# Effects of Introducing Classroom Performance Assessments on Student Learning

Lorrie A. Shepard, Roberta J. Flexer, Elfrieda H. Hiebert, Scott F. Marion, Vicky Mayfield, and Timothy J. Weston
*University of Colorado, Boulder*

*How can performance assessments be used as part of regular instruction? Will this raise student performance on external achievement measures? What aspects of examinee performance improve on the assessment exercises?*

Arguments favoring the use of performance assessments make two related but distinct claims. Performance assessments are expected, first, to provide better measurement and, second, to improve teaching and learning. Although any measuring device is corruptible, performance measures have the potential for increased validity because the performance tasks are themselves demonstrations of important learning goals rather than indirect indicators of achievement (Resnick & Resnick, 1992). According to Frederiksen and Collins (1989), Wiggins (1989), and others, performance assessments should enhance the validity of measurement by representing the full range of desired learning outcomes, by preserving the complexity of disciplinary knowledge domains and skills, by representing the contexts in which knowledge must ultimately be applied, and by adapting the modes of assessment to enable students to show what they know. The more assessments embody authentic criterion performances, the less we have to worry about drawing inferences from test results to remote constructs.

The expected positive effects of performance assessments on teaching and learning follow from their substantive validity. If assessments capture learning expectations fully, then when teachers provide coaching and practice to improve scores, they will directly improve student learning without corrupting the meaning of the indicator. Resnick and Resnick (1992), Frederiksen and Collins (1989), and Wiggins (1989) all argue that it is natural for teachers to work hard to prepare their students to do well on examinations that matter. Rather than forbid "teaching to the test" which is impossible, it is preferable to create measures that will result in good instruction even when teachers do what is natural. The reshaping of instruction toward desirable processes and outcomes is expected to occur both indirectly, as teachers individually imitate assessment tasks in a variety of ways, and directly, because expectations and criteria for judging performances will be shared explicitly.

These anticipated benefits of performance assessments have been inferred by analogy from research documenting negative effects of traditional, standardized testing. Under conditions of high-stakes accountability pressure, it has been demonstrated that elementary teachers in particular align instruction with the content of basic skills tests, often ignoring science and social studies and even untested objectives in reading and mathematics. In addition, instruction on tested skills comes to resemble closely the format of multiple-choice tests (Madaus, West, Harmon, Lomax, & Viator, 1992; Shepard, 1991; Smith, Edelsky, Draper, Rottenberg, & Cherland, 1990). Such practices may limit the generalizability of test score results and ultimately harm learning if students have not really mastered the intended skills. For example, Cohen (1987) cited an experiment (Koczor, 1984) where students who drilled on translating Roman numerals to Arabic numerals had 40% greater mastery when tested in the same format than when asked to convert from Arabic to Roman numerals. Some measurement specialists have speculated that this type of narrow teach-

*Lorrie A. Shepard is a Professor of Education at the University of Colorado, Campus Box 249, Boulder, CO 80309-0249. Her specialization is educational measurement.*

*Roberta J. Flexer is an Associate Professor of Education, University of Colorado, Campus Box 249, Boulder, CO 80309-0249. Her specialization is mathematics education.*

*Elfrieda H. Hiebert is a Professor of Education at the University of Michigan, 4204E SEB, 1259, Ann Arbor, MI 48109. Her specializations are literacy instruction and assessment.*

*Scott F. Marion is a PhD Candidate at the University of Colorado, Campus Box 249, Boulder, CO 80309-0249. His specialization is educational measurement.*

*Vicky Mayfield is a PhD Candidate at the University of Colorado, Campus Box 249, Boulder, CO 80309-0249. Her specialization is mathematics education.*

*Timothy J. Weston is a PhD Candidate at the University of Colorado, Campus Box 249, Boulder, CO 80309-0249. His specialization is educational measurement.*

ing to the test during the 1980s may explain inflated test score gains on state accountability tests compared to more modest gains on the National Assessment of Educational Progress (Linn, Graue, & Sanders, 1990).

Based on negative examples and evidence, advocates of performance assessments assume that parallel mechanisms will work to produce positive effects once limited tests are replaced by more desirable measures. However, to date little research has been done to evaluate the actual effects of performance assessments on instructional practices or on student learning. Although some extreme views hold that authentic performance measures are valid by definition and will automatically produce salutary effects, we would argue in contrast that the effects of performance assessments should be evaluated empirically following a program of inquiry closely parallel to the studies undertaken to examine the effects of standardized tests. We concur with Linn, Baker, and Dunbar (1991) that validity criteria for alternative assessments should address intended and unintended effects as well as more substantive features such as cognitive complexity, content quality and comprehensiveness, generalizability of knowledge from assessed to unassessed tasks, and the like. Although we are committed to developing performance assessments because they can more completely represent challenging subject matter, their demonstrated effects on teaching and learning remain an open question.

The purpose of the present study was to examine the effects of performance assessments on student learning. If teachers begin to use more open-ended and authentic tasks as part of regular instruction with appropriate feedback and instructional support for students, will student performance on independent measures of achievement be improved? Note that some arguments favoring the use of performance assessments to leverage educational reform presume that the high-stakes accountability pressures are still needed to drive instructional change. Other advocates focus more on the informational and modeling effects of classroom-embedded assessments. In this study, we adopted the second perspective. Performance assessments are expected to improve learning in two ways: (a) Content will be improved by use of challenging tasks consistent with curricular goals, and (b) teachers will have clearer knowledge of their students' understandings to inform their instruction. We were interested in the effects of using new forms of assessment as part of instruction but without the incentives and context created by an externally mandated system.

A year-long project was undertaken to help teachers in 13 third-grade classrooms begin to use performance assessments as a part of regular instruction in reading and mathematics. Other parts of the research project focused on changes in teachers' beliefs and instructional practices in reading and mathematics (Borko, Davinroy, Flory, & Hiebert, 1994; Borko, Mayfield, Marion, Flexer, & Cumbo, in press; Davinroy & Hiebert, 1993; Flexer, Cumbo, Borko, Marion, & Mayfield, 1994); but here research questions are focused on student achievement in reading and mathematics. Did students learn more or develop qualitatively different understandings because performance assessments were introduced into classrooms? Achievement results were compared both to the performance of third-grade students in the same schools the year before and to third-grade performance in matched-control schools.

## Study Methods

### *Setting*

The study was conducted in a working-class and lower-to-middle-class school district on the outskirts of Denver, Colorado. The district was selected in part because of the willingness of central office administrators to participate and in part because of its ethnically diverse student population. In the 1980s the district was known for its extensive mastery learning and criterion-referenced testing system, but more recently curriculum guidelines in language arts and mathematics were revised to reflect more constructivist conceptions of these disciplines, consistent with national standards (Anderson, Hiebert, Scott, & Wilkinson, 1985; National Council of Teachers of Mathematics, 1989).

