Insights Gained from a Classroom-Based Assessment Project

CSE Technical Report 451

Lorrie A. Shepard

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)/
University of Colorado at Boulder

December 1997

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA Catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

INSIGHTS GAINED FROM A CLASSROOM-BASED ASSESSMENT PROJECT

Lorrie A. Shepard
CRESST/University of Colorado at Boulder

The rhetoric of assessment reform makes it sound as if new performance assessments will automatically improve teaching and learning. For example, Resnick and Resnick (1992) used research documenting the negative effects of traditional standardized testing to argue that because teachers teach toward high-stakes tests, assessments should be built intentionally to serve as the targets of instruction. In the past, teachers have emphasized tested subjects (math and reading) to the exclusion of social studies and science and have tailored worksheets, quizzes, and day-to-day instruction to closely resemble standardized test questions (Shepard, 1991). Therefore, advocates of assessment reform assume that if assessments are changed to reflect thinking and valued performances, instruction will be driven in the right direction.

According to the Resnicks (1992), "Assessments must be designed so that when teachers do the natural thing—that is, prepare their students to perform well—they will exercise the kinds of abilities and develop the kinds of skills and knowledge that are the real goals of educational reform" (p. 59). Wiggins (1989) said almost the same thing: "If tests determine what teachers actually teach and what students will study for—and they do—then the road to reform is a straight but steep one: test those capacities and habits we think are essential, and test them in context" (p. 41).

In this paper, I describe the experiences of teachers and researchers in a classroom performance assessment project. Our research team agreed with the Resnicks and Wiggins that introducing performance assessments aimed at thinking and problem-solving goals could be an important inducement for instructional improvement. We disagreed, however, that high-stakes consequences should be used to leverage change. Even authentic measures are corruptible and, when practiced for, can distort curriculum and undermine professional autonomy. We were more interested in a "bottom up" approach where teachers tried out performance measures in the context of classroom instruction.

