Instructional Effects in Elementary Schools

CSE Technical Report 577

Joan L. Herman
CRESST / University of California, Los Angeles

September 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.5 Technical and Functional Quality of Assessment Systems: Validity, Equity, and Utility
Robert L. Linn, Project Director, CRESST/University of Colorado at Boulder

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this publication do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

INSTRUCTIONAL EFFECTS IN ELEMENTARY SCHOOLS[1]

Joan L. Herman
CRESST/University of California, Los Angeles

Abstract

Standards-based reform represents not only high expectations for student performance, but also equally high expectations for how assessment-based accountability policies can influence teaching and learning in schools. Much is expected of standards-based assessment at the policy level: Such assessments are expected to serve as both a lever for improvement and a measure of such improvement. Based on available research, this report explores how well assessment serves these functions from the perspective of elementary schools. The report begins with the basic vision of what standards-based assessment is expected to accomplish and then reviews major themes emerging from the literature that show the extent to which this vision is being realized. The report concludes with recommendations for improving policy and practice.

A Vision of Standards-Based Assessment Reform

In brief, the basic vision of standards-based assessment starts with consensus on what is important for *all* students to know and be able to do if they are to be successful in the 21st century. The idea is that if society and its stakeholders are clear on what is expected, it is possible to hold everyone in the system—from policymakers to educators and students—accountable for meeting those expectations. What is particularly new in standards-based assessment reform is being clear not only on the "what" of what is expected (the content standards), but also on "how well" it should be accomplished (the performance standards) (Linn & Herman, 1997).

The multiple functions of assessment. The performance standards really come to life as large-scale assessments are developed and put into place. Emanating from the state and/or local level, the assessments make explicit what kinds of learning are expected and, as performance levels and minimum passing scores are established, make clear how well students have to do to meet the standard. The assessments thus become a primary vehicle for communicating what the standards really mean and provide a strong signal to teachers and schools about what they should be teaching and what

[1] A revised version of this paper will appear as a chapter in *Redesigning Accountability Systems*, edited by Richard Elmore and Susan Fuhrman, Teachers College Press.

students should be learning. Unique to standards-based assessment as well is the intention not only to signal to teachers what to teach but also, with the use of multiple types and forms of assessment, to provide clues on how to teach. That is, with the incorporation of more performance-based and open-ended items, assessments are also expected to communicate models of good teaching and learning practice.

The results from these assessments are supposed to provide information of value to schools and policymakers by measuring the status and progress of student learning. The results are intended to support important insights on the nature, strengths, and weaknesses of student progress relative to the standards, and educators are expected to use this feedback to understand and to direct their efforts toward improving relevant aspects of student learning.

Policymakers try to strengthen the accountability aspects of the system by establishing specific goals for school performance and attaching incentives and sanctions to achieving, or not achieving, or surpassing these results. Across the country, and spurred at least in part by federal policy, states have created sizeable incentives for performance—substantial cash awards for schools and teachers who meet or exceed their goals; and at the other extreme, schools that don't make the grade are threatened with takeover. Dramatic incentives for students also have been added to the mix, as a growing number of states adopt policies that require students to meet a performance standard to be promoted to the next grade or to be granted a high school diploma. Through such rewards and sanctions, policymakers seek to motivate teachers, students, and the community to pay attention—to the standards, to the assessment results, and to the analysis of results to improve subsequent performance. The system thus promotes a continuous improvement model aimed at enabling all children to reach the standards: Establish and monitor goals and benchmarks, assess progress, use results on goal attainment to improve performance.

Essential alignment with standards. The idea is not really to teach to the test, but to motivate everyone in the system to focus on the standards and enable children to reach them (see Figure 1). Reaching the goal requires the broad alignment of system components and the specific alignment of the assessment with the standards, but more importantly, and of special importance to the content of this report, reaching the goal requires the alignment of classroom instruction with the standards and their assessments. It is only when the content and process of teaching and learning correspond to the standards that students indeed have the opportunity to learn what they need to be successful. Under these conditions, too, an assessment provides

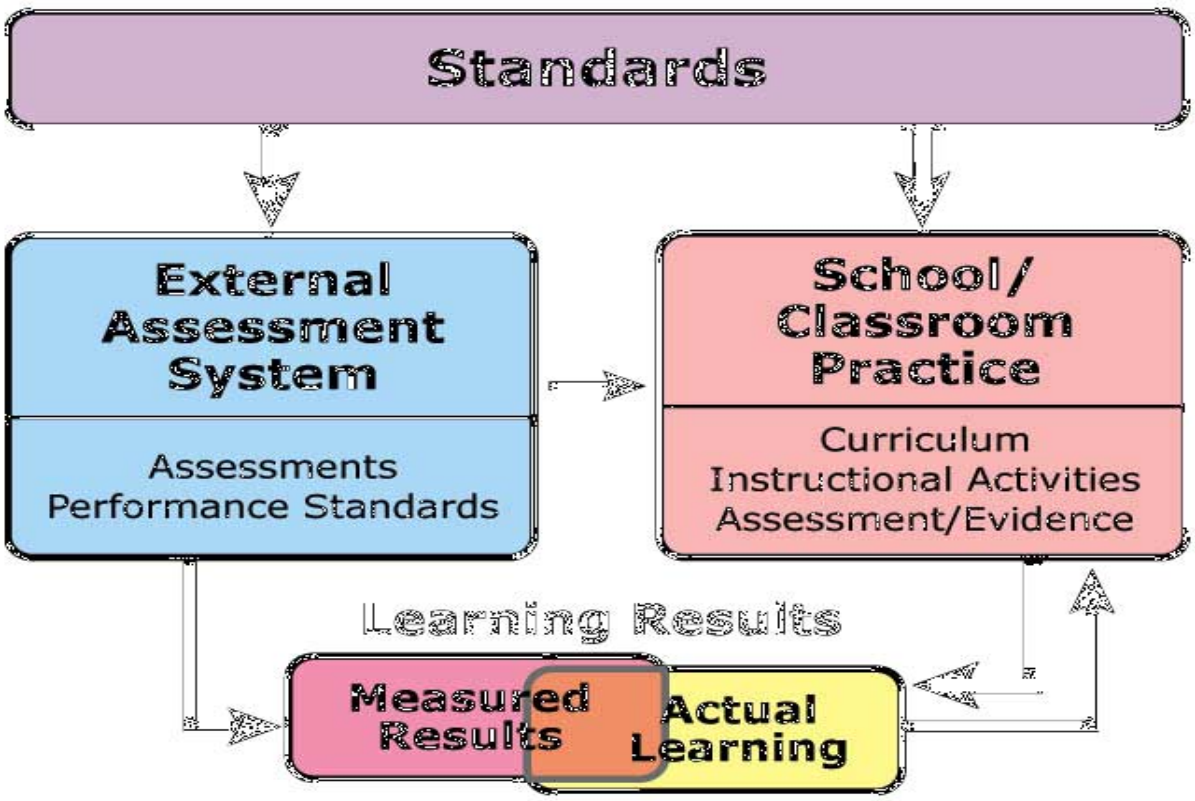

Figure 1. A model of standards-based assessment reform.

information on how well students are doing relative to the standards and on the extent to which classroom teaching and learning are helping students to attain the standards. All parts of the system are focusing on the same or a similar conception of standards and are in sync with a continuous improvement model.

Without such a correspondence, the logic of the standards-based system falls apart. The inferences that can be drawn from assessment results about how well schools are doing and what progress is being made also become tenuous. For example, if there is little alignment between what is being taught and what is being tested, the value of using results to determine the strengths and weaknesses or overall effectiveness of teaching and instruction is significantly undermined. That is, if what is tested is not taught, the information can tell us little, if anything, about what students learned in school, because that which they might have learned was not assessed.

Similarly, if the assessment and the standards are not aligned, the results can provide little information about whether students are attaining specified standards or

whether instruction is helping them to make the grade. Worse yet, rather than being mutually reinforcing, the standards and the assessment may push teachers and schools in different directions. With incentives or sanctions attached to performance results, there is little doubt about which direction teachers and schools are most likely to heed.