### *Sample and Research Design*

Third grade was selected as the target grade level because the Comprehensive Test of Basic Skills (CTBS) is administered district-wide in Grade 3. Because project participation would involve a 2-year waiver from standardized testing and an extensive commitment of teacher time, prospective schools were asked to submit proposals documenting support for the project from the principal, the parent accountability committee, and the entire third-grade team of teachers. Although 10 schools were represented at an initial information workshop, only three schools completed the formal application and were accepted as participants. In the 1992–1993 study year, there were 13 third-grade classrooms in the three schools combined involving approximately 335 third graders.

Three control schools were identified to be used for comparison when analyzing teachers' beliefs and parents' opinions as well as students' achievement. The control and participating schools were matched on free- and reduced-lunch percentages, percentage of minority children, and other knowledge of neighborhood similarities, such as type of housing. Data in Table 1 show the socioeconomic differences among the three participating schools as well as their matches to control schools. The two sets of schools were also compared on CTBS achievement test scores from prior years and 1992 premeasures. Even with 20 elementary schools to choose from, we found that it was not possible to match schools on both socioeconomic factors and CTBS scores because they diverged too much. Several possible control schools had higher test scores than the participating schools. Because we could not know whether sharp differences in achievement scores meant more able populations, more able teaching, or even more test-score inflation in some control schools, we elected to match only on socioeconomic data. As we gathered more data, the most likely explanation for the divergence in test scores appeared to be stronger mathematics instruction in the control schools at the start of the study rather than differences in student

**Table 1**

*Demographic and Achievement Characteristics of Participating and Control Schools*

| Demographic characteristics | Participating schools | | | Control schools | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 |
| Free and reduced lunch | 61% | 9% | 6% | 55% | 13% | 3% |
| Percent minority | 37% | 16% | 14% | 45% | 19% | 10% |
| Student turnover | 27% | 7% | 11% | 30% | 11% | 10% |
| 5-Year average 3rd-grade CTBS scores (1987–1991) | | | | | | |
| Total reading | 47.8 | 48.8 | 52.7 | 48.9 | 50.4 | 54.7 |
| Total mathematics | 52.5 | 47.5 | 51.3 | 49.3 | 60.9 | 58.1 |
| Baseline year 3rd-grade CTBS scores (Spring 1992) | | | | | | |
| Total reading | 44.6 | 51.9 | 55.5 | 43.1 | 54.5 | 57.2 |
| Total mathematics | 53.9 | 53.8 | 62.8 | 47.9 | 66.5 | 68.1 |
| Fall 1992 pretest measures for entering 3rd graders | | | | | | |
| Reading | 9.8 | 12.2 | 13.2 | 9.0 | 11.6 | 10.9 |
| Mathematics | 14.6 | 21.9 | 22.4 | 16.7 | 21.9 | 20.4 |

populations. In addition to the matched socioeconomic data, premeasures administered to entering third graders in the Fall of 1992 show how similar the populations were in the matched-pairs of schools. End-of-year data on the prior years' third graders, however, showed superior math performance for the control schools not only on the CTBS but also on our independently administered baseline achievement measures (Spring 1992, see Table 2).

The research design called for two separate comparisons. Outcome measures in reading and mathematics selected for administration in May 1993 were also administered as baseline measures in May 1992. In addition, premeasures appropriate for entering third graders were administered in September 1992 and used as covariates to evaluate 1993 outcomes.

**Table 2**

*1992 Versus 1993 Comparisons in Reading and Mathematics for Participating and Control Schools*

| Outcome measures | 1992 Mean (*n*) | 1993 Mean (*n*) | 1992–1993 Mean difference | 1992 Pooled w/in school *SD* | ES* of difference |
|---|---|---|---|---|---|
| Maryland reading total | | | | | |
| Participating | 27.7 (290) | 26.1 (305) | –1.6 | | –.14 |
| | | | | 11.7 | |
| Control | 28.9 (210) | 26.5 (228) | –2.4 | | –.21 |
| Maryland math total | | | | | |
| Participating | 12.2 (288) | 13.0 (305) | 0.8 | | .13 |
| | | | | 5.94 | |
| Control | 15.3 (210) | 13.6 (231) | –1.7 | | –.29 |
| Alternative math total | | | | | |
| Participating | 12.7 (288) | 12.9 (305) | 0.2 | | .06 |
| | | | | 3.5 | |
| Control | 13.3 (208) | 13.5 (229) | 0.2 | | .06 |

*Effect size calculations are based on pooled within-school 1992 standard deviations using both participating and control group schools.

*Assessment Project "Intervention"*

An adequate description of the project "Intervention" requires an understanding of both original intentions and of changes that were made in response to the reality of teachers' practices and perspectives. Because the district had in place curriculum frameworks consistent with emerging national standards in reading and mathematics, and because teachers had volunteered to participate in the project after seeing examples of the kinds of assessments envisioned, we assumed that their views about instruction would be similar to our own. Our intention was not to make wholesale changes in instruction, and we did not arrive with a predesigned curriculum and assessment package. Rather, we proposed to work with teachers to help them develop (or select) performance assessments congruent with their own instructional goals. Four faculty researchers offered expertise in mathematics, reading, teacher change, and assessment. After-school workshops were held each week for the entire 1992–1993 school year, alternating between reading and mathematics so that subject-matter specialists could rotate among schools.

Once in the schools, we learned that our assumptions about instructional practice being congruent with the district frameworks were not accurate except for a few classrooms. Not all teachers were true volunteers; some had been "volunteered" by their principals or had acceded to pressure from the rest of the third-grade team. More importantly, even some teachers who were willing and energetic project participants were happy with the use of basal readers and chapter tests in the math text and, especially in mathematics, were not necessarily familiar with curricular shifts implied by the district frameworks.

Beginning-of-the-year interviews with teachers documented similar instructional practices in reading across participating and control schools. All teachers were familiar with the district "significant learnings," adopted 3 years before, which were "whole-language based." Nearly all reported using Readers' Workshop and mini-lessons to teach skills and strategies in context. However, practices could best be described as a blend of old and new. Many teachers were still primarily using basals with literature books as supplements, while others had developed extensive libraries of chapter books and picture books. Consistent with a whole-language approach, teachers said first that they want children to enjoy reading and were helping children learn to select appropriate books, but they also emphasized more traditional approaches to comprehension with plenty of structured practice on identifying story elements: plot, characters, setting. Although some individual teachers were further along in implementing whole-language literacy instruction, there were no systematic differences between participating and control classrooms, with the exception of one participating teacher who began the year with phonics workbooks.

