Making Accountability Work to Improve Student Learning

CSE Report 649

Joan Herman
CRESST/University of California, Los Angeles

March, 2005

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Project Director: Joan Herman

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

MAKING ACCOUNTABILITY WORK TO IMPROVE STUDENT LEARNING

Joan L. Herman
National Center for Research on Evaluation, Standards and Student Testing (CRESST)
UCLA Graduate School of Education

Abstract

That *No Child Left Behind* (NCLB) places unprecedented demands on districts and schools to improve student performance is trite and a truism. That NCLB places unprecedented demands on the design and productive use of accountability systems may be less well appreciated. In this article, I consider how accountability is supposed to work to support the improvement of student learning; how it does work, based on available research evidence; and finally what might be done to make it work better.

The Model of Standards-Based Reform

Standards-based reform is built on the assumption that being explicit about learning goals and measuring students' progress toward these same goals will help to improve student learning. The logic seemingly is straightforward and linear, even if the reality is not: we agree on standards for what students ought to know and be able to do, agree as a society and as communities of educators and schools that we will work with and expect *all* students to achieve these standards, develop measures that tell us how well we (educators, students) are doing, and then use feedback from the measures to analyze the quality of programs and the strengths and weaknesses of student learning and subsequently to improve educational opportunities and to help assure every student's success. Though our conceptions of goals have grown more ambitious and our theories about effective pedagogy have been dramatically transformed, the basic outline of establishing goals, making plans to attain them, measuring progress, and then revising and refining our efforts based on results is a time worn and familiar process, with roots harkening back to such concepts as Skinner's behaviorism and programmed instruction, Bob Glaser's early articulation of criterion-referenced measurement (1959), Benjamin's Bloom's mastery learning, and relatively more recent renditions of assessment- driven

reform (Popham et al. 1985) and the power of tests worth teaching to (Resnick & Resnick, 1992). These core ideas also have gained currency and credibility from private sector applications, for example as reflected in total quality management and benchmarking strategies commonly popular in business.

However, while often conveyed as technical steps in a familiar problem solving process, accountability at its heart is a system for motivating performance. The intent is not only to provide a technical system that can measure performance and provide data to support improvement, but more importantly, the system is intended to stimulate purposeful reform to achieve agreed upon standards. As a policy lever, the system serves symbolic purposes in: establishing the target for reform efforts; communicating to educators, administrators and parents what is expected; providing incentives and/or sanctions; and thereby motivating all levels of the education system to focus on achieving the policy goals. In the case of *No Child Left Behind*, prime among these goals is the assurance of schools' adequate yearly progress toward all children being proficient by the year 2014.

Figure 1 shows one view of how accountability is supposed to work, focusing particularly on the quality of classroom teaching and learning necessary to enable students to reach the standards. While the full and coordinated support of all levels and resources of the educational systems may be needed to achieve policy goals, it seems axiomatic that students cannot be expected to become proficient unless and until the content and process of their classroom instruction well prepares them to do so. As the figure shows, standards are the basis for accountability assessments and likewise are the targets of classroom and learning. Feedback is used to improve learning opportunities for student and to increase their attainment of standards.

What should be apparent from the figure is the importance of several technical features of the system. First, the alignment of standards, assessments, and classroom instruction is critical to the validity of the system. For example, if external assessments do not match the standards well, using feedback from them to make adjustments may well distort curriculum and divert attention from the important goals. Yet even with tight alignment, the figure tries to makes clear that a test can only measure a portion of what students are learning and therefore is imperfect. Tests can only assess that which can be measured in whatever finite periods of time are allocated to testing and through the types of formats that are

included in the tests—meaning that it is impossible for tests to assess everything that is important. Furthermore, all measures also contain error and thus provide only an imperfect *estimate* of student performance relative to standards—meaning that we should never base any important decision on the basis of a single test.