The Assessment Project

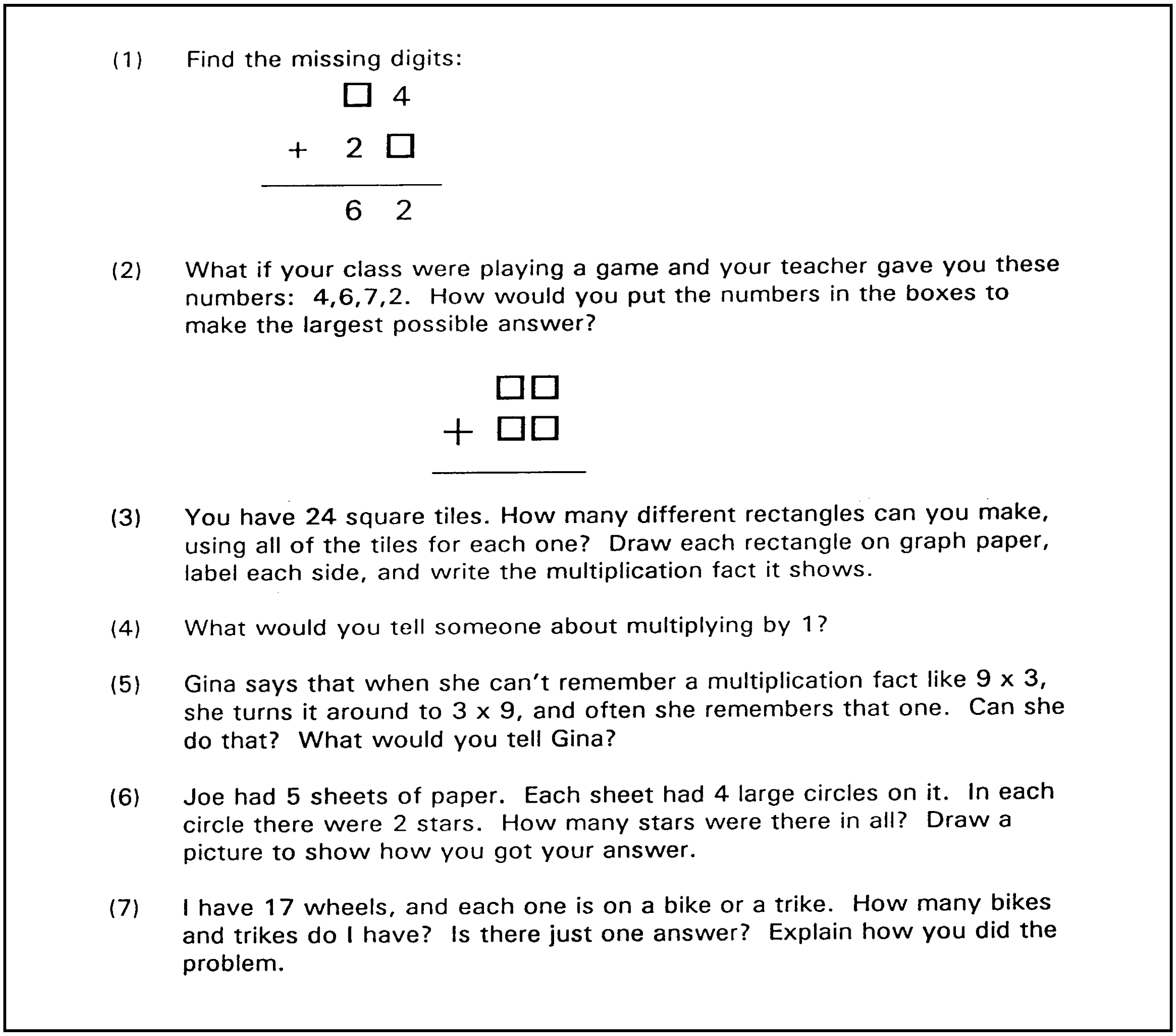
The intention of this project was not to introduce to teachers an already-developed curriculum and assessment package. Rather, we proposed to work with teachers to help them develop (or select) performance assessments in reading and mathematics congruent with their own instructional goals. I served as an assessment specialist on the project, and other members of the research team included Roberta Flexer, a specialist in mathematics education; Elfrieda Hiebert, a specialist in reading; and Hilda Borko, whose specialty is teacher change.

The study was conducted in a mixed lower and middle-class school district on the outskirts of Denver. We worked with teams of third-grade teachers in three schools. After a preliminary workshop, each school team submitted a proposal indicating its desire to participate. To ensure that teachers were free from the worry of preparing students for the CTBS, a 2-year waiver from standardized testing was obtained from the state; the waiver required a host of approvals from district officials, the teachers' union, and each school's parent accountability committee.

We met with teachers in spring 1992 and at the start of the 1992 school year to establish project goals. Then we met weekly for after-school workshops for the entire 1992-93 school year, alternating between reading and mathematics so that subject-matter specialists could rotate among schools.

In reading, teachers identified meaning making and fluency as the instructional goals to be assessed. "Running records" were used to assess fluency especially for below-grade-level readers. This meant listening to a child read individually and recording any misread words, with attention to the types of errors made. Initially, teachers used photocopied pages of text to keep track of errors but eventually could keep a running record on a separate notebook page. Written summaries were selected as a means to assess comprehension. In the fall, workshop discussions focused on issues of scoring written summaries and, because most students could not write adequate summaries at first, strategies to use for teaching students to write them. In the spring, ideas about meaning making and written summaries were extended to expository texts which had not previously been a part of the third-grade reading curriculum.

In mathematics, teachers identified place value, addition, subtraction, and multiplication as key instructional goals. They also made extensive requests throughout the year for materials and ideas for teaching in ways suggested by the new mathematics curriculum and for including new topics such as geometry and probability. Open-ended and nonroutine problems like those in Figure 1 were provided for teachers to try in their classrooms along with hands-on materials for modeling problems.