In mathematics, the range of instructional practices at the start of the project was much greater. Teachers in both control and participating schools tended to be less familiar with the substance of the district framework adopted the previous spring. Some teachers described relatively traditional approaches to mathematics instruction: whole-group instruction followed by individual practice supported by manipulatives, math facts and story problems, textbook problems and worksheets. Teachers who reported instruction congruent with the NCTM standards and the district framework were rare, but they were represented equally in control and participating schools. Their goals included instruction in measurement, geometry, and probability as well as numeration and computation; they talked about number sense, estimation, reasonableness, hands-on activities, and modeling; instruction involved centers, small groups, and materials such as Marilyn Burns's Math Solutions, and Activities Integrating Mathematics and Science (AIMS). In the middle of the range were teachers who were following a traditional text but supplementing it with problem-solving strategies and possibly a Marilyn Burns unit. For the most part, this range characterized both participating and control schools. However, in one matched pair of schools, control teachers had already tried using Marilyn Burns, while two participating teachers began the year teaching children to copy problems from the book according to a specific format.

Given some dissonance between researchers' and teachers' views about subject matter instruction, we looked for areas of agreement as the place to focus our joint efforts. In reading, teachers identified meaning-making and fluency as instructional goals for which they would like to develop more systematic assessments. Hiebert suggested that running records be used especially with below-grade-level readers to assess both fluency and meaning-making. Teachers requested help with the logistical difficulties in finding time for one-on-one assessment; this was especially difficult for one teacher who used only a whole-class approach. They also asked for help with strategies for teaching word-attack skills to third graders who were not yet decoding fluently. Written summaries were used to assess comprehension. In workshop sessions, teachers discussed students' first attempts at writing summaries, developed scoring rubrics, and over time presented the various activities they had devised to help students get better at writing summaries—discussing good summaries on book jackets and bad ones written by the teacher, writing a summary together as a class, and having students critique one another's work. In the spring semester, ideas about meaning-making and written summaries were extended to expository texts.

In mathematics, teachers identified place value, addition and subtraction, and multiplication as foci for assessment. Flexer began by providing examples of more open-ended problem-oriented, and hands-on tasks that could be used for either instruction or assessment. (Our position is not that the exact same task should be used to teach and then to measure, because it is important to check for generalizability and transfer. However, good tasks can be used for either purpose.) Early in the year, teachers made extensive requests for materials and methods to support the kind of teaching advocated by us and the district framework, and these were provided—for example,

materials were distributed for making base 10 blocks for modeling numbers and operations. Some teachers had not previously worked with place-value mats or manipulatives and introduced them for the first time. Games and "family math" activities were supplied in response to questions about how to provide practice without relying on rote memorization.

Asking students to write explanations about how they did a problem or why an answer seemed reasonable became a regular part of many open-ended problems. In workshop sessions, scoring rubrics were developed by first discussing student work, then identifying features of excellent work, and then formally defining levels of the rubric. Based on the positive experiences of those who had already used the Marilyn Burns multiplication unit, others decided to try it as well. These activities, engaged in over several weeks, help children visualize multiplication as repeated addition, as the geometry of arrays and areas, and as patterns (counting by 5s) and functions (6 cows, how many legs?). Teachers also asked for ideas about teaching new topics in the third-grade curriculum, such as geometry and probability. In one school, each of the teachers agreed to create a center dealing with some aspect of probability that then all the classes could use.

*Outcome Measures and Covariates*

For obvious reasons, we did not wish to use a multiple-choice standardized test to measure the project's effects. At the same time, a compendium of performance tasks used throughout the project would also not be a fair outcome measure. The 1991 Maryland School Performance Assessment Program was selected to measure achievement because it provided broad coverage of the district curriculum frameworks in both reading and mathematics and had a mixture of right–wrong and open-ended questions. In literacy, students read extended stories and informational texts in a separate reading book and then wrote responses about what they read, completed tables, drew story webs, and so forth. In mathematics, the three tasks selected each involved a series of problems all related to the same information source or application. Students had to solve problems that involved identifying and extending patterns, comparing and ordering quantities, estimating, multiplying and adding, doubling and halving, using calculators, computing area, and explaining how they got their answers.

To be sure to assess a range of skills in mathematics, we supplemented the three tasks from the Maryland assessment with a portion of an alternative measure developed for another study (Koretz, Linn, Dunbar, & Shepard, 1991). This test consisted of 15 short-answer and multiple-choice items that assess problem solving in, and conceptual understanding of, functions and relations, patterns, whole-number operations, probability, and data and graphs.

Covariate measures were needed for entering third graders to assess their initial abilities in reading and mathematics. In reading, portions of a Silver, Burdett and Ginn 2/3 Reading Process Test and 2/3 Skills Progress Test were used with permission from the publisher. In mathematics, open-ended problems were developed to measure students' ability to discern patterns and number relations. This subtest was combined with three subtests from the second-grade level of the Iowa Test of Basic Skills covering math concepts, estimation, and data interpretation.

*Scoring and Reliability*

All of the measures used in the study required scoring of open-ended student responses. Scorers worked from the scoring guides provided by the Maryland School Performance Assessment Program with slight modifications made by the respective subject-matter experts. Day-long training sessions were held in Summer 1992 and again in 1993 to ensure that scorers were familiar with the scoring rules and able to apply them consistently.

Inter-rater reliability was assessed both within year (are all of the scorers rating consistently?) and between years (were the scoring rules implemented consistently in 1992 and 1993?). For the within-year studies, three student booklets in reading and three in mathematics were chosen at random from each classroom. This resulted in more than a 10% sample with 55 to 60 booklets being rescored in each set of 500. Booklets were scored independently by the scorer-trainer. Pearson correlations between total scores assigned by other raters and by the standard rater were quite high in both years for both reading and mathematics; values ranged from .96 to .99.

The Maryland reading measure was composed of 61 scored *items* or task subparts; the Maryland mathematics measure had 31 scorable entities. The high correlations between raters simply mean that, with sufficient numbers of task subscores, raters can rank students quite accurately. Absolute agreements on total score provided a sterner test of rater consistency. Within years, raters agreed with the standard rater on total score within 1 or 2 points for 97% or 98% of cases in reading and for 90% to 91% of cases in mathematics. These agreement rates are respectable for subjectively scored instruments but nonetheless introduce noise into the evaluation of effects.