Of course, even under the best of circumstances, a test measures only a part of what students are learning—that which can be measured in a finite and limited period of time and by the types of formats that are included in the test. All measures also are fallible and include error; they thus provide only an imperfect measure of student performance. Well recognizing the limits of the information that can be derived from any single measure, measurement experts advise that good assessment systems really need to include multiple measures to assess the range of knowledge and skills we really want children to achieve.

The Research Base

Interestingly, the current vision of standards-based assessment reform and the high hopes it holds for large-scale, standards-based assessment has its roots in research conducted during the late 1970s and 1980s showing the unfortunate effects of traditional, standardized tests. The research showed the power of these tests, built to assess general achievement and based solely in multiple-choice items, to influence teachers and schools.

Pre-reform literature. For example, a number of researchers, using surveys of teachers, interview studies, and extended case studies, provided evidence that traditional, standardized tests were having adverse effects on the quality of curriculum and classroom learning. Under pressure to help students do well on such tests, teachers and administrators tended to focus their efforts on test content, to mimic the tests' multiple-choice formats in classroom curriculum, and to devote more and more time to preparing students to do well on the tests (Corbett & Wilson, 1991; Dorr-Bremme & Herman, 1986; Kellaghan & Madaus, 1991). The net effect was a narrowing of the curriculum to the basic skills assessed and a neglect of complex thinking skills and subject areas that were not assessed.

Furthermore, the research suggested that schools and teachers used the test format as a model for curriculum and instruction. Preparing students for the test meant lots of practice with test-like, multiple-choice items, with more and more of the curriculum given over to test preparation as the pressure to do well increased. To many, testing was encouraging "drill and kill" worksheets and outmoded, behaviorist pedagogy.

Such pedagogy viewed students as black boxes to be filled with discrete bits of knowledge, learning as a linear progression of discrete skills from rote to complex, and connections to students' existing knowledge and experience as unimportant (Resnick & Resnick, 1992; Shepard, 1991). There also was concern that an overreliance on testing gave short shrift to content areas such as science, social studies, and the arts, which often were not the subject of testing (Darling-Hammond & Wise, 1985; Shepard, 1991). Herman and Golan (1993), among others, noted that such narrowing was likely to be greatest in schools serving at-risk and disadvantaged students, because test scores in these schools were typically very low, and educators in these schools were likely to be under great pressure to improve their scores.

Effects on instruction, however, appeared very different when tests or other assessments used more performance-oriented items, rather than multiple-choice formats. Direct writing assessment—asking students to actually compose an essay rather than answer multiple-choice questions about the quality or grammar of a given piece—was a first example. Large-scale writing assessment had begun to gain popularity in the late 1970s with its inclusion in NAEP; then gradually throughout the 1980s, more and more states and locales moved to include this type of assessment in their programs. At the time, arguments for this mode of testing were based primarily on evidence of validity—evidence suggesting that multiple-choice tests did not provide accurate measures of students' ability to write (Quellmalz & Burry, 1983). However, as experience with these direct measures grew, their potential for influencing teaching and learning became more apparent. Studies of the effects of California's eighth-grade writing assessment program, for example, indicated that the program encouraged teachers both to require more writing assignments of students and to give students experience in producing a wider variety of genres. Beyond impact on instruction, furthermore, studies showed that student performance in some states and districts improved over time with the institution of the new assessment programs (Chapman, 1991; Quellmalz & Burry, 1983).

Post-reform studies. Armed with the research, educational reformers aimed to use the power of assessment intentionally to achieve their goals, first in promoting the use of performance assessment in large-scale assessment during the 1990s and more recently moving to the adoption of standards-based assessment systems. Coincident with these reforms have been a number of studies of their implementation and impact. These studies cross states and locales and represent significant variation in terms of the nature of tests used, the strength of incentives and sanctions, and research

methodology. For example, at the state level, there have been studies of the effects of systems in Arizona (Smith & Rottenberg, 1991), California (Herman & Klein, 1996; McDonnell & Choisser, 1997;), Kentucky (Borko & Elliott, 1998; Koretz, Barron, Mitchell, & Stecher, 1996; Stecher, Barron, Kaganoff, & Goodwin, 1998; Wolf & McIver, 1999), Maine (Firestone, Mayrowetz, & Fairman, 1998), Maryland (Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000; Goldberg & Rosewell, 2000; Lane, Stone, Parke, Hansen, & Cerrillo, 2000), New Jersey (Firestone et al., 2000), North Carolina (McDonnell & Choisser, 1997), Vermont (Koretz, McCaffrey, Klein, Bell, & Stecher, 1993), and Washington (Borko & Stecher, 2001; Stecher, Barron, Chun, & Ross, 2000).

Major Themes in Recent Research

Echoing themes from earlier studies, findings from these post-reform studies provide a surprisingly consistent picture of how these new assessment systems are working and the extent to which they are working as intended, in the sense of encouraging good teaching and learning and promoting progress toward students achieving the standards.

Teachers Listen to the Signal

Results from nearly every study indeed indicate that teachers pay attention to what is tested and adapt their curriculum and teaching accordingly. For example, Lane et al. (2000), in a survey of a representative sample of Maryland elementary and middle schools (*n = 90*), found teachers and principals reporting that the Maryland State Performance Assessment Program (MSPAP) was having substantial impact on curriculum and instruction in reading and mathematics. The researchers' composite index of MSPAP impact, including teachers' responses to the overall influence of MSPAP on classroom activities, its influence on subject area instruction and assessment, and teachers' use of MSPAP-type problems, showed at least moderate impact (means of 2.8 to 3.3 out of a possible 4) across the two subject areas and school levels.

A recent statewide study of the education reform in Washington State similarly showed the seriousness with which educators respond to testing. One hundred percent of the surveyed principals reported that they had developed schoolwide plans for improving performance on the Washington Assessment of Student Learning (WASL) and implemented test preparation activities (Stecher et al., 2000). Moreover, nearly three quarters of the principals indicated that they had instituted schoolwide policies to address curriculum gaps revealed by the test. Moving to the classroom level, Stecher et al. found nearly two thirds of surveyed elementary school teachers reporting that the

WASL had had a moderate or great effect on their teaching of writing and three quarters reporting a moderate or great effect on their teaching of mathematics. These findings mirror earlier studies in Kentucky that found principals strongly encouraging teachers to focus their instruction on the content and skills likely to be on the Kentucky Instructional Results Information System (KIRIS) and teachers reporting an increase in the match between the content of their instruction and that of the assessment (Koretz, Barron, et al., 1996).

Teachers Model Test Content and Pedagogy

Research shows furthermore that in addition to modifying their classroom curriculum and instruction to include the content of what is tested, teachers tend to model the pedagogical approach represented by the test. Thus, when a large-scale assessment is composed of multiple-choice tests, teachers tend to use multiple-choice worksheets in their practice, but when the assessments use open-ended items and/or extended writing and rubrics to judge the quality of student work, teachers incorporate these same types of activities in the classroom work. Dan Koretz's early study of Vermont's statewide portfolio assessment, for instance, found more than 80% of elementary school teachers reporting a moderate or large increase in the amount of class time they devoted to teaching problem solving due to the assessment (Koretz, Stecher, & Deibert, 1992). Similarly, because the assessment also stressed communication, more than two thirds of the teachers reported having their students spend somewhat more or much more time than in previous years writing reports about mathematics, and more than 60% assigned mathematics applications, which were required by the portfolios, at least weekly. Subsequent studies in Kentucky similarly found teachers reporting that that state's innovative assessment system stimulated teachers to focus more on tested subjects and to increase their use of instructional practices intended by the test reformers (Stecher et al., 1998).

Findings from Maine and Maryland echo these trends. Firestone and colleagues (Firestone et al., 1998) found teachers *adding to* their curriculum the types of problem-solving tasks the teachers expected to be on the statewide assessment. In the case of Maryland, these were extended projects that asked students to apply mathematics concepts, reason mathematically, and use multiple forms of representation.