(1) Find the missing digits:

$$\begin{array}{rcc} & \square & 4 \\ + & 2 & \square \\ \hline & 6 & 2 \end{array}$$

(2) What if your class were playing a game and your teacher gave you these numbers: 4,6,7,2. How would you put the numbers in the boxes to make the largest possible answer?

$$\begin{array}{rcc} & \square & \square \\ + & \square & \square \\ \hline & & \end{array}$$

(3) You have 24 square tiles. How many different rectangles can you make, using all of the tiles for each one? Draw each rectangle on graph paper, label each side, and write the multiplication fact it shows.

(4) What would you tell someone about multiplying by 1?

(5) Gina says that when she can't remember a multiplication fact like 9 x 3, she turns it around to 3 x 9, and often she remembers that one. Can she do that? What would you tell Gina?

(6) Joe had 5 sheets of paper. Each sheet had 4 large circles on it. In each circle there were 2 stars. How many stars were there in all? Draw a picture to show how you got your answer.

(7) I have 17 wheels, and each one is on a bike or a trike. How many bikes and trikes do I have? Is there just one answer? Explain how you did the problem.

Figure 1. Sample open-ended math problems used for instruction and assessment in project classrooms.

Some teachers had not previously worked with place-value mats or manipulatives and introduced them for the first time. Discussions at biweekly meetings included dialogue about using good problems interchangeably for instruction and assessment, making observations and keeping track of them, analyzing student work, and developing rubrics for scoring problem solving and explanations.

The research part of the project also included baseline and follow-up measures of student learning in both participating and matched-control schools, interviews with parents, and interviews with students. Data about changes in instruction and assessment practices were gathered through teacher interviews, transcripts from the workshops, and samples of student papers, scoring rubrics, and the like.

Struggles

Despite successes reported later in this paper, it would be misleading to act as if the project encountered only smooth sailing. I report these struggles and difficulties because they are relevant to making practical suggestions for staff development and to informing policy discussion nationally about school delivery and opportunity-to-learn standards.

One of the first things we discovered was that not all 14 teachers in the project were true volunteers. Some had been implicitly volunteered by their principals or gently coerced by colleagues who wanted their team to participate. In addition, we were unprepared for the large conceptual differences between us and many of the teachers. Because teams had volunteered after hearing our project rationale and, because the district had newly developed curriculum frameworks consistent with emerging national standards in literacy and mathematics, we assumed that teachers' views about instruction would be similar to those reflected in the district curriculum and therefore similar to our own. In fact, even some teachers who were willing and energetic project participants were happy with the use of basal readers and chapter tests in the math book and were not necessarily familiar with curricular shifts implied by the new district framework in mathematics.

Dissonance between our and teachers' views about subject-matter instruction was sometimes acknowledged and joked about in workshops at two of the schools, but for the most part we avoided confrontations about differences in beliefs and did not propose radical changes in instruction. We suggested reading and mathematics activities that departed from a strictly skills-based approach, and these were adopted or adapted as teachers saw fit (Flexer, Cumbo, Borko, Mayfield, & Marion, 1994). We remained adamant about refusing to include timed tests on math facts as part of project portfolios but, given implied school policies and pressure from teachers in other grades, timed tests continued to coexist with project activities. In other areas where we did not confront differences, project-derived activities sometimes took on a character we would not have endorsed. For example, for some teachers writing summaries moved away from being a measure of how well a student understood the "gist" of a story and became more and more focused on features of the written product (Borko, Davinroy, Flory, & Hiebert, 1994). When some students were interviewed, they said that a good summary meant that handwriting was neat and all the words were spelled right (Davinroy, Bliem, & Mayfield, 1994). In math, teachers sometimes made problems easier or taught specific strategies for getting the answer that reduced the conceptual challenge of the problem.

For their part, teachers were equally disconcerted by project demands for which they had not bargained. Yes, everyone knew about the weekly workshops. But "trying out assessments in classrooms" had enormous implications. Early in the project, teachers' constant complaints were about *time*. There was not enough time to do extra planning, not enough time to meet as a team to talk about scoring or do the scoring (of reading summaries and math explanations), and most of all not enough time in the instructional day to teach reading both the old way and the new way and to add new math problems on top of old instructional routines. For some, a commitment to whole-class instruction made it impossible to manage time for individual assessments of below-grade-level readers.