To check for consistency of scoring across years, test booklets from 1992 were seeded into 1993 classroom sets without scorers being aware of which booklets were being rescored. A total of 57 booklets were rescored in both mathematics and reading. The between-year agreements were not so high as the within-year agreements. In mathematics, 79% of total scores were within 2 points of the score assigned to the same booklet the year before. In reading, 72% were within 4 points (which is comparable in standard deviation units to a 2-point difference on the mathematics assessment). The between-years analysis also revealed some systematic biases, with raters tending to become more stringent in 1993 than raters had been in 1992. In reading, there was an average mean score shift downward of 2.47 points for the 57 1992 booklets rescored in 1993. In math, the greater stringency created a downward shift of .25 points. Because the reading score shift was both statistically and practically significant, 1993 reading scores were adjusted to correct for the systematic bias. Average biases varied for indi-

vidual raters from 1.13 to 3.63, all in the direction of greater stringency; these specific corrections were applied to the sets of booklets scored by each rater.

Internal consistency coefficients provide another indicator of the psychometric adequacy of research instruments. Based on the entire sample, which varied by instrument from 487 to 524, the respective 1992 and 1993 coefficients were .90 and .90 for the Maryland reading assessment, .84 and .83 for the Maryland mathematics assessment, and .78 and .80 for the alternative mathematics assessment.

## Results

Data comparing the 1992 and 1993 end-of-year assessments for both participating and control schools are reported in Table 2. Overall, the predominant finding is one of no-difference or no gains in student learning following from the year-long effort to introduce classroom performance assessments. Although we argue subsequently that the small year-to-year gain in mathematics is real and interpretable based on qualitative analysis, honest discussion of project effects must acknowledge that any benefits are small and ephemeral. For example, improvements occurred in some project-teachers' classrooms but not in all, and the gain from 1992 to 1993 for the participating schools on the Maryland mathematics assessment had an effect size (ES) of only .13.

In reading, there were no significant differences (alpha = .05) between 1992 and 1993 results or between participating and control schools. Both groups of schools appeared to lose ground slightly (.9 and 1.9 points, respectively). Analysis of covariance tables is not shown because, in both reading and mathematics, matched schools were so similar on the premeasures that adjustments did not alter the findings of essentially zero difference between participating and control schools at the end of the project on all three outcome measures.

In mathematics, the alternative test also showed no effects. However, the Maryland assessment in mathematics, which requires students to do more extended problems and explain their answers, showed an improvement in the participating schools. We interpret this change, albeit small, as a "real" gain based on the following arguments. First, CTBS results for 1993 showed declines district-wide and in two of the control schools. Against a backdrop of declining achievement, slight gains in the participating schools are more important. Although the populations of the participating and control schools are quite similar as evidenced by socioeconomic variables and fall pretest measures, third graders in the control schools had traditionally outperformed third graders in the participating schools. This was apparent in 5 years of CTBS data and on the 1992 baseline measure in mathematics. In fact, the only significant differences in Table 2 are the differences between the control and participating schools on the Maryland mathematics ($p < .0001$) and alternative mathematics assessments ($p = .053$) the spring before the study began. Therefore, one way of interpreting the between-year and covariance analyses together is to say that the assessment project helped participating students catch up to the control students in math achievement. From all indications, this would not have occurred without the project.

In an effort to understand the substantive nature or character of the change on the Maryland math assessment, qualitative analyses were conducted of student responses. Coding categories were developed for each task or task subpart based on a sample of student papers. Then these categories were applied systematically to all of the papers in the two or three participating classrooms per school with large effect sizes and to their matched controls. Analyses of this type were carried out for two of the three multiquestion tasks.

From the qualitative analyses, we noted consistent changes in students' answers to math problems which suggest that at least in some project classrooms whole groups of students were having opportunities to develop their mathematical understandings that had not occurred previously. Figures 1 and 2 and Table 3 were constructed to provide a qualitative summary of student responses to a task subpart and to illustrate what small improvements in student scores may mean substantively. The two classrooms that showed the greatest gains from 1992 to 1993 in the low socioeconomic participating and control schools were one of the matched pairs selected for comparison (Table 3). Both teachers' classrooms showed an effect-size gain of .27 from 1992 to 1993 on the Maryland mathematics assessment. However, for this particular problem, there was a noted improvement in partial credit for students in the participating classroom that did not occur in the matched class. This shift suggests that a greater proportion of this teacher's classroom of typically poorly performing students could recognize patterns and complete numeric tables than could do so in the previous year. At the top of the scale, there were no more right answers in 1993 than in 1992. However, in 1993, 84% of the children in the participating classroom could complete the table (Categories I–V), whereas in 1992 only 34% of the same teacher's students could complete this part of the problem. The percentage of students in the participating classroom who could write explanations describing a mathematical pattern or telling how they used the table (Categories I, III, or IV) also increased substantially, from 13% to 55%. Even students who took the *wrong* answer from their table could describe the pattern:

- I counted by fours which is 60 the[n] I went in the ones which is 15.
- I counted by 4 and ones and came to 60.
- First I went up to 15 pitchers. Then I made 60 cups.
- First I cont'd by one's then I contid by fors. (Answer 60)
- First I saw that the[y] where counting by 4s So I counted by fours. until there was no rome and got the answer 57.
- I counted by 4s and I lookt at the top one. (Answer 15)

In the matched control low-SES classroom, the percentage of students writing explanations actually declined from 39% to 23%. For these two teachers to have had the same positive gain in total score, there

**Table 3**

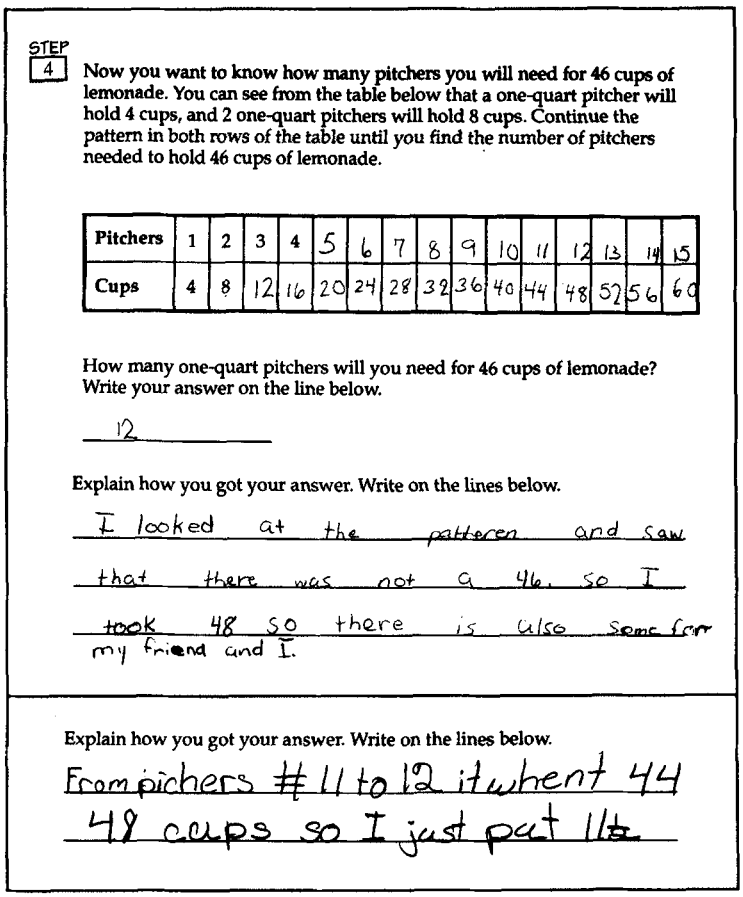
*Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) From the Classrooms With the Greatest Gains in Low Socioeconomic Participating and Control Schools*