Test Preparation Merges Into Instruction

The match between test format and instructional format is most apparent in direct test preparation activities. Here the intent is to engage students in practice activities

explicitly designed to mirror the given assessments as closely as possible, with the explicit purpose of getting students familiar with the test format and enabling them to do better on the test. Such practice activities are typically derived from sample items and practice materials provided by the state or district and from commercially available materials developed by test publishers.

The extent and nature of such test preparation vary considerably from study to study. Mary Lee Smith's case study of Arizona elementary schools found regular curriculum virtually shutting down in some schools for several weeks prior to the mandated standardized test period, as teachers directly prepared their students for the coming test (Smith, Edelsky, Draper, Rottenberg, & Cherland, 1990). Smith and colleagues viewed this as an obvious interruption and detraction from regular instruction.

Similar at the extreme, but different in process, Herman and colleagues' study of California's then eighth-grade mathematics assessment found that virtually all surveyed teachers reported using sample items with their students (Herman, Klein, Heath, & Wakai, 1995). The assessment emphasized complex thinking and problem solving, and the sample items were open-ended, requiring extended time. On average, teachers spent three to five class periods on these practice items, but notably, about one third of the respondents reported spending nine or more class periods, the equivalent of nearly 2 weeks, in such practice. Anecdotal evidence suggested that in some classrooms these practice items were amassed near testing time, but in other cases, they were distributed throughout the school year.

More recently, Stecher et al.'s (2000) study of Washington explicitly documented how time spent in test preparation may vary with the time of the year. That study found that teachers increased the amount of time they spent in direct preparation for the WASL as the test approached in the spring. Near the beginning of the year, in November, about one half of the teachers reported spending 1 to 2 hours a week preparing for the WASL, and about a quarter reported spending no time at all in test preparation. Not surprisingly, however, the picture changed as the testing dates approached. Near testing time in April, one third of fourth-grade teachers and one fifth of seventh-grade teachers reported spending more than 4 hours per week preparing for the test, and less than 10% reported spending no time on test preparation (see Figure 2). The results were similar for writing teachers in Washington.

Firestone et al. (2000) also found a similar pattern of increased attention to test preparation just prior to testing in New Jersey and noted sizeable socioeconomic

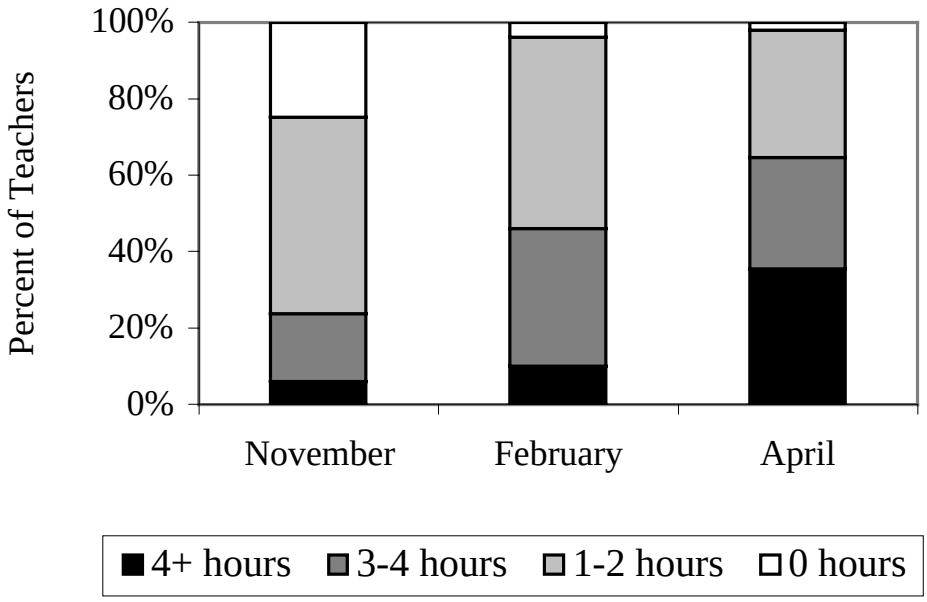

Figure 2: Time devoted to test preparation: Fourth-grade mathematics (Stecher et al., 2000).

differences in such practices as well. Teachers from schools in high-poverty districts reported substantially more time devoted explicitly to test preparation activities than those in wealthy districts. The estimates of teachers from middle-class districts fell between these two.

Of course, it can be difficult to differentiate between special test preparation efforts and "regular" curriculum and instruction activity that has been influenced by the standards and assessments mandated by external authorities. Part of the issue is one of intent: The former is enacted especially to increase test performance; its value in real learning is not a primary issue. The latter is ongoing curriculum activity that is influenced in content and format by important assessments but is intended to promote student learning.

Sometimes an activity may start with one intent and merge into the other. For example, Herman et al.'s (1995) study of eighth-grade mathematics found a number of teachers instituting a "[state assessment] problem of the week," which initially was intended to prepare students for the test. However, some teachers reported anecdotally that, over time, the attention to problem solving that this practice represented became more integrated with regular instruction and part of teachers' routine repertoire.

Many states' and locales' experiences incorporating state-assessment rubrics into their instruction tell a similar story. For example, Hilda Borko and colleagues' study

(Stecher & Borko, 2002) of exemplary sites in Washington revealed teachers consciously using WASL rubrics for writing and mathematics to prepare their students for the test. Borko described the example of Ms. Alexander [pseudonym], who asked her students repeatedly over the course of the year to write to the sample prompts provided by the state, scored their pieces using the state rubrics, and engaged students in discussion about what skills they used, where they might have gotten stuck, and what strategies might help them to do better. In a mathematics example from this study, another teacher commented, "I have conversations with them [her students] about what the scoring on the WASL is going to look like. Early in the year we talk about what it looks like to score at the level of a one or a two. And then we talk about what to do to raise that score. If we use threes and fours as examples, we get together in pairs or as a whole class and talk about their justification for that level" (pp. 25-26).

Tests Draw More Attention Than Standards

The time many teachers acknowledge spending in test preparation makes obvious that the test, rather than the standards, may become the primary target in teachers' curricular plans, at least at some times during the year. That the test rather than the standards may get primary attention *throughout* the year is a point Brian Stecher and colleagues make forcefully using data from their Washington State study (Stecher et al., 2000). For example, when principals were asked about the alignment of their school's curriculum with the state standards in various subjects, the answers were very different depending on whether or not the subject is tested. Uniformly, more than 90% of the principals reported alignment for tested subjects of reading, mathematics and writing, but dramatically fewer reported alignment for other subjects. Similarly, there were large discrepancies in teachers' responses when asked about their understanding of the Washington Assessment of Student Learning (WASL), the Essential Academic Learning Requirements (EALR), and aligning curriculum and instruction with EALRs. Ninety four percent of the fourth-grade teachers were confident of their understanding of the WASL, but only about three quarters were similarly confident about the alignment of curriculum and instruction with the EALRs. Finally, survey responses found two thirds of teachers identifying their teaching as more like "teach(ing) to the WASL," than "teach(ing)" to the EALRs" (Stecher & Borko, 2002, p. 21).

That teachers may pay more attention to the tests than to the standards and/or curriculum frameworks that underlie them also is evident in teacher reports on their use of instructional time. In their study of Kentucky, Stecher and Barron (1999) examined how teachers allocated classroom time as a function of what was tested on

the now defunct Kentucky Instructional Results Information System (KIRIS) at their grade level.[2] Figure 3, taken from their work, shows that the amount of time teachers engaged their students in a subject each week seemed to be highly related to whether the subject was tested at their grade level. Teachers shifted their use of curriculum time from one grade to the next: Fourth-grade students on average spent 16.2 hours a week engaged in reading, writing, and science, the subjects on which they were tested by KIRIS, as compared to the 12.2 hours fifth-grade students spent on the same subjects. In contrast, fifth-grade students on average were involved for 16.8 hours a week in mathematics, social studies, arts and humanities, and practical living/vocational education, the subjects in which they were assessed, compared to 11.3 hours a week for fourth graders. Combined across subjects, this indeed represented a sizeable shift in curricular time. Asked why they reallocated their use of time, teachers, in responses to open-ended items, stated that KIRIS was the reason.