Problems about time were negotiated and resolved in different ways. Instead of expecting that both reading and math activities would be tried out in classrooms every week on an ongoing basis, we slowed the pace for class-time demands by moving to a schedule of alternating weeks. We also modeled the use of running records during regular reading groups and discussed management strategies to make time for individual assessments of students most in need. We arranged university course credit, which had not originally been planned, and teachers used their usual team strategies to share the load; for example, in one school each member of the team developed a center for a unit on probability. The district also agreed to help by providing a half-day released time per month.

By January, some of the tensions about time in the instructional day had dissipated, perhaps because teachers were more comfortable incorporating new activities in place of the old. For other teachers, competition between instructional time and time spent on assessment increased as they struggled to keep written records of observations of students (Borko, Mayfield, Marion, Flexer, & Cumbo, in preparation). Other time pressures also increased as teachers “caught on” to the ideas of the project and needed more time to think and plan. For example, in February one team used an extra district inservice day to review the new math frameworks and decide “what the kids needed to know and how we were going to teach it.” In the words of one teacher, “that’s when we really sat down and started on this process” (not in September when the project nominally began).

Before turning to project successes, a word of caution is needed. Generalizations describing trends are the most useful to policy makers and teachers in other schools and districts, but generalizations about either difficulties or successes do not necessarily represent the experiences of individual teachers. In fact, the 14 project teachers represented a tremendous range of teaching styles and abilities. A few were already teaching in ways that coincided with the new curriculum frameworks; some pursued traditional content but were excellent classroom managers and challenged us intellectually about why we thought explanations and open-ended problems were a good thing. A few teachers participated in the project to the minimum extent possible and therefore were affected little by either the struggles or the successes.

Successes

The majority of teachers in our project could be described as effective teachers. When faced with new content, they did what good teachers do: They invented ways to help their students master the new material. (The Resnicks are right about this.)

For example, when third graders were first asked to write summaries of what they had read, they tended to produce long lists of events instead of focusing on the main points of the story or problem resolution. In workshop conversations, teachers concluded that helping students get better at summaries was worthwhile because it would help students pay attention to what was important in a story and because writing well-organized summaries was a good writing task as well. Teachers involved students in the development of scoring criteria and in grading each others' summaries. When students were at first indiscriminate in their scoring (everyone got a 3 on the 4-point scale), one teacher prompted a more thoughtful class discussion about "what makes a good summary" by writing some bad summaries for the students to score and then comparing these to summaries on book jackets and in advertisements. Teachers had their entire class read the same story and then develop a summary as a group. A class-authored summary elicited debate, suggestion by suggestion, as to the proper order of things and whether or not specific details needed to be included. Eventually students got much better at writing summaries because of teachers' effective use of modeling and class discussions around application of the scoring rubrics.

Teachers were also willing to try out a wide array of hands-on and problem-based activities in mathematics. One team invented its own activities for content that had not previously been included in third grade—geometry and probability—and it seemed to us that teachers were more inventive in devising challenging problems for students in these new content areas. One school had already used Marilyn Burns' (1991) 5-week multiplication program the year before, another school team decided to try it, and a third school used it in conjunction with help from a district math specialist. As one teacher explained, having the Burns' materials was especially useful because someone had already thought through the process and organized feasible instructional activities and accompanying assessments. This left teachers free to focus on implementation and their students' responses.

By the end of the year, most of the teachers were using math activities more closely aligned with the NCTM standards (National Council of Teachers of Mathematics, 1989) to replace and supplement more traditional practices of text-based work, and they had extended the range of mathematical challenges they thought feasible to attempt with third graders (Flexer et al., 1994). In end-of-year interviews, teachers reported that students now had a clearer understanding of why teachers graded papers the way they did (grading was less "mysterious") and, in both reading and math, teachers felt they had greater knowledge about what their students could do.