| Qualitative categories | Participating | | Control | |
|---|---|---|---|---|
| | 1992 | 1993 | 1992 | 1993 |
| I. Extends table, answers correctly, explains (explains either pattern or point in chart) | 13% | 13% | 31% | 19% |
| II. Extends table, answers correctly, inadequate explanation | 4% | 0 | 8% | 12% |
| III. No answer but stops table at right place, explanation describes pattern | 0 | 0 | 0 | 0 |
| IV. Extends table, wrong answer (60, 15, 11, other), explanation describes pattern | 0 | 42% | 8% | 4% |
| V. Extends table, wrong answer (60, 15, 11, other), inadequate explanation | 17% | 29% | 8% | 35% |
| VI. Cannot extend table | 63% | 8% | 46% | 31% |
| VII. Blank | 4% | 8% | 0 | 0 |

must be other problems where the control class gained relatively more. However, qualitative analyses of other tasks did not reveal any large, systematic gains in the control classroom like the distinct shift just described; students in the control class picked up a few more points here and there, but there were no big changes compared to the same teacher's class the previous year. Thus, we are inclined to attribute systematic shifts in the distribution observed in the participating classroom to changes in instruction.

Similar analyses were carried out for the best and next-best pairs of classrooms in the higher socioeconomic schools. In these pairings, however, the best classes in the control schools were the classes with the smallest decline on the Maryland mathematics assessment because all classrooms in these schools declined from 1992 to 1993. In contrast, the best classroom in the highest socioeconomic participating school showed a substantial improvement (ES = .53) and caught up to where the best control classrooms had been the year before.

Although the level of student performance was much higher in both the participating and control class-


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

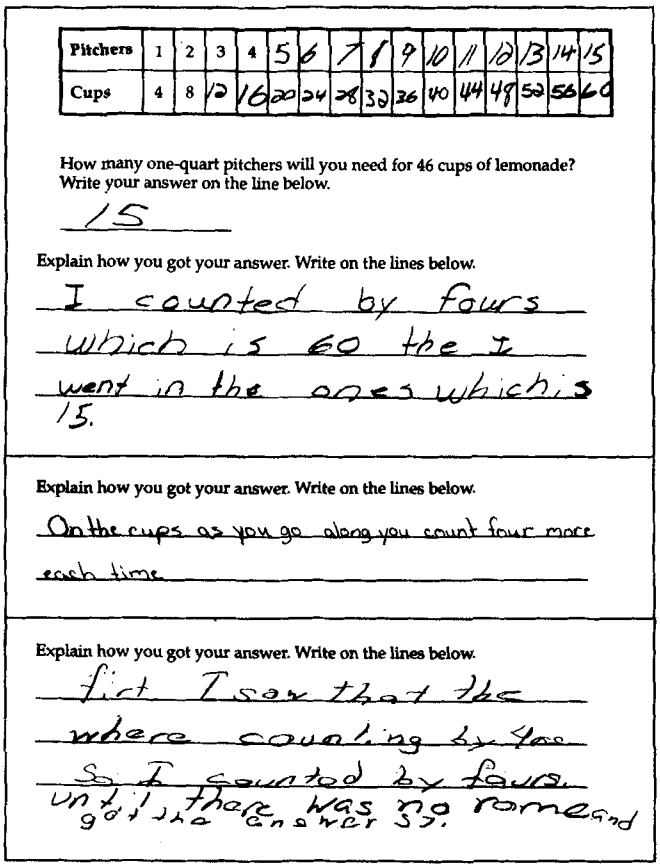
Explain how you got your answer. Write on the lines below.

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.

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

From pichers # 11 to 12 it went 44 48 cups so I just put 11+

FIGURE 1. *Sample student responses on Maryland mathematics assessment, Problem Set Two (Lemonade Step 4) illustrating qualitative Category I: Extends table, answers correctly, explains either pattern or point in chart*



| 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.

15

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

I counted by fours which is 60 the I went in the ones which is 15.

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

On the cups as you go along you count four more each time

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

fist I saw that the where coualing by 4s so I counted by fours. until I got the answer 57. there was no rome and

FIGURE 2. *Sample student responses on Maryland mathematics assessment, Problem Set Two (Lemonade Step 4) illustrating qualitative Category IV: Extends table, wrong answer (60, 15, 11, other), explanation describes pattern*

rooms in the higher socioeconomic schools, the best participating classrooms still showed specific improvements in student performance that could be associated with the project intervention (Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1995). For example, there were more right answers (Category I)—43% in 1993 versus 19% in 1992—in one top participating classroom. More importantly, however, in 1993, 77% of the children in this participating classroom wrote mathematically adequate explanations (Category I, III, or IV) about how they solved the problem. This proportion is in contrast to 31% who wrote explanations in the same teacher's classroom the year before. In the paired best control classroom, 95% wrote adequate explanations in 1992, but only 52% could do so in 1993. As explained previously, we are more inclined to attribute these declines to population changes rather than to a decline in the quality of teaching, especially because all classrooms in the control school were affected. Data from this same pair of classrooms, as in some other participating classrooms, also showed an increased ability for students to extend a mathematical pattern or complete a function table. In the baseline year, only 70% of the children in the participating teacher's classroom could extend the table (Categories I–V), but this percentage increased to 86% in 1993, making the participating classroom more comparable to the high levels achieved in the control classroom both years (95% and 86%, respectively).