Further, state assessments are not the only assessments of importance in the system. The continuous improvement model that accountability envisions means that educators must keep their eyes on student learning; regularly assess how students are doing relative to the standards; use the information to understand what students need; and take appropriate, meaningful action based on results.

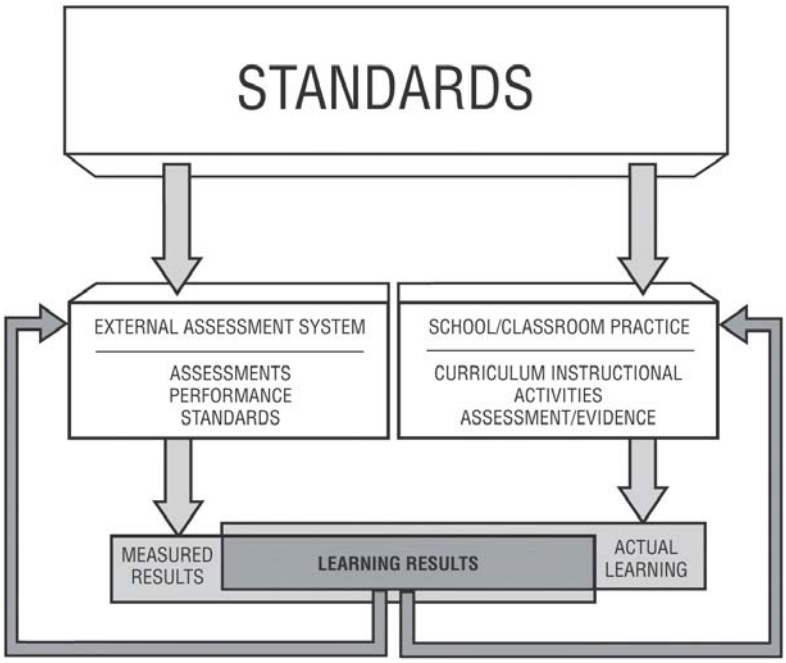

figure 1

How Is Accountability Working? Research Evidence on Effects on Teaching and Learning

If accountability systems are intended to serve both symbolic and technical functions, one can ask both how well the motivation system is working to stimulate desired actions and how well the technical information system is working to provide appropriate inferences. Selected research findings related to each of these questions follow.

Ample research suggests that accountability systems can be powerful in communicating expectations and stimulating teachers and schools to modify their teaching; educators actively work to attain the goals that have been established for student performance. Studies conducted in numerous states, among them 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 et al., 1998; Goldberg & Rosewell, 2000; Lane, Stone, Parke, Hansen, & Cerrillo, 2000), New Jersey (Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000), North Carolina (McDonnell & Choisser, 1997), Vermont (Koretz, McCaffrey, Klein, Bell, & Stecher, 1993), and Washington (Borko & Stecher, 2001; Stecher, Barron, Chun, & Ross, 2000) using a variety of quantitative and qualitative methodologies have shown quite consistent results.

Accountability Tests Serve to Focus Instruction

Teachers and principals indeed pay attention to what is tested and adapt their curriculum and teaching accordingly. Principals, sometimes with and sometimes without the involvement of their staff, analyze test results and develop school plans to concentrate on areas where test results show a need for improvement. Almost all principals also take action to assure that students at their schools engage in direct test preparation activities during classroom instruction. Teachers consistently report that state tests have a substantial effect on what they teach and how they assess student learning.

Teachers Model What is Assessed

Moreover, teachers tend to model the pedagogical approach reflected in the test. When a state or district assessment is composed of multiple-choice tests, teachers tend to rely heavily on multiple-choice worksheets in their classroom

instruction. However, when the assessments use open-ended items and/or extended writing and rubrics to judge the quality of student work, teachers prepare students for the test by incorporating these same types of activities in their classroom practice. Direct test preparation activities, which capture significant time in many schools, also directly mimic the content and format of the test. Such modeling of test content and pedagogical approach provides an opportunity to stimulate important changes in teachers' practice.

