

National Board for Professional Teaching Standards. (2006b). *Certification areas.* Retrieved on June 29, 2006 http://www.nbpts.org/for_candidates/certificate_areas

National Board for Professional Teaching Standards. (2006c). *NBPTS early childhood generalist portfolio instructions.* Retrieved July 20, 2006, from http://www.nbpts.org/for_candidates/certificate_areas?ID=17&x=49&y=6

National Board for Professional Teaching Standards. (2006d). *NBPTS middle childhood generalist portfolio instructions.* Retrieved July 20, 2006, from http://www.nbpts.org/for_candidates/certificate_areas?ID=27&x=48&y=6

National Center for Education Statistics. (1999). *Teacher quality: A report on the preparation and qualifications of public school teachers.* (NCES 1999-080). Washington, DC: U.S. Department of Education/OERI.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform: a report to the nation and the Secretary of Education.* Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

National Evaluations Systems, Inc. (2006). *California subject examinations for teachers.* Amherst, MA: Author.

Phillip, R.A., Ambrose, R., Lamb, L.C., Sowder, J.R., Schappelle, J.C., Sowder, L., et al. (in press). Effects of early field experiences on the mathematical content knowledge and beliefs of prospective elementary school teachers: An experimental study. *Journal for Research in Mathematics Education.*

Pimm, D. (1987). *Speaking mathematically: Communication in mathematics classrooms.* London: Routledge.

Porter, A.C., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of research on teaching,* (4th ed., pp. 259–297). Washington, DC: American Educational Research Association.

Rotberg, I. C., Futrell, M. H., & Lieberman, J. M. (1998). National Board certification: Increasing participation and assessing impacts. *Phi Delta Kappan, 79*(6), 462–466.

Rowan, B., Chiang, F., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education, 70*, 256–284.

Rowan, B., Harrison, D., & Hayes, A. (2004). Using instructional logs to study elementary school mathematics: A close look at curriculum and teaching in the early grades. *Elementary School Journal, 105*, 103–127.

Sanders, W. L., Ashton, J. J., & Wright, S. P. (2005, March 7). *Comparison of the effects of NBPTS certified teachers with other teachers on the rate of student academic progress.* Retrieved June 21, 2006, from http://www.nbpts.org/pdf/sas_final_report.pdf

Sawada, D., & Pilburn, M. (2000). *Reformed teaching observation protocol.* (Tech. Rep. No. IN00-1). Arizona Collaborative for Excellence in the Preparation of Teachers: Arizona State University.

Schön, D. A. (1995). *The reflective practitioner: How professionals think in action.* Aldershot, England: Arena.

Schweingruber, H. A., & Nease, A. A. (2000, April). *Teachers' reasons for participating in professional development programs: Do they impact program outcomes?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Serafini, F. (2002). Possibilities and challenges: The National Board for Professional Teaching Standards. *Journal of Teacher Education, 53*(4), 316–327.

Shechtman, N., Roschelle, J., Knudsen, J., Vahey, P., Rafanan, K., Haertel, G., et al. (2006). SRI *Teaching survey: Rate and proportionality.* Unpublished manuscript, SRI International, Menlo Park, CA.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1–22.

Simon, M. (1993). Prospective elementary teachers' knowledge of division. *Journal for Research in Mathematics Education, 24*(3), 233–254.

Sowder, J. T., Phillip, R. A., Armstrong, B. E., & Shappelle, B. P. (1998). *Middle-grade teachers' mathematical knowledge and its relationship to instruction.* Albany, NY: SUNY Press.

SRI International. (2004, May). *Exploring differences in minority and majority teachers' decisions about and preparation for NBPTS certification* (SRI Project—P12209). Arlington, VA: Author.

Stein, M. K., Baxter, J. A., & Leinhardt, G. (1990). Subject-matter knowledge and elementary instruction: A case from functions and graphing. *American Educational Research Journal, 27*(4), 639–663.

Strauss, R. P., & Sawyer, E. A. (1986). Some new evidence on teacher and student competencies. *Economics of Education Review, 5*, 41–48.

Summers, A. A., & Wolfe, B. L. (1977). Do schools make a difference? *American Economic Review, 67,* 639–652.

Sumner, A. (1997). The toughest test. *Techniques, 71*, 28–31, 65.

Tatto, M. T., Nielsen, H. D., Cummings, W., Kularatna, N. G., & Dharmadasa, K. H. (1993). Comparing the effectiveness and costs of different approaches for educating primary school teachers in Sri Lanka. *Teaching and Teacher Education, 9*, 41–64.

TIMSS-R video math coding manual. (2003). Retrieved June 26, 2006, from http://www.lessonlab.com/TIMMS/download/TIMSS%201999%20Video%20Coding%20Manual.pdf

Tirosh, D., & Graeber, A. O. (1989). Preservice elementary teachers' explicit beliefs about multiplication and division. *Educational Studies in Mathematics, 20*, 79–96.

Tirosh, D., & Graeber, A. (1990). Evoking cognitive conflict to explore preservice teachers' thinking about division. *Journal for Research in Mathematics Education, 21*(2), 98–108.

Tracz, S. M., Sienty, S., Todorov, K., Snyder, J., Takashima, B., Pensabene, R., et al. (1995, April). *Improvement in teaching skills: Perspectives from National Board for Professional Teaching Standards field test network candidates.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46). Retrieved March 10, 2006, from http://epaa.asu.edu/epaa/v12n46/

Wakefield, D. (2003). Screening teacher candidates: Problems with high-stakes testing. *The Educational Forum, 67*(4), 380–388.

Walsh, K. (2001). *Teacher certification reconsidered: Stumbling for quality.* Baltimore: Abell Foundation.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73, 89–122.

Wilson, S., & Youngs, P. (2005). Research on accountability processes in teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 591–644). Mahwah, NJ: Erlbaum.

Wilson, S. M., Shulman, L. S., & Richert, A. (1987). 150 different ways of knowing: Representations of knowledge in teaching. In J. Calderhead (Ed.), *Exploring teachers' thinking* (pp. 104–124). Sussex, England: Holt, Rinehart & Winston.

Wineburg, S. (2004). Crazy for history. *The Journal of American History, 90*(4), 1401–1414.

AUTHOR NOTE

Work on this chapter was supported by grants from the National Science Foundation REC-0207649, EHR-0233456, and EHR-0335411. We would like to acknowledge the assistance of Richard Askey, Timothy Boerst, Catherine Brach, and Seán Delaney. We also thank Pamela Moss, Judith Sowder, Suzanne Wilson and Peter Youngs for their critical and helpful reviews of an earlier draft of this chapter. Errors and omissions remain the property of the authors.