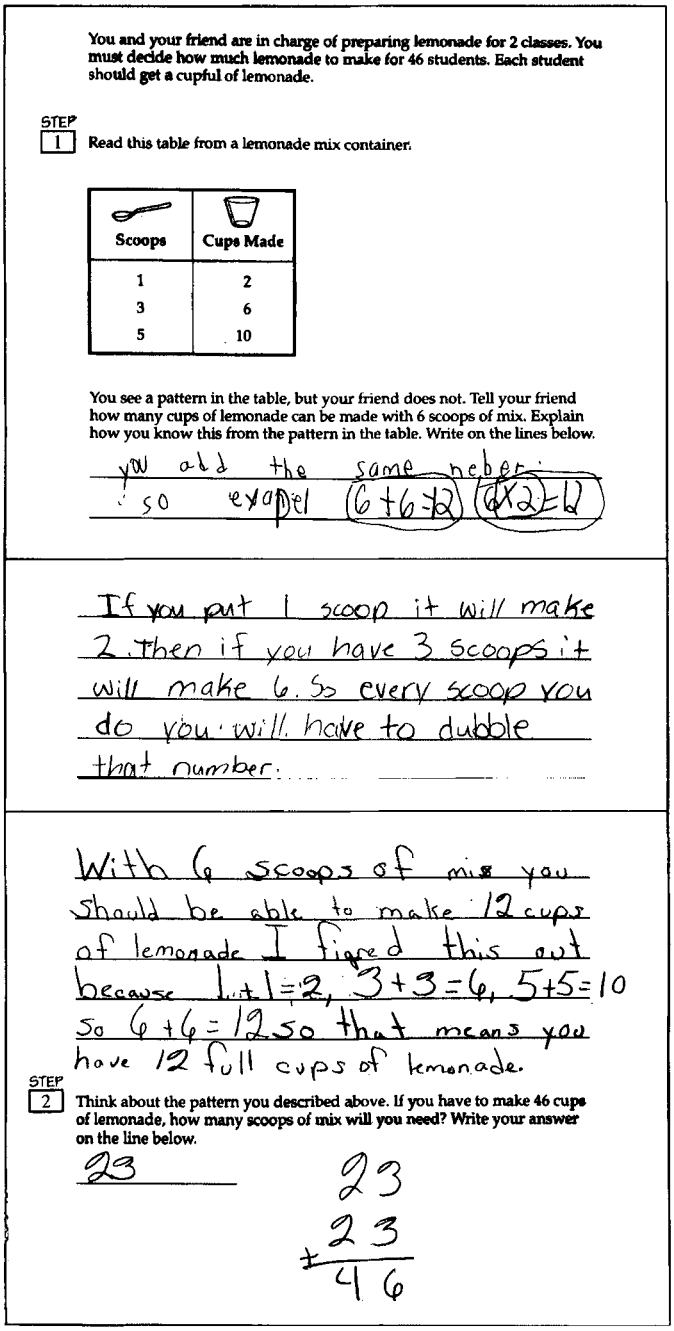
Samples of student responses to a different subpart of the lemonade task are presented in Figures 3 and 4. Again, we have chosen to illustrate the qualitative categories where students wrote explanations; these answers received either whole or partial credit in the quantitative scoring. This subpart was much more difficult for children across schools and did not show much of an improvement for the best low-socioeconomic classroom. There were no more right answers than in 1992, but 27% of students in the best low-socioeconomic classroom wrote mathematically adequate descriptions of the pattern (Category V, shown in Figure 4) compared to 0% in 1993. An improvement in the number of students writing explanations on this problem also occurred in the matched, low-SES classroom.

Category V responses show some of the richness of the students' answers and also help us to understand why many students found this problem more difficult. In every classroom, there were some students who could count by fours when they got to Step 4 but had trouble with Steps 1–2 because they extended the table downward without looking at the left–right correspondence. They were able to explain what they were thinking mathematically in a way, in fact, that revealed their misconception:

- Yes I do see a pattern, on the side with the spoon it counts by 2's were there's a cup it counts by fours.
- because on scoops it's go 1, 3, 5, I saw that their doing all odd so I put odd why cups was all even and 4 in the mitel. What I mean is 2 + 4 = 6 and 6 + 4 = 10 and so on.

The best high-SES participating classroom showed a substantial gain on this problem (the data for the comparison participating and control classroom are shown in Table 4). From 1992 to 1993, the percentage of students who wrote mathematical explanations (and extended the table) increased from 27% to 57% (Categories I, III, V). The corresponding change in the control classroom was a decrease from 45% to 35%.

The qualitative analyses of student answers on the Maryland mathematics assessment were not intended to refute or contradict quantitative findings of little or no difference. In fact, patterns suggested by the qualitative coding could be confirmed using quantitative scores. For example, the overall gain was paralleled by a gain in



You and your friend are in charge of preparing lemonade for 2 classes. You must decide how much lemonade to make for 46 students. Each student should get a cupful of lemonade.

STEP 1 Read this table from a lemonade mix container.

| Scoops | Cups Made |
|---|---|
| 1 | 2 |
| 3 | 6 |
| 5 | 10 |

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

you add the same nebet so exapel (6+6=12) (6x2=12)

If you put 1 scoop it will make 2. Then if you have 3 scoops it will make 6. So every scoop you do you will have to dubble that number.

With 6 scoops of mix you should be able to make 12 cups of lemonade I figred this out because 1+1=2, 3+3=6, 5+5=10 so 6+6=12 so that means you have 12 full cups of lemonade.

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23

23
+23
46

FIGURE 3. *Sample student responses on Maryland mathematics assessment, Problem Set One (Lemonade Steps 1–2) illustrating qualitative Category I: Right answers, explanation describes pattern (includes minimal explanation 6 + 6 = 12)*

points on the explanation portion of problems. Apparent gains at the lower end of the distribution were confirmed by significant shifts out of the lowest two quintiles (as defined in the baseline year) for two of the three participating schools. Most importantly, effect size calculations showed that about half of the participating classrooms gained a great deal (.25 to .50) while the other half of classes gained zero or lost ground consistent with the pattern in control schools. What the qualitative analyses helped to do is illustrate the substantive nature of improvement in student learning when it did occur. Significant shifts were observed on specific aspects of problems in participating classrooms, but not in control classrooms, and were associated with the kinds of mathematical activities introduced as part of the project. In many cases, this meant that students in the middle and bottom of the class were able to do things that their counterparts in participating classrooms had not been able to do the previous year.