Similarly, when Stecher et al. (2000), looked within subjects to see what teachers were teaching relative to what was tested, they found different patterns by grade level. Thus, although standards are supposed to be continuous across grade levels, teachers tended to involve their students in more extended writing and address a greater number of writing objectives in tested grades than in the grades that were not part of the writing portfolio assessment. There were similar findings in mathematics, where teachers tended regularly to teach a greater number of mathematics topics when their grade was assessed in mathematics.

However, Borko's (Borko & Stecher, 2001) case studies of *exemplary* sites in the same state suggest that the picture of test-focused curriculum may not be as stark as Stecher et al.'s (2000) findings suggest. At these sites, principals and teachers certainly paid close attention to test results, analyzed them class by class, and used them to help identify curriculum strengths and weaknesses, but the analysis was a point of departure for reflecting on practices and identifying concrete ways to improve instruction. As one principal commented, "[WASL scores] raised our awareness level in terms of where we need to put our energies," but did not dictate the what and how of instruction (Stecher & Borko, 2002, p. 24).

[2] The Kentucky Instructional Results Information System tested seven school subjects, each tested at one grade level each in elementary, middle, and high school. At the elementary school level, three subjects were tested in the fourth grade; the remaining four were tested in the fifth grade. At the middle school level, subjects were split between the seventh and eighth grades.

	Grade	
	Fourth	Fifth
Subjects Tested in **Fourth** Grade		
Reading	**5.2**	4.7
Writing**	**5.8**	4.0
Science**	**5.2**	3.5
Subjects Tested in **Fifth** Grade		
Mathematics**	4.9	**6.4**
Social Studies**	3.5	**5.6**
Arts & Humanities**	1.5	**2.4**
Practical Living / Voc. Ed.**	1.4	**2.4**

Mean hours per week of instruction

***Significant at α = .05; **significant at α = .01.**

Figure 3. Mean hours per week allocated to instruction in various topics by subjects tested at each grade level (Stecher & Barron, 1999).

Nontested Content Gets Short Shrift

A focus on the test rather than the standards also means that what gets tested gets taught, and what does not get tested may get less attention or may not get taught at all. WYTIWYG—what you test is what you get—is a continuing truism in the world of standards-based assessment. Again, the Stecher et al. (2000) survey data from their Washington study provide a strong case. In Figure 4, we see teachers' reports of how, and how much, their time allocations to various subjects changed from prior years. Note that reading, writing, communication, and mathematics were subjects then tested on WASL, whereas the other subjects were not included on the test. The pattern is clear: Teachers increased the time they spent on tested subjects at the expense of nontested subjects. Moreover, teachers attributed the cause of these changes to WASL. Again, this mirrors earlier findings from Kentucky, where the great majority of teachers agreed that because of KIRIS, they were de-emphasizing or neglecting content that was not on the test (Koretz, Barron, et al., 1996).

The findings thus suggest that teachers and schools may focus overly on what is tested to the neglect of both the broader domain of the tested discipline and important subjects that are not tested. To the extent that a state or district test represents a well-balanced picture of its standards, this focus on the test may represent little problem.

However, the reality is that there are limits to how much time can be spent testing, and there are limits to the kinds of academic and intellectual capacities that can be well, efficiently, and accurately assessed with the most commonly used test formats. Recent reports about the nature of current state assessment programs, for example, show a retreat toward more traditional types of tests. The rich performance assessment experiments of the 1990s seem to have devolved, at best, into some attention on state assessments to limited open-ended, short-answer items. Multiple-choice items continue to predominate. *Quality Counts 2001*, for example, shows only eight states including extended response items outside of English / Writing (see chart "Measurement of Student Performance," Orlofsky & Olson, 2001).

Furthermore, the alignment between states' standards and what actually is tested continues to be problematic. Despite test developers' assurances that their tests matchspecified standards, relatively few states have undergone serious alignment review. The Achieve studies of nine states' systems represent an exception, but these

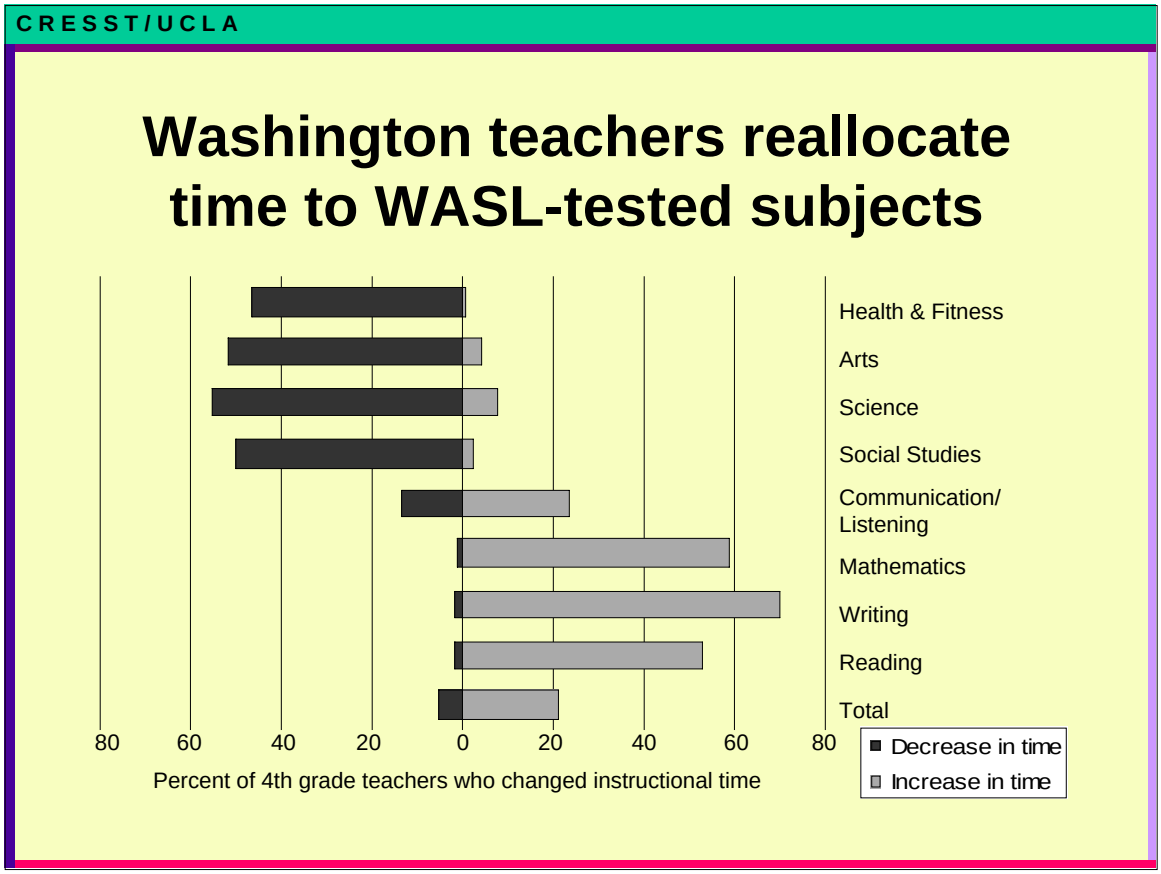

Figure 4: Teachers' allocation of time to WASL-tested and nontested subjects (from Stecher et al., 2000).

show uneven results (Rothman, Slattery, Vranek, & Resnick, 2000). Even at the simplest level of alignment where results tend to be the strongest—the extent to which items on an assessment could be matched to a state standard—results were variable. For example, though 95% of the items on the English tests in one state matched content and skills found in their standards, only 65% of the mathematics items were so aligned. Furthermore, the tests tended mostly to measure the lowest level objectives and not the depth of complexity articulated by the standards. Nor did the Achieve studies tend to find that the tests were balanced in their representation of each state standard. So teaching to the test does not necessarily mean teaching to the standards, and with increasingly great incentives and punishments attached to test performance, there is little doubt about whether the standards or the tests are the greater focal point.