Test Scores Show Initial Increases

Such sustained attention to test content and format tends to show up in test performance. In state after state, when new assessments and accountability provisions are put into place, student scores show an increase, at least for the first few years. For example, California elementary school students showed a 12-point gain in the percentage of students scoring at or above the 50th percentile from 1998 to 2001, as second graders in 1998 progressed through to the fifth grade in 2001 (Herman & Perry, 2002).In Texas, the percentage of students passing the Texas Assessment of Academic Skill (TAAS) rose from 55.6% in 1994 to 85.3% in 2002, an increase of nearly 30% more students passing (Texas State Department of Education website).

While these first three points demonstrate that some aspects of accountability are working as intended, the research also suggests areas where there may be unintended consequences that need to be remedied.

Schools Focus on the Test Rather Than the Standards

At least initially, educators appear to give their primary attention to what is tested and how it is tested, rather than to the standards themselves. Teachers in Washington, for example, reported that their instruction tended to be more like that Washington state assessment than the state's standards (Stecher & Borko, 2001) and emphasized the specific knowledge and skills they expected to be tested, while elementary school teachers in Washington and Kentucky accorded priority to particular subject matters and topics depending on whether the subject was assessed at their grade level (Stecher and Barron, 1999; Stecher et al., 2000). As a result, math received relatively more time and attention relative to language arts in grade levels at which math was tested and vice versa for grade levels at which language arts was tested. In short, *what* was tested and *when* it was tested caused significant shifts in teachers' use of classroom time both within

and across subjects, and these changes were not motivated by any coherent sense of curriculum nor driven by the need to continuously develop students' learning within and across grade levels. Such test-based decision-making has the potential to distort curriculum.

What is not Tested becomes Invisible

As a corollary, focusing on the test rather than the standards also means that what does not get tested tends to get less attention or may be ignored all together. This seems true both within and across subjects. For instance, using math as an example, if extended math problems are not included on the test, instructional time may go to the computation or other problem types that are on the test. Similarly, as more time is devoted to the tested subjects—typically reading, language arts and mathematics—such time must come from other areas of the curriculum. Both the broader domain of the tested disciplines and important subjects that are not tested may get short shrift.

Is Accountability Working? Selected Technical Issues

As the research clearly shows that accountability testing is working to influence educators' behavior, it underscores the importance of assuring quality tests whose results are worthy of attention. Unfortunately, the research shows some problems, among them the relationship between state accountability tests and the standards they are supposed to measure, and questions about whether the increases in state assessment scores reflect real learning. The feasibility of attaining ambitious goals for adequate yearly progress also gives pause.

State Assessments Show Uneven Alignment with Standards

Alignment is the lynchpin of standards-based reform. As noted above, the alignment between a state's standards and its accountability assessments is essential. But alignment is a term that can have many different meanings. At the simplest level, for example, one can ask whether the items on a state's test correspond to or are relevant to any of that state's standards for a particular subject area. While the vagueness and lack of specificity in the standards of many states can make this determination difficult, most state tests do quite well by the simple relevance criterion. Available evidence suggests that most test items on state tests do reflect some standard on the state's list. However, by this definition a state test would be considered aligned if all the items on the test addressed a single standard and all other standards were ignored.

Comprehensiveness and balance in alignment are trickier to achieve, yet both are essential if a test is to well represent a state's standards. Comprehensiveness is the extent to which all standards or benchmarks are addressed by a test or set of assessments, and balance reflects the extent to which some standards or benchmarks may be privileged over others in terms of relative emphasis or number of items. Not surprisingly, studies conducted in more than 10 states by Achieve and those led in a number of other states by Norman Webb suggest that existing tests do not fully cover intended standards. Moreover, it appears that existing tests tend to emphasize lower levels of knowledge and skills. What is tested seems at least as much a function of the items particular item writers are most adept at producing and those that survive psychometric field-testing—e.g., items that are at appropriate levels of item difficulty and relate in empirically coherent ways to other items—as of what sets of items will provide the most comprehensive and balanced view of how students are achieving relative to standards.