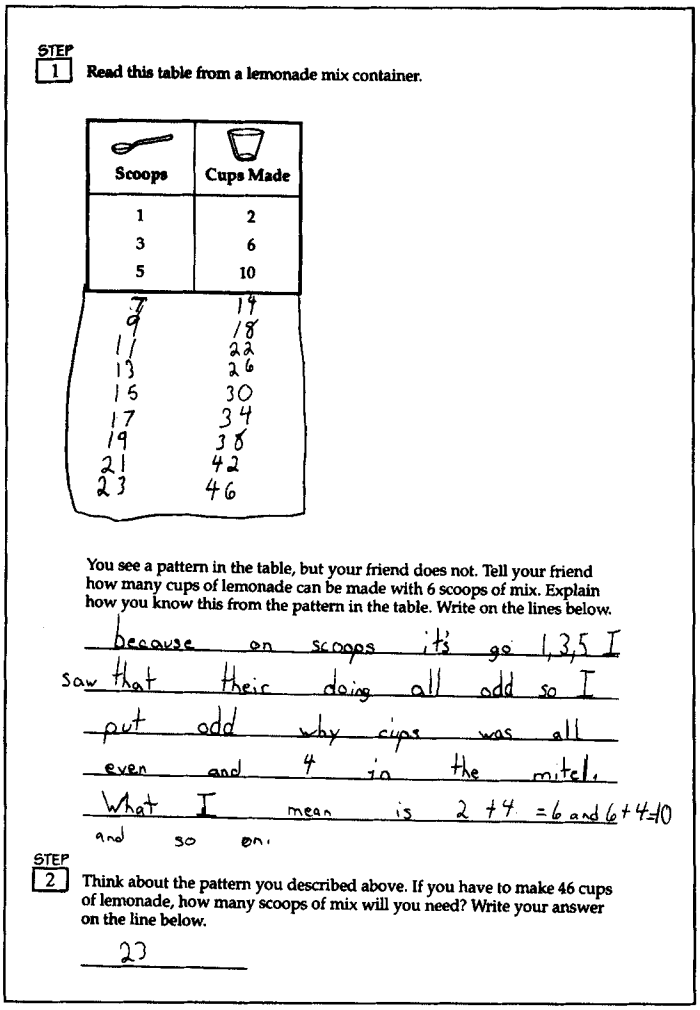
## Discussion

A fairly elaborate research design was implemented to evaluate the effect of a year-long performance assessment project on student learning. Maryland third-grade assessments in reading and mathematics and another alternative mathematics test served as independent measures of student achievement, separate from the classroom assessments developed as part of the project. 1993 end-of-year results were compared to baseline administrations of the same measures in 1992 and to control-school performance using analysis of covariance.

Results in reading showed no change or improvement attributable to the project. Third graders in the participating schools did about the same on the Maryland reading assessment as third graders had done the year before, and there were no significant differences between participating and control schools. In mathematics, there were also no gains on the alternative assessment measure. However, small and potentially important changes did occur on the Maryland mathematics assessment.

It is possible to offer both pessimistic and optimistic interpretations of the study results. Most significantly, from a negative perspective, it is clear that introducing performance measures did not produce immediate and automatic improvements in student learning. This finding should be sobering for advocates who look to changes in assessment as the primary lever for educational reform.

Of course, there were mitigating factors that help to explain and contextualize the lack of dramatic effects. First, we did not teach to the project outcome measures. For example, the classroom use of written summaries to assess meaning-making should have given students more



STEP 1 Read this table from a lemonade mix container.

| Scoops | Cups Made |
|---|---|
| 1 | 2 |
| 3 | 6 |
| 5 | 10 |
| 7 | 14 |
| 9 | 18 |
| 11 | 22 |
| 13 | 26 |
| 15 | 30 |
| 17 | 34 |
| 19 | 38 |
| 21 | 42 |
| 23 | 46 |

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

because on scoops it's go 1,3,5 I saw that their doing all odd so I put odd why cups was all even and 4 in the mitel. What I mean is 2 + 4 = 6 and 6 + 4 = 10 and so on.

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23

FIGURE 4. *Sample student response on Maryland mathematics assessment, Problem Set One (Lemonade Steps 1–2) illustrating qualitative Category V: Attempts to extend table but focuses on left or right column, not left: right pattern OR sees 1:2 pattern but can't apply to get answers*

experience with the open-ended format on the Maryland reading assessment. However, the Maryland reading tasks did not require skill at summarizing, and we did not introduce any other item formats from the outcome measure, such as comparative charts or story webs. We should also note that the level of text difficulty in the Maryland assessment was quite high. In retrospect, we might have included additional, easier texts to be more sensitive to gains by below-grade-level readers who were the focus of the running-record assessments.

Similarly, in mathematics, we worked on explanations, consistent with the NCTM standard that students be able to communicate mathematically, which should and did help students write about their thinking on various open-ended problems. In other respects, we did not use formats or problem types that conformed specifically to the Maryland assessment. For example, some classroom activities addressing patterns and functions were conceptually similar to the table-extension problems seen in the Maryland Lemonade tasks (What's my rule? 67, 56, 45, . . . from Dale Seymour; or, using pictures or models, if it takes 4 toothpicks to make one square and 12 toothpicks to make two squares, one around the other, how many toothpicks does it take altogether to make a third square around the others? Show the number of squares and the number of toothpicks in a table.) These pattern and function problems were quite varied and were only a small part of the dozens of problems provided addressing a wide variety of topics. It is reasonable to assume that teachers might have behaved differently and imitated the outcome measures more closely, if our 1992 baseline administration and anticipated 1993 measure had been imposed by an external agency for accountability purposes. Such practices could very likely have heightened the improvement of outcome scores, but then the question would arise as to whether the increased scores validly reflected improvement in students' understanding.

When we showed project teachers the outcome findings (in Fall 1993), they were disappointed but offered an explanation regarding the intervention that jibes with our own sense of the project's evolution. Despite the level of workshop effort throughout 1992–1993, by Christmas, project assignments still had not been assimilated into regular instruction. Although we have evidence of changes beginning to be made in the spring term (Flexer et al., 1994), many teachers said that they did not "really" change until the next year (1993–1994) (beyond the reach of the outcome measures). Several teachers argued that they did not fully understand and adopt project ideas and assessment strategies until they began planning and thinking about what and how to teach the next year. This view is consistent with the literature on teacher change. Fundamental and conceptual change occurs slowly. Furthermore, changes in student understandings must necessarily come last, after changes in teacher thinking and changes in instruction.