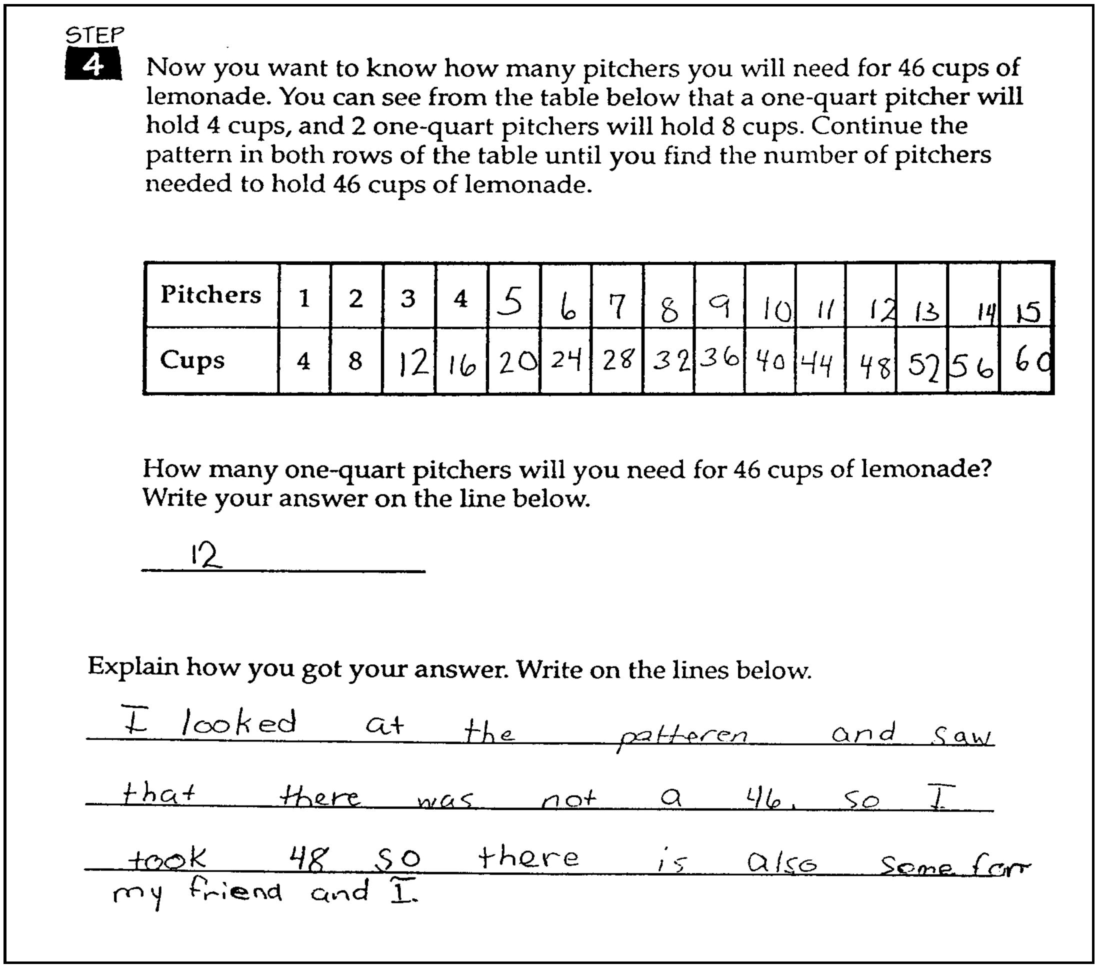
For many teachers, these "successes" continued into the next year. Specifically in two schools, by mutual agreement, we continued to meet with teachers on a less demanding schedule. In the spring of the second year, at an inservice given by one team for other teachers in their school, teachers made comments such as the following, which showed a thorough understanding of project issues and ownership of original project aims.