Are the Learning Gains Real?

The potential narrowing of the curriculum to focus on only what is tested also has implications for validity of the assessment results and the credibility of gains that almost always appear in the first several years of a new state assessment program. We care about students' performance on a test, after all, because we believe that it represents something more than performance on the specific items and content covered by the test. It is not just that a student got these particular items correct, but rather that the score *generalizes* to some large domain of knowledge or skill and tells us something important about what students know and can do—in the current context, the content and performance standards that have been established. We want to infer how well students have achieved the standards from their performance on the particular sample of items included on the test.

However, if teaching and learning focuses, in the extreme, on only what is tested and the formats in which it is tested, then the test ceases to be a sample of performance. The test becomes the domain, and the generalizability of the results—and what meaning can be drawn from students' test performance, other than that they scored at a certain level on this particular set of items—becomes suspect.

This raises the question of whether the gains shown on state assessments represent real improvement in learning, or reflect narrow test preparation activities that do not generalize beyond the test itself and inflate actual improvement. Dan Koretz posited that if improvements in learning are genuine and meaningful, one should expect the increases in performance on the high-profile state assessment to show up on other similar measures of student achievement. Using data from fourth-grade mathematics in 1992 and 1996, he compared standardized gains on KIRIS with state performance on the

National Assessment of Educational Progress (Koretz & Barron, 1998). Granted, one would expect to see higher growth on KIRIS, which was customized to Kentucky's learning objectives, than on the more general and thereby less curricularly sensitive NAEP measure; but still, the magnitude of the difference gives pause: The KIRIS results showed 3.6 times the growth shown by NAEP. [3]

Steve Klein and colleagues found similar disparities when they examined Texas students' performance on the Texas Assessment of Academic Skills (TAAS) and on the NAEP for the period 1994 to 1998 (Klein, Hamilton, McCaffrey, & Stecher, 2000). For example, the NAEP analysis showed an increase in fourth graders' reading performance from 1994 and 1998, comparable to national trends and with effect sizes of .13 standard deviation units and .15 standard deviations units for students of color. However, TAAS results for fourth graders over this same period showed dramatically greater gains, with effect sizes ranging from .31 to .49 standard deviation units, and with Black and Hispanic students showing substantially greater gains than White students. Thus, NAEP results confirmed neither the slope of the increase in TAAS scores, nor the claim that the achievement gap between White students and those of color was closing.

Beyond these empirical data, it is interesting as well to note that teachers are skeptical about the broader meaning that should be ascribed to score increases. For example, teachers in Stecher and colleagues' survey study of Kentucky (Stecher, et al., 1998) were much more likely to attribute changes in their students' test performance to test-taking skills and test preparation practices than to broad improvements in students' knowledge and skills. These beliefs mirror those found in an earlier study by Koretz and colleagues (Koretz, Mitchell, Barron, & Keith, 1996).

Is There a Relationship Between Intended Changes in Practice and Student Performance?

A parallel question to the question about the meaningfulness of gains is the extent to which desired changes in practice are associated with the improvement of student learning—or at least with observed test score gains. If tests are intended to signal desirable content and pedagogy in which to engage students, does implementing such changes result in intended student effects? The picture here is mixed, but results seem promising.

---

[3] Koretz also analyzed eighth-grade performance. Though the difference was most dramatic at the fourth-grade level, the eighth-grade results also showed differences in trend lines.

A first example comes again from Brian Stecher's work in Kentucky (Stecher et al., 1998) in which he compared teachers' reports of practices in high- and low-gain schools based on second biennium KIRIS gains. Though the study reports few consistent findings across grade and subject areas, and some contradictory results, the bulk of the positive findings do show a relationship between the standards-based practices that KIRIS was intended to stimulate and performance—particularly at the middle school level. For example, significantly more seventh-grade writing teachers in high- versus low-gain schools reported integrating writing with other subjects and increasing their emphasis on various aspects of the writing process, including writing in a variety of genres, use of effective language, sentence structure, logical organization, tone/voice, and idea development. However, these teachers also reported greater attention to specific skills, including regular practice in grammar and English mechanics. Similarly, in eighth-grade mathematics, teachers in high-gain compared to low-gain schools reported a greater increase in their coverage of mathematics topics, particularly statistics and probability; more frequent use of calculators; and an increased use of classroom assessment, including extended investigations. At the eighth-grade level, there also were associations between KIRIS gains and school support for mathematics reform. At the same time, more eighth-grade teachers in high-gain schools relative to low-gain schools reported having their students practice computations on a daily basis. Might it be that teachers in high-gain schools are incorporating reform practices, but in the process, have not lost the basic skills of traditional instruction? Perhaps they have found a way to integrate the two. Finally, one of the few consistent findings across grade levels is worth noting: Mathematics teachers in high-performing schools were more likely than those in low-performing schools to report increases in *both* remedial and enrichment activities for students.

Clement Stone and Suzanne Lane examined similar issues using structural equation modeling and data from the Maryland State Performance Assessment Program (MSPAP; Stone & Lane, 2000). They examined the relationship between changes in MSPAP scores from 1993 to 1998, school demographics and survey-based measures of classroom instruction and assessment practices, student perspectives on their instruction and motivation, and beliefs and attitudes about MSAP. The researchers composed models to explain variability in 1997 or 1998 school performance[4] and rates

[4] Data collection in the various subjects spanned 2 years. Math and language arts survey data were collected in 1996-1997, and science and social studies data were collected in 1997-1998. Thus, 1997 performance was used as the measure of "current" achievement for the first two subjects and 1998 performance data for the latter. Note also that the school samples differed for the two data collections.

of change in performance over time in four subject areas: mathematics, language arts (reading and writing), science and social studies. As with Stecher et al.'s (1998) findings, few results were consistent across subject areas. As would be expected, there was a persistent, strong relationship between socioeconomic status (SES), as measured by free lunch eligibility, and student performance, but no relationship between SES and rates of change in performance. Of particular interest, instruction-related variables that assessed the extent to which practice was aligned with reform goals consistently explained differences in performance across subject areas, except for social studies. That is, practices that were more reform-oriented were associated with higher levels of performance in reading, writing, mathematics, and science. However, only in language arts (writing and reading) were these variables significantly related to improvements in performance over time, and even in these subjects, the practical impact was very small. Other significant results were scattered and quite small. For example, teachers' reports of the impact of MSPAP on their instruction were associated with performance differences in reading and writing and with improvement of performance in math and science. Students' reports of how often they engaged in MSPAP-type tasks and of their perceptions of the importance of MSPAP were *negatively* related to performance in science and social studies. Stone and Lane hypothesized that the student perceptions may have been a proxy for test preparation, in that lower performing schools may have engaged in more test preparation just prior to the test, leading students to give high ratings to the frequency of MSPAP-like activities and to the importance of the test.

Are Instructional Changes Sufficient to Influence Performance?

That teachers' reports of their use of reform practices show limited relationship to student learning should be expected. Decades of research show the difficulty of changing practices and admonish that meaningful change takes time (Cuban, 1993; McLaughlin, 1990). Moreover, available research suggests that teachers' responses to standards and assessments may be initially fairly shallow. That is, teachers indeed listen to the signal sent by standards and assessment and attempt to model them in their practices, but understandably, their initial attempts may be just that—initial, mimicking the superficial features of the intended reform but not incorporating deep understanding or quality implementation.