We also note that small gains in mathematics compared to zero gain in reading might have occurred because teachers had "further to go" in mathematics than in reading. If we take district curriculum frameworks as the standard, which are consistent with emerging professional standards in the respective disciplines, most teachers in the participating schools had already implemented some instructional strategies focused on meaning-making. In mathematics, the district frameworks were newer, and teachers were less familiar with them. Two teachers had tried out the Mar-

**Table 4**

*Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set One (Lemonade Step 1-2) From the Classrooms With the Greatest Gains in the High Socioeconomic Participating and Control Schools*

| | Participating | | Control | |
|---|---|---|---|---|
| Qualitative categories | 1992 | 1993 | 1992 | 1993 |
| I. Right answers, explanation describes pattern (includes minimal explanation, 6 + 6 = 12) | 19% | 24% | 39% | 9% |
| II. Right answers, no explanation (but may show 23 + 23 = 46) | 0 | 0 | 6% | 9% |
| III. Gets 12 cups with adequate explanation but cannot extend to 46 cups | 8% | 9% | 0 | 0 |
| IV. Gets 12 cups, inadequate explanation, (wrong or no extension) | 4% | 0 | 0 | 4% |
| V. Attempts to extend table but focuses on L or R column, not L/R pattern, OR 1:2 correspondence without answers, explains thinking | 0 | 24% | 6% | 26% |
| VI. Wrong answers, explanation not based on chart or only restates answer | 58% | 9% | 33% | 48% |
| VII. Wrong answers, no explanation | 4% | 29% | 0 | 0 |
| VIII. Blank | 8% | 5% | 17% | 4% |

ilyn Burns (1991) multiplication unit the year before, but several more teachers decided to try it during the project year. Several were using manipulatives for the first time; several adopted materials to teach problem-solving strategies for the first time, and one group of teachers worked to develop new units in geometry and probability. Even when teachers did not understand them well or use materials optimally, these brand-new activities represented substantial shifts in the delivered curriculum.

In contrast to these apologies and caveats about why change did not occur, the cause for optimism comes from the small but real gains in mathematics. Because of the project, most of the teachers in the participating schools spent class time on written explanations (especially what makes a good explanation) and on mathematical patterns and tables, which they had never done before. As a consequence, there were specific things that a large proportion of third graders in these classrooms could do on the outcome assessments, where before only the most able third graders had been able to intuit how to do them.

Our concluding advice is that reformers take seriously the need for sustained professional development to implement a thinking curriculum. Performance assessments—even with the diligent effort of most project teachers and the commitment of four university researchers—did not automatically improve student learning. When positive changes did occur, however, they supported our beliefs that less able students can develop conceptual understandings presently exhibited by only the most able students—if only they are exposed to relevant problems and given the opportunity to learn. Performance assessments that embody important instructional goals are one way to invite instructional change, and assessments have the added advantage of providing valuable feedback about student learning. However, we would not claim that performance assessments are necessarily the most effective means to redirect instruction. When teachers' beliefs and classroom practices diverge from new conceptions of instruction, it may be more effective to provide staff development to address those beliefs and practices directly. Performance assessments are a key element in instructional reform, but they are not by themselves an easy cure-all.

## Note

Work reported herein was supported by the Office of Educational Research and Improvement (OERI), U. S. Department of Education, through the Center for Research on Evaluation, Standards, and Student Testing (CRESST). The findings and opinions expressed do not reflect the position or policies of OERI or the U. S. Department of Education. We thank the Maryland Department of Education for allowing us to use tasks from the Maryland School Performance Assessment Program as outcome measures for the study. We also thank the Riverside Publishing Company for permission to use portions of the second grade Iowa Test of Basic Skills as a premeasure.

## References

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the Commission on Reading.* Champaign, IL: National Academy of Edu-

cation and National Institute of Education, Center for the Study of Reading.
Borko, H., Davinroy, K. H., Flory, M. D., & Hiebert, E. 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: LEA.
Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (in press). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education.*
Burns, M. (1991). *Math by all means: Multiplication grade 3.* Sausalito, CA: The Math Solution Publications.
Cohen, S. A. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher, 16*(8), 16–20.
Davinroy, K. H., & Hiebert, E. H. (1993, December). *An examination of teachers' thinking about assessment of expository text.* Paper presented at the annual meeting of the National Research Conference, Charleston, SC.
Flexer, R. J., Cumbo, K., Borko, H., Marion, S., & Mayfield, V. (1994, April). *How "messing about" with assessment affects instruction.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.
Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.
Koczor, M. L. (1984). *Effects of varying degrees of instructional alignment in posttreatment tests on mastery learning tasks of fourth grade children.* Unpublished doctoral dissertation, University of San Francisco.
Koretz, D. M., Linn, R. L., Dunbar, S. B. & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Research, 20*(8), 15–21.
Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "Everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5–14.
Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4–12: Executive summary.* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.
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.
Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 72,* 232–238.
Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1995). *Effects of introducing classroom performance assessments on student learning* (CRESST Report). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.
Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1990). *The role of testing in elementary schools.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.
Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70,* 703–713.

# CALL FOR AWARD NOMINATIONS

## Nominations for the NCME Award for Career Contributions to Educational Measurement

NCME members are encouraged to nominate individuals for the NCME Career Award, which honors living persons whose contributions over a career have had a widespread positive impact on the field of educational measurement. These influential contributions may include theoretical or technical developments; conceptualizations of information on educational measurement that has enhanced public understanding; and applications of theory through procedures, instruments, or programs that have had broad influence on the nature and practice of educational measurement.

Recipients of the award since its inception in 1988 are: Melvin Novick, Ralph Tyler, Frederic Lord, Albert Hieronymus, T. Anne Cleary, Ronald Hambleton, Leonard Feldt, Robert Linn, and Jason Millman.

Nominations should consist of a 1- or 2-page written statement describing the nature, significance, and breadth of impact of the nominee's work. Mail nominations to David A. Frisbie, 316 Lindquist Center, University of Iowa, Iowa City, IA 52242. Inquiries may be made by E-mail (dfrisbie@uiowa.edu) or phone (319-335-5410).

The deadline for receipt of written nominations for the 1997 award is January 13, 1997. The award recipient will be recognized at the Annual Meeting in Chicago.

## NCME Award for Technical Contribution to Educational Measurement

This year (1996–1997) the National Council on Measurement in Education (NCME) will make its fifth award for outstanding *technical* or *scientific* contribution to the field of educational measurement.

Examples of technical contributions include, but are not limited to, innovative ways of solving practical and theoretical measurement problems, inventive instrument development techniques, creative testing procedures or products, and scientific contributions to measurement research methodology. Selection criteria are quality, inventiveness, and positive impact of the technology on the field of educational measurement. The past award recipients are Kikumi K. Tatsuoka, University of Illinois (1985); Robert Mislevy, Albert Beaton, Eugene Johnson, and Kathleen M. Sheehan, Educational Testing Service (1988); Fumiko Samejima, University of Tennessee (1991); and William F. Stout, University of Illinois (1994).

To be eligible for this award, the technical contribution must have occurred initially during 1994, 1995, or 1996. The work must have appeared in a research publication, but not necessarily an NCME publication. One may nominate his or her own technical contribution or, with permission, someone else's. A nomination should consist of five copies of a 3- to 5-page statement describing the technology, application area, and products or results of the effort. Finalists may be requested to submit additional information.

Mail nominations to: Cynthia B. Schmeiser, Awards Committee Chair, ACT, 2201 North Dodge Street, Iowa City, IA 52243-0168. Nominations for the award must be received by January 10, 1997. The award will be presented at the 1997 NCME Annual Meeting in Chicago.