Questions Arise about Whether Increases in Test Score Increases Signal Real Increases in Learning

As noted above, test scores in the first years of a state accountability test are likely to show substantial increases. However, results tend to level out in subsequent years. For example, in the first 3 years of California's current accountability system, elementary schools showed the greatest increase in the first year, less so in the second and less still in the third. Scores at middle and high school levels tended to level out even sooner. Some have interpreted such leveling off of performance to mean that schools can get an initial boost in performance through test preparation and concentrated focus on test content, but that sustaining progress in the longer term requires more substantial changes in teaching and learning processes in schools than has thus far occurred.

Those who question whether increases in test scores reflect real improvement in teaching and learning also point to disparities between student performance on state accountability tests and that on other achievement measures that are intended to measure similar areas of learning. If test performance represents real learning, then we expect that learning to generalize to or show up on other measures of students' achievement. For example, if students score well on a state's reading test, we interpret that to mean students are doing well in reading and thus expect them to do well in other situations that

require good reading skills. For example, we might expect them to be able to read and understand grade-level books and to score well on other independent measures of reading. However, Bob Linn has shown that when states change from one test to another, their test scores typically plummet, raising questions about validity of previously observed gains. (See, *Standards based accountability: Ten suggestions*, CRESST Policy Brief.)

Moreover, the dramatic upward trends found in state accountability test results are not mirrored in the results of tests that hold less substantial consequences for local educators or students, such as the National Assessment of Educational Progress (NAEP). Dan Koretz and colleagues found Kentucky fourth grade students showed nearly four times the growth on Kentucky's math assessment over the period 1992 to 1996 than was evident in Kentucky's students on NAEP over the same period. Stephen Klein and colleagues found similar disparities between TAAS results in Texas and NAEP. Granted one would expect students to perform better on a state test that is customized to a state's curriculum priorities than on a curriculum free measure such as NAEP. However, these disparities are sizeable enough to question whether the state test score increases are inflated.

We care about students' performance on a test, after all, because we believe that it represents something larger than 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, only on what is tested and on the formats in which it is tested, the test ceases to be a sample of performance. The test becomes the domain and the generalizability of the results—to 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.

The Reliability of School Score Changes from Year to Year is Uncertain

The reliability of test scores is an issue regardless of any potential inflation. As noted above, all test scores are fallible and contain error. The test scores that students achieve on a particular day and time reflect their actual capability as well as errors introduced by how the students felt on the day of the test, how attentive they were on a moment-by-moment basis to the cues and questions in the tests, how carefully they completed their answer sheet, how much they studied or were prepared on the specifics of what was actually tested as opposed to other content and items that might have been on the test, and many other factors. Test scores at the school level similarly are an amalgam of students' actual knowledge and skills and error, including fluctuations associated with sampling error, i.e., who was actually tested one year to the next. One may imagine a substantial difference in a school's test result depending only on whether all children were tested or whether unusual proportions of high or low ability students were absent on the day of the test. While *No Child Left Behind* tries to control the effects of such sampling error by insisting that virtually all students are tested, it cannot ameliorate the problem in year-to-year comparisons. Students who are tested from one year to the next can change substantially because they may move in and out of their school neighborhoods and because, particularly when the certain tests are given only at certain grade levels, the students who actually take the tests are different. For example, if a school's results for Grades 3 to 5 in 2002 are compared to those for Grades 3 to 5 in 2003, it is clear that, at best, there only could be approximately 2/3's overlap in the two samples—those who were in grades 3 and 4 in 2002 and moved to grades 4 and 5 in 2003. As a result, schools scores can bounce around from year to year, irrespective of any change in student learning, a phenomenon having important implications for meeting the annual yearly progress goals of NCLB. For example, Linn and Haug (2002) find fewer than 5 percent of Colorado's schools consistently grew at least one percentage point on the Colorado Student Assessment Program from 1997 to 2000, even though schools on average showed nearly a 5 percent increase over the three year period in the number of students deemed proficient. Combining school scores over several years and establishing minimum group sizes (as allowed by NCLB) reduces the volatility, but does not eliminate the problem.