That initial changes tend to be superficial is evident in the work of a number of researchers employing various methodologies in studies of a number of states. For example, William Firestone and colleagues' (Firestone et al., 1998) study of Maryland and Maine looked in depth at teachers' classroom practices in mathematics. They found

examples in which the superficial features of a task matched the state standards and assessment goals, though in general, they concluded that the reform intent was not fully realized. For example, they cited instances in which teachers tried to incorporate extended projects that asked students to apply mathematical concepts, reason mathematically, and/or use multiple forms of representation; but on deeper examination, the classroom tasks and criteria did not demand the intended depth of understanding and complex thinking. The authors concluded: "Such assessments generate considerable activity focused on the test itself. This activity can promote certain changes, like aligning subjects taught with the test. It appears less successful, however, in changing basic instructional strategies" (p. 95). Lorraine McDonnell and Craig Choisser (McDonnell & Choisser, 1997) came to a similar conclusion based on evidence from their study of Kentucky and northern California. Although teachers implemented new instructional approaches, the depth and complexity of their content and pedagogy did not change in meaningful ways.

Yet additional confirmation comes from Gail Goldberg and Barbara Rosewell's study of MSPAP effects, looking at effects on writing instruction (Goldberg & Rosewell, 2000). They followed up a sample of elementary and middle school teachers who had been involved in scoring state writing assessments to examine the effects of the experience on their instructional practice and to see how well these teachers were implementing the state's vision of standards-based writing reform. The study drew on multiple data sources, including teacher surveys before and after the training and scoring experience, semi-structured interviews and subsequent classroom visits, and analysis of classroom artifacts. Echoing many studies of the impact of large-scale scoring, teachers were highly enthusiastic about the value of their scoring experience. Their responses across data sources indeed showed impact on classroom practices, particularly in terms of eliciting writing for varied purposes, integrating academic content into writing and cueing for complex thinking. However, although teachers were struggling mightily to understand the kinds of tasks and meaning of the rubric that MSPAP demanded, the quality of their implementation was "incomplete and superficial" (p. 257).

What Factors May Influence Effects?

There of course are innumerable factors that may influence how statewide assessment and accountability systems affect classroom instruction and student learning. Here we are interested primarily in those factors that may be part of the accountability system itself—for example, stakes attached to performance, efforts to

support low-performing schools, and district- and school-level leadership and support for improvement. Currently more is known about the variation in these elements across states and localities than is known about their influence on schools, teaching, and student learning.

Stakes provide one example: By attaching consequences to performance, states hope to motivate additional effort and improved learning. However, the nature of the stakes varies from state to state—from publishing test results, to financial and other rewards for schools and/or teachers, to sanctions for principals, teachers, and/or students who do not meet their targets. There is ample evidence to suggest that state assessment systems do create pressure for teachers and principals (see, for example, Aschbacher, 1994; Koretz et al., 1996; Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993), but little clear evidence on how various stakes have differential effects on teachers, their curriculum and instruction, or, ultimately, student learning. In general, studies of teachers' and principals' reactions in states with higher stakes for schools (e.g., Kentucky) show results similar to those in which there currently are no special consequences for schools associated with test performance (e.g., Washington). However, deeper qualitative studies show that there may be differences in how teachers respond (Firestone et al., 1998). Given that some states are making sizeable investments in cash incentives, it seems important to know whether and how such incentives may work, and to investigate more fully the intended and unintended consequences of various rewards and sanctions. Little is known about how stakes for students may interact with those for teachers and schools.

Similarly, states and districts differ in how they respond to low-performing schools, but evidence on whether and how their various responses influence classroom teaching, test performance, and student learning is limited. As Peg Goertz and her colleagues (Goertz, Duffy, & Le Floch, 2001) documented, states are implementing a variety of strategies to help such schools, including support for school improvement or corrective action planning, financial assistance, expert assistance in curriculum planning and instruction, and state-sponsored or regionally sponsored professional development opportunities. States also vary in the resources they make available to aid in these processes, from support teams composed of state and/or local officials, to distinguished educators and regional service centers and external providers. The nature and quality of strategies employed, as well as the level of expertise represented, by available resources are likely to be highly influential in how schools respond to low test scores and the quality of changes they are able to make in the teaching and learning process.

Finally, it goes without saying that one would expect that local school and district leadership would affect whether and how assessment influences teaching and learning. James Spillane, for instance, has explored how the various models of district support may differentially affect success and the ways in which structural constraints, local circumstances, and competing demands on teachers may lead to fragmentation and less-than-optimal improvement efforts (Spillane, 2000), but further research is necessary to identify optimal approaches. Needed, too, is additional research on how schools can best orchestrate their improvement efforts. For instance, Hilda Borko's (Borko, Elliott, & Uchiyama, 1999) and Shelby Wolf's (Wolf, Borko, McIver, & Elliott, 1999) qualitative studies of exemplary sites in Kentucky identified the importance of professional development time and money, coupled with the development of curriculum and assessment activities strongly linked to standards (Borko et al., 1999). Common themes characterizing these exemplary sites included a strong sense of identity as a school, a cooperative view of leadership, strong but reflective alignment with the Kentucky standards and reform agenda, and an unwavering emphasis on students and a commitment that all decisions and actions at the school ought to be for the benefit of children (Wolf et al., 1999). However, not all the sites identified as exemplary at the beginning of the study continued to be identified as exemplary during the term of the study.

Conclusions and Recommendations

A consistent picture emerges from these collective findings: Standards-based assessments can serve to stimulate reform and encourage schools and teachers to focus on teaching specified content, but clearly an assessment or accountability system itself is no panacea for the difficulties of ensuring that all children achieve the standards. Furthermore, there are challenges in the design of current assessment accountability systems that will need continued attention.

The fact that assessment systems encourage teachers to adopt new content and pedagogy and bring their classroom and instruction into alignment with valued knowledge and skills is decidedly good news.[5] Assessment appears instrumental in initiating change and movement from existing practices in schools toward new expectations, including desired standards and pedagogy. It should come as no surprise that simply modeling test content and pedagogy is insufficient to achieve teaching

[5] Ideally, a test represents the standards, but research shows this is not consistently the case. Certainly, however, regardless of whether it matches the standards, what is on state assessments represents valued content knowledge and skills—otherwise, why has the state chosen to use it?

expertise or high-quality implementation of new practices and that there are imperfections in the current systems. Simply getting the system moving is no mean feat, and we need to capitalize on the existing momentum and continue to move forward productively toward the vision of standards-based assessment. There are implications here for the types of assessment systems we need to design and the types of capacity we need to help teachers and schools develop.

Multiple Measures

First, from the assessment side, the findings underscore the importance of having assessment systems that are aligned with our standards and, as Lauren Resnick put it, long ago, "tests worth teaching to" (Resnick, 1996). The evidence is strong: Teachers respond to what we ask of them and teach what is tested. If we are serious about standards and want teachers to teach them, our assessment systems simply must measure the depth and breadth of those standards.

As measurement experts, we know that a single measure cannot serve all purposes or fully cover a domain or discipline, nor can it be responsive to the reality of individual differences. Students and schools need multiple and diverse opportunities to show what is being learned. Multiple-choice measures can go only so far in tapping the complex thinking, communication, and problem-solving skills that students need for future success. With multiple measures in the system, teachers also could not overly fixate on a narrow range of content. The school-based inquiry that standards-based reform seeks to encourage could not devolve so easily into a microanalysis of how students perform relative to the specific knowledge, skills, and formats covered by the test, nor could implications for action become a curriculum of test preparation. Rather, the existence of multiple measures, assuming they reflected a coherent, standards-referenced system, might encourage teachers and schools to reflect on what the standards really mean and to internalize an overall framework into which the multiple measures fit. The multiple measures themselves would help to communicate the range and complexity of expectations for student performance.

Coordinated Systems of Local and Classroom Assessment

This is not to say that all "multiple measures" must emanate from the top down or be part of an annual state "test." There are limits, of course, to how much time and other resources can be devoted to such testing. There are limits, as well, to the depth of knowledge and information such tests can provide. To truly understand why student performance is as it is and to get to the root of whatever teaching and learning issues

may exist, schools and teachers really need to move to a more detailed level of assessment and analysis than annual state tests afford. Schools and teachers need to be able to supplement the external assessment results with other, local data. No matter how well aligned and how sensitively crafted, these assessments can offer only a limited perspective on what children really know and can do relative to standards and what factors may be working against their progress.