> **Teacher 1.** The [researchers] felt that if you taught the strategy that was just like teaching an algorithm. So we wanted the kids to pull their own resources from their head. Each day we gave them problems that would be different and use a different strategy. We didn't want them to think that yesterday we used a table and then look at a problem today and automatically make a table. . . .
>
> During my debriefing time on the following day I put on a poster paper three different ways that three different children solved the problem to show them that there is more than one way to solve the problem.

> **Teacher 2.** Confession time here. I remember five years ago, three years ago, we lived with that textbook. Well, we haven't used our math book too much this year. Maybe a little bit as a resource, but it's not like, page 36 today, 38 tomorrow, and the next day is more on 40. . . . So what I'm going to show you are just some examples that two years ago I would have said, "No way, third graders cannot handle this." But it's amazing, they can when they're exposed to it.

Teachers eventually developed greater sophistication about scoring criteria and revisited assessment issues that had been problematic early on. For example, the difficulties some students have with writing had been a concern for both written summaries and explanations in math. (Indeed, the problem of writing and language skills having an undue influence in more authentic assessments is an issue nationwide.) Initially some teachers allowed specific students other opportunities to "show their thinking" by asking for oral retellings or letting a student explain a picture in math. At the same time, early scoring rules for some teachers included spelling and handwriting and other features of the written product. After a year, teachers were much more aware that scoring rules should depend on what you were *scoring for* (i.e., the intended construct in measurement terms). Working with expository text in the spring semester, which third-grade teachers had never done before, seemed to help refocus attention on the purpose of summarizing (Davinroy & Hiebert, 1994). Furthermore, teachers became clearer about the multiple dimensions buried in their scoring rubrics, a tension that some resolved by giving two scores—for example, one for reading comprehension and one for writing, or one for the explanation and one for getting the right answer. Rubrics in math also shifted to focus more on the reasonableness of the answer and explanation rather than the accuracy of the calculation.

Last but not least, project successes also included appreciable gains in student learning in mathematics. Independent outcome measures showed no change in reading performance, possibly because both participating and control schools were further along in implementing new language arts curricula at the start of the project than they were in math. The specific gains in mathematics could be tied directly to project activities. Especially apparent were students' abilities to recognize and extend patterns and to write mathematical explanations. Figure 2 shows one student's answer to one segment of a problem from the 1991 Maryland math assessment used as an outcome measure.



STEP **4** Now you want to know how many pitchers you will need for 46 cups of lemonade. You can see from the table below that a one-quart pitcher will hold 4 cups, and 2 one-quart pitchers will hold 8 cups. Continue the pattern in both rows of the table until you find the number of pitchers needed to hold 46 cups of lemonade.

Pitchers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cups	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60

How many one-quart pitchers will you need for 46 cups of lemonade? Write your answer on the line below.

12

Explain how you got your answer. Write on the lines below.

I looked at the patteren and saw
that there was not a 46, so I
took 48 so there is also some for
my friend and I.

Figure 2. Sample student response to one step from a Maryland math assessment task.

At the top end of the distribution, in the participating schools, many more students could give a complete answer after the project than could do so in the baseline year. More importantly, there were demonstrable gains at the low end of the distribution as well. In the baseline year, most low-scoring students could not fill out the table in Step 4 and could not write an explanation. After the project year, low-performing students in the participating schools most frequently gave wrong answers of either "15" or "60"; nonetheless, they completed the table, and gave explanations with real mathematical content (for which they received partial credit):

> "I counted by fours which is 60 then I went in the ones which is 15."
>
> "On the cups as you go along you count four more each time."

Implications for Staff Development

What implications for staff development can be drawn from our classroom-based assessment project? Superficial change is easy; fundamental changes are much more difficult. Current calls for assessment-driven reform acknowledge the need for staff development but tend to underestimate the extent and depth of what is needed. While teaching toward open-ended tasks might be an immediate improvement over worksheets designed to mimic standardized tests, our experience shows that well-intentioned efforts to help students get good at assessment tasks can be misdirected if teachers do not understand the philosophical and conceptual bases of the intended curricular goals. "Why should our kids have to write explanations, if we already know whether they know it or not?"