The Colorado figures also demonstrate the ambitiousness of NCLB goals. Having *all* students proficient by 2014 would require a substantial acceleration over the improvements that thus far have been made, to say the least. Assuring that all subgroups uniformly meet goals for adequate yearly progress clearly magnifies the volatility issues associated with small group size and represents a gargantuan educational challenge. This is particularly the case with English learners who must learn English in addition to acquiring the knowledge and skills they need to meet content standards, and students with disabilities who would not be so identified if they did not have serious learning problems.

How Can Accountability Work Better? Some Advice

All of us certainly may agree on the underlying goals of *leaving no child behind*, but we also must recognize that current accountability systems are imperfect mechanisms for achieving the intended goals. While some aspects of the system may be working as intended to motivate performance, others may function to undermine ultimate success. The more we can recognize these potential problems, the more we may be able to deal with them and to use accountability and NCLB to achieve serious progress for children. Gaming the system or focusing solely on test preparation, as the research makes clear, will not achieve the goals of NCLB. Dealing realistically with current circumstances and potential dangers, in light of available research, may help guide the way.

Work to Assure that Accountability Tests are Aligned with Clear Standards for Students' Performance

The research strongly suggests that educators, particularly those in poor schools who are under pressure to show improvement, are teaching to the test, not the standards. Accountability tests are the lens through which the standards are interpreted: they define the standards. Standards in subjects not tested and standards that are not included in subject matter tests seem to get, at most, weak treatment in classroom teaching and learning. As the stakes associated with test performance rise under *No Child Left Behind*, and in the absence of policies and procedures to dissuade it, such curriculum distortions are likely to increase.

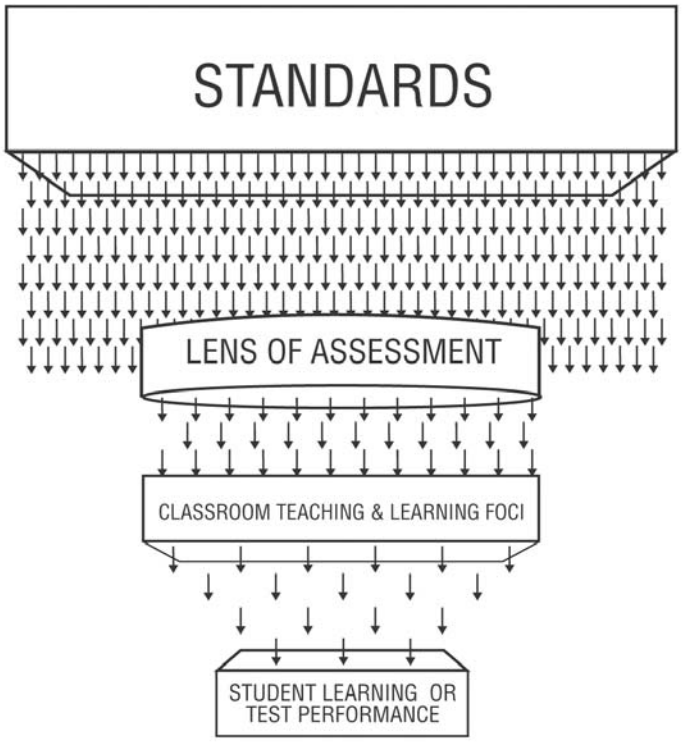

figure 2

As Figure 2 attempts to makes clear, a focus on tests rather than standards has serious consequences for students. Rather than being exposed to the full breadth of knowledge and skills that society, through its standards, has determined are important for future success, students have the opportunity to learn a relatively narrow, test-based curriculum. With the specifics of the test—rather than the essentials of the discipline or meaningful learning—as a primary focus, there also is danger of test score inflation. Students may only be learning what is tested, and increases in test scores may not generalize to other situations. Test results may only be telling us how well students did on a particular test. Moreover, potential mismatches between tests and standards can lead educators and policy makers to misinterpret test results and fail to address genuine needs. Such a system could take us in the wrong direction.