How do we assure a picture closer to the vision as intended? The answer lies at least partially in coordinated systems of local assessments: district, school, and/or classroom assessments that are aligned with standards and that can provide educators with the diverse forms of evidence they need to understand and improve their students' learning. Moreover, integrated with classroom curriculum and/or administered periodically over the course of the year, such local assessments are also necessary to provide teachers with essential, ongoing information to gauge student progress and adjust teaching and learning opportunities accordingly. Ultimately, these are the "multiple measures" that really can make a significant difference in student learning. Good teaching is a process of continual assessment and adjustment—waiting until the external results show up annually or even semi-annually just is not enough. Such multiple measures also could provide a safeguard against simply "teaching the test" and a potential wealth of data against which the validity of gains could be judged—by parents and students as well as by external authorities. Developing the capacity for local and classroom assessment, furthermore, should help to build and support the credibility of teacher judgment, because such measures and judgments could be based on sound, visible evidence and ultimately reduce the enormous pressure that now rides on external, one-time assessments.

Capacity Building

The findings reported here and future directions for improvement show strong implications for teacher and school capacity building. That the content and pedagogical signals sent by an assessment are insufficient to enable teachers' mastery of new approaches, as indicated earlier, should come as no surprise. Moreover, the vast majority of teachers, schools, and districts lack the capacity to engage in the vision of coordinated local and classroom assessments that have just been described.

Surely, most states provide some attention to professional development along with their assessment systems, but such professional development is likely to deal with the tests themselves and their administration, and/or the mechanics of understanding scoring and scoring reports, and be of limited duration—the kind of one-shot

opportunities that we know are of limited valued. Even more intensive involvement in state or district scoring, the professional development value of which has been highly touted (Aschbacher, 1994; Falk & Ort, 1997; Sheingold, Heller, & Storms, 1997), is insufficient for meaningful change. As Goldberg and Rosewell (2000) characterized the effects of such experiences, "like Socrates the wise man who knows that he does not know all, teachers report that the experience [of training and rubric-based scoring of state writing assessments] highlights for them the as yet unfulfilled need for resources and professional support to meet demands and expectations that only grow greater and more complex with their increased understanding of the issues and implications of performance-based instruction and assessment" (p. 286).

Generally absent are the types of sustained, intensive, and ongoing professional development opportunities that would enable teachers to engage in standards-based reform and well use assessment within that context. Ample research shows that such opportunities are embedded in, and responsive to, the local environment; permit teachers to gain, apply, and progressively appropriate new content and pedagogical knowledge in supportive circumstances that provide coaching and mentoring; and encourage active reflection and problem solving (see, for example, Cohen & Ball, 1999; Darling-Hammond & Ball, 1998).

The instinct to simply "teach to the test" may in part be a survival instinct. Lacking alternative strategies or effective avenues for acquiring them, teachers do what they can and what they know how to do to reach targeted goals. Just as we need coordinated systems of assessment, so, too, do we need coordinated systems of professional development that align preservice and in-service professional development programs with a comprehensive and integrated understanding of the requirements of standards-based instruction and assessment.

Ongoing Evaluation to Support Validity and Positive Consequences

That good intentions are insufficient to assure good consequences from assessment is a lesson that has been learned repeatedly over the last century—whether the topic has been admissions testing, objective-referenced testing, minimum competency testing, or performance assessment. That the stakes associated with performance in standards-based assessment systems are on the increase across the country, furthermore, also increases the likelihood of corruption of test results. That is, such stakes may create incentives for some schools and teachers to teach only to the test—or worse—and such actions, as was documented earlier, can invalidate the meaning of the results and the inferences about student learning and progress that can be drawn from such results.

These possibilities underscore the importance of ongoing evaluation of standards-based systems, as advocated by current standards for accountability systems (see Baker, Linn, Herman, Koretz, & Elmore, 2001).

Validity of scores. Questions about the validity of gains and whether score increases truly signal increases in learning,[6] coupled with the high stakes attached to test results, make it essential that safeguards be built into accountability and assessment systems. If we are not confident that substantial increases in test performance really signify meaningful improvement in student learning, it is difficult to justify delivering substantial rewards or meting out severe punishments based on test scores alone. Rather, there need to be additional checks and balances in the system to verify the quality or level of learning in identified schools and to assure that schools get what they deserve. Evidence derived from coordinated systems of local assessment could be used in such a verification process, as could spot checks or monitoring of the quality and comprehensiveness of classroom curriculum and instruction. Promising approaches to assessing the quality of classroom practice exist (see, for example, Aschbacher, 1999; Clare, 2000) and should be considered as components in an accountability system. Such checks not only could assure the fairness of rewards and sanctions but furthermore should mitigate against teaching solely to the test.

Similarly, at the state level, there need to be ongoing studies of the validity of state assessment results and convincing evidence mounted that increases in test scores translate into meaningful improvements in student learning for all students. The validity of gains for traditionally underperforming subgroups deserves special scrutiny, as closing the gap is a prime goal of standards-based reform, and disadvantaged subgroups are the ones who are at most risk of curricular corruption. Studies cited in this review provide possible models for looking at the relationship between gains on a particular state assessment and gains on NAEP and/or other measures of performance that may be better aligned with that state's standards. One would want also to assure test scores were sensitive to quality instruction and well-honed improvements in standards-based, classroom practice.

Consequences of assessment systems. Beyond issues of the validity of gains, the findings cited in this report make it clear that there are gaps between the vision and current practice. It therefore is essential that we continue to evaluate the claims

[6] Teaching to the test is but one reason why changes in test scores may or may not signal true improvement or decrements in student learning or school effectiveness. The instability of scores from year to year (Kane & Staiger, in press; Linn & Haug, 2001) and issues of accuracy and classification error (Rogosa, 2000) are also important problem sources.

supporting standards-based assessment systems and regularly examine the actual consequences of such systems. The accountability standards advocate regularly assessing system effects on capacity building, resource allocation, instructional effects, equity and access to education, teacher quality, recruitment and retention, and unanticipated outcomes. For example, is there sufficient capacity at the district, school, and classroom levels to support standards-based reform? How and to what extent are the accountability system and its results being used to marshal capacity to support improvement? How and to what extent are results used for resource allocation and to assure that resources and attention get to the children and standards that are most in need of attention? Equity in resources and capacity to deliver effective standards-based programs should be an important, continuing issue, based on findings cited in this report (Firestone et al., 2000; Herman & Golan, 1993; Smith & Rottenberg, 1991).

That the accountability system will focus teachers and schools on teaching to the standards and improving instructional practice has been the prime focus of this report, and indeed the research on instructional effects shows both good news and bad news. There needs to be continuing study, particularly of schools serving students at risk. We need to check our assumptions about the effects of accountability systems on equity and providing all children access to opportunity. Both history and specific studies cited in this report provide cause for concern: Despite intended consequences, is the gap increasing or decreasing between economically advantaged and disadvantaged groups? Between Caucasian children and those of color? How are English language learners faring? Students with disabilities? What of equity in the curriculum and instruction offered in schools serving traditionally underperforming groups and other students? Is instruction for the former devolving into test preparation, while schools serving wealthier students benefit from more varied instructional resources and a richer curriculum that provides better opportunities to develop the complex thinking and communication skills students will need for future success?

The evaluation questions are complex and varied and deserve continuing inquiry. Just as standards-based assessment is intended to improve the quality of the educational system, so, too, should the evaluation of assessment and accountability systems lead to their continued improvement. Building on the current momentum, such improvement should enable systems that can better deliver on the promise of standards-based reform.

References

Aschbacher, P. R. (1994, June). Helping educators to develop and use alternative assessments: Barriers and facilitators. *Educational Policy, 8*, 202-223.