Losing track of written summaries as an assessment of meaning making and oversimplifying mathematical problem solving are examples of how the original purpose of assessment tasks could be distorted. Given ongoing project discussions, teachers eventually recognized that these issues were problematic and addressed them with deepening understanding. It is hard to imagine how such detailed project insights could have been handed out in a one-time inservice session or inferred by teachers alone if external assessments were the only mechanism for instructional reform.

To make changes that are conceptually meaningful, teachers need support on an ongoing basis.

1. They need appropriate materials to try out and adapt.

2. They need *time* to reflect and to develop new instructional approaches.

3. They need support from experts to learn (and challenge) the conceptual basis behind intended reforms.

The need for materials poses several interesting dilemmas. Professional, autonomous teachers do not need canned curriculum packages or scripted lessons. However, if we want teachers to try significantly different content and modes of instruction, teachers in our project would argue that they have neither the time nor the know-how (initially) to invent their own materials. It did not even help to have an abundant supply of materials in the curriculum library, because teachers did not have time to review them and they did not know which were good and which were not. What worked best in our project was for us to supply good examples in response to teacher-identified topics. Then teachers were excellent at extending the examples and inventing entire instructional units.

Teachers learned the most by trying new, challenging content with their students and by being surprised by what their students could do. Just as constructivist pedagogy would allow students the opportunity to develop their own understandings, teachers need the opportunity to try new instructional strategies, observe what works and what doesn't, and then talk with colleagues about both logistics and underlying rationale.

It is perhaps not surprising that, as a university researcher, I also make a place in this conversation for talking with "experts." However, experts include curriculum specialists and lead teachers in school districts as long as they have a thorough understanding of the conceptual basis behind content standards and curriculum frameworks. If teachers are unfamiliar with new curriculum expectations, they cannot be expected to appreciate from first-time tellings why making connections and communication are important mathematical goals or what the proper place of skill instruction should be in developing literacy. But these are the kinds of questions that are best returned to and debated with specialists after teachers have had some first-hand experience with new content in their own classrooms. For example, one team of teachers attended a district inservice on assessment at the start of the second school year; they came back and asked why we had never told them about scoring for more than one dimension. The answer was we had "told them" (one school was already using double scoring), but expert advice had not made sense until teachers had had relevant experience with that issue in their own classrooms.

The "successes" of our assessment project support the claims of assessment reform advocates, albeit on a much more modest and tentative scale. Performance assessments have great potential for redirecting instruction toward more challenging and appropriate learning goals. Open-ended assessment tasks not only prompted teachers to teach differently, but criteria were made explicit, and students learned more. However, the concomitant "struggles" give the lie to the presumption that new assessments will automatically improve instruction. If teachers are being asked to make fundamental changes in what they teach and how they teach it, then they need sustained support to try out new practices, learn the new theory, and make it their own.

References

Borko, H., Davinroy, K. H., Flory, M. D., & Hiebert, G. H. (1994). Teachers' knowledge and beliefs about summary as a component of reading. In R. Garner & P. A. Alexander (Eds.), *Beliefs about texts and instruction with text* (pp. 155-182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Borko, H., Mayfield, V., Marion, S. F., Flexer, R., & Cumbo, K. (in preparation). *Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development*. Denver: University of Colorado, School of Education.

Burns, M. (1991). *Math by all means: Multiplication grade 3*. Sausalito, CA: The Math Solution Publications.

Davinroy, K. H., Bliem, C. L., & Mayfield, V. (1994, April). "*How does my teacher know what I know?": Third-graders' perceptions of math, reading, and assessment*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Davinroy, K. H., & Hiebert, E. H. (1994). An examination of teachers' thinking about assessment of expository text. In D. J. Leu & C. K. Kinzer (Eds.), *43rd NRC Yearbook* (pp. 60-71). Chicago: National Reading Conference.

Flexer, R. J., Cumbo, K., Borko, H., Mayfield, V., & Marion, S. F. (1994, April). *How "messing about" with performance assessment in mathematics affects what happens in classrooms*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan*, *72*, 232-238.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership*, *46*(7), 41-47.