Aschbacher, P. R. (1999). *Developing indictors of classroom practice to monitor and support school reform* (CSE Tech. Rep. No. 513). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., Linn, R. L., Herman, J. L., Koretz, D., & Elmore, R. (2001, April). *Holding accountability systems accountable: Research-based standards.* Symposium presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Borko, H., & Elliott, R. (1998). *Tensions between competing pedagogical and accountability commitments for exemplary teachers of mathematics in Kentucky* (CSE Tech. Rep. No. 495). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Borko, H., Elliott, R., & Uchiyama, K. (1999). *Professional development: A key to Kentucky's reform effort* (CSE Technical Report No. 512). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Borko, H., & Stecher, B. M. (2001, April). Looking at reform through different methodological lenses: Survey and case studies of the Washington state education reform. In J. Manise (Chair), *Testing policy and teaching practice: A multi-method examination of two states.* Symposium conducted at the annual meeting of the American Educational Research Association, Seattle, WA.

Chapman, C. (1991, June). *What have we learned from writing assessment that can be applied to performance assessment?* Presentation at ECS/CDE Alternative Assessment Conference, Breckenridge, CO.

Clare, L. (2000). *Using teachers' assignments as an indicator of classroom practice* (CSE Tech. Rep. No. 532). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing .

Cohen, D., & Ball, D. L. (1999, June). *Instruction, capacity, and improvement* (CPRE Research Rep. No. RR-043). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion.* Norwood, NJ: Ablex Publishing.

Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1890-1980* (2nd ed.). New York: Teachers College Press.

Darling-Hammond, L. (1995). Equity issues in performance based assessment. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 89-114). Boston, MA: Kluwer.

Darling-Hammond, L., & Ball, D. L. (1998, November). *Teaching for high standards: What policymakers need to know and be able to do* (CPRE Research Rep. No. JRE-04). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Darling-Hammond, L., & Wise, A. E. (1988). *The evolution of teacher policy*. Santa Monica, CA: The RAND Corporation.

Dorr-Bremme, D., & Herman, J. (1986). *Assessing student achievement: A profile of classroom practices* (CSE Monograph Series in Evaluation No. 11). Los Angeles: University of California, Center for the Study of Evaluation.

Falk, B., & Ort, S. (1997, April). *Sitting down to score: Teacher learning through assessment*. Presentation at the annual meeting of the American Educational Research Association. Chicago.

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000, April). State standards, socio-fiscal context and opportunity to learn in New Jersey. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. *Educational Policy Analysis Archives 8*(35). Retrieved August 23, 2002, from http://epaa.asu.edu/epaa/v8n35/

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998, Summer). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20,* 95-113.

Goldberg, G. L., & Rosewell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance based instruction and classroom practice. *Educational Assessment 6,* 257-290.

Goertz, M. E., Duffy, M. C., & Le Floch, K. C. (2001). *Assessment and accountability systems in the 50 states: 1999-2000* (CPRE Research Rep. No. RR-046). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Herman, J. L., & Golan, S. (1993). Effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12*(4), 20-25, 41-42.

Herman, J. L., & Klein, D. (1996). Evaluating equity in alternative assessment: An illustration of opportunity to learn issues. *Journal of Educational Research 89,* 246-256.

Herman, J. L., Klein, D. C. D., Heath, T. M., & Wakai, S. T. (1995). *A first look: Are claims for alternative assessment holding up?* (CSE Tech. Rep. No. 391). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kane, T. J., & Staiger, D. O. (in press). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings Papers on Education Policy*. Washington, DC: Brookings Institution Press.

Kellaghan, T., & Madaus, G. (1991). National testing: Lessons for America from Europe. *Educational Leadership, 49*(3), 87-93.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). *What do test scores in Texas tell us?* (RAND Issue Paper IP-202). Santa Monica, CA: RAND.

Koretz, D., & Barron, S. (1998). *The validity of gains in scores on the Kentucky Instructional Results System (KIRIS)*. Santa Monica, CA: RAND.

Koretz, D., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)* (MR-792-PCT/FF). Santa Monica, CA: RAND.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (CSE Tech. Rep. No. 355). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Mitchell, K. J., Barron, S., & Keith, S. (1996). *Perceived effects of the Maryland State Assessment Program* (CSE Tech. Rep. No. 409). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Stecher, B., & Deibert, E. (1992). *The Vermont Portfolio Assessment Program: Interim report on implementation and impact, 1991-92 school year* (CSE Tech. Rep. No. 350). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience* (CSE Tech. Rep. No. 371). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Lane, S., Stone, C. A., Parke, C. S., Hansen, M. A., & Cerrillo, T. L. (2000, April). *Consequential evidence for MSPAP from the teacher, principal and student perspective.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Linn, R. L., & Haug, C. (2001). *Stability of school building accountability scores and gains* (CSE Tech. Rep. No. 561). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L., & Herman, J. L. (1997). *Standards-led assessment: Technical and policy issues in measuring school performances* (CSE Tech. Rep. No. 426). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing; Denver, CO: Education Commission of the States (ECS).

McDonnell, L. M., & Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments* (CSE Tech. Rep. No. 442). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

McLaughlin, M. W. (1990). The RAND Change Agent Study revisited: Macro perspectives and macro realities. *Educational Researcher, 19*(9), 11-16.

Orlofsky, G. F., & Olson, L. (2001, January 11). The state of the states [Electronic version]. *Education Week. Quality Counts 2001,* pp. 86-88. Retrieved September 24, 2001, from http://www.edweek.org/sreports/qc01/articles/qc01story.cfm?slug=17states.h20

Quellmalz, E., & Burry, J. (1983). *Analytic scales for assessing students' expository and narrative writing skills* (CSE Resource Paper No. 5). Los Angeles: University of California, Center for the Study of Evaluation.

Resnick, L. B. (1996). *Performance puzzles: Issues in measuring capabilities and certifying accomplishments* (CSE Tech. Rep. No. 415). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Rogosa, D. (2000). *Accuracy of year-1, year-2 comparisons using individual percentile rank scores: Classical test theory calculations* (CSE Tech. Rep. No. 510). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2000). *Benchmarking and alignment of standards and testing* (CSE Tech. Rep. No. 566). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Sheingold, K., Heller, J., & Storms, B. (1997, April). *On the mutual influence of teachers' professional development and assessment quality in curricular reform.* Presentation at the annual meeting of the American Educational Research Association, Chicago.

Shepard, L. (1990). *Inflated test score gains: Is it old norms or teaching the test?* (CSE Technical Report No. 307). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*(7), 2-16.

Smith, M. L. (1997). *Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program (ASAP): Equity and teacher capacity building* (CSE Tech. Rep. No. 425). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Smith, M. L., Edelsky, C., Draper, K., Rottenburg, C., & Cherland, M. (1990). *The role of testing in elementary schools* (CSE Tech. Rep. No. 321). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Smith, M. L., & Rottenberg, C. (1991, Winter). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7-11.

Spillane, J. (2000). *District leaders perceptions of teaching learning* (Res. Rep. No. OP-05). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Stecher, B., & Barron, S. L. (1999). *Quadrennial milepost accountability testing in Kentucky* (CSE Tech. Rep. No. 505). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B., Barron, S. L., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classroom.* (CSE Tech. Rep. No. 525). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B., Barron, S. L., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing* (CSE Tech. Rep. No. 482). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B., & Borko, H. (2002). *Combining surveys and case studies to examine standards-based educational reform* (CSE Tech. Rep. No. 565). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stone, C., & Lane, S. (2000, April). *Consequences of a state accountability program: Relationships between school performance gains and teacher, student, and school variables.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Wolf, S. A., Borko, H., McIver, M., & Elliott, R. (1999). *No excuses: School reform in exemplary schools of Kentucky* (CSE Tech. Rep. No. 514). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S. A., & McIver, M. C. (1999). When process becomes policy: The paradox of Kentucky state reform for exemplary teachers of writing. *Phi Delta Kappan, 80,* 401-406.