While as educators and administrators we may not be able to control the contour or details of our states' tests, we can act as informed consumers and advocate for quality tests. We must continue to ask for evidence that tests which are used for accountability purposes are balanced and sufficiently comprehensive to support intended inferences and that they indeed tell us the extent to which students are making progress in attaining established standards. To the extent that state assessments give short shrift to the meaning of standards, the instruction and teaching of children is probable to do likewise. Moreover, we must insist on evidence that test results are sufficiently accurate and reliable to support intended decisions.

Integrate a Variety of Local and State Measures to Understand and Support Students' Attainment of Standards

That a single test cannot address all that is important for students to know and be able to do is axiomatic. Multiple measures are needed to address the full depth and breadth of our expectations for student learning. The multiple choice and short answer type items that tend to predominate in large scale accountability tests can only go so far in tapping the complex thinking, communication and problem solving skills that students will need for future success. Other types of performance measures—essays, applied projects, portfolios, demonstrations, etc.—are needed to guide students' progress. Moreover, multiple types of measures can better respond to the reality of individual differences than can a single test. Just as not all students learn in the same way, not all students can demonstrate their proficiency in the same way. Some may be do better in some formats and contexts than in others.

Regularly and Richly Assess Students' Progress Toward Standards

While the notion of multiple measures may seem unrealistic, these are essential for meeting NCLB goals. No matter how well aligned and how sensitively crafted, accountability assessments can only offer a limited perspective on what children really know and can do relative to standards and what factors may be impeding their progress. In order to 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 must 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, both to acquire the deep understanding they need to improve the

learning process and to get regular information during the school year about whether and how students are progressing. Waiting for the once or twice a year test results is too late to make a difference in student learning. Rather, assessment must be on-going, teaching must be tailored to students specific needs, and students who are faltering need to be given special attention. District, school, and/or classroom assessments that are aligned with standards and coordinated with accountability tests are needed to provide educators with the diverse and regular forms of evidence that they need to understand and improve their students' learning. Such evidence also has potential in providing alternative sources of information to document progress for various audiences.

Empower Educators to Improve Teaching and Learning

While accountability is a top-down, policy strategy to promote improvements in student learning, it will not work in the absence of talent and creativity at the local level. Accountability may help to provide the motivation to change, but educators must be assisted to acquire the capacity they need to increase student learning. Good teaching is an intense problem-solving endeavor: it requires sophisticated content knowledge and pedagogical finesse, including sensitivity to students' needs and motivations, to take students from where they are to where they need to be. There is no single cookie-cutter approach that will meet the needs of all students. Policy and capacity building efforts must respect educators as professionals and help them develop the expertise they need to do well by students.

The area of assessment is a special need for teachers. While accountability underscores the importance of assessment in improving student learning, the reality is that most educators are little prepared to regularly well assess their students.

Social as well as Academic Capital Needs Attention

While students' academic progress may be the primary goal in NCLB, relationships may make a big difference in whether and how academic goals are achieved. Research, particularly at the middle school and high school levels, shows the importance of students' feeling a sense of connection and commitment to schooling, safety, positive norms, and efficacy also are essential. (See, *Forum for Youth Investment, 2003.)*

Social capital is the glue that holds a community together. It creates a sense of mutual obligation and a network of support for reaching goals. As we plan to help students achieve high academic standards, we must do so in ways that develop their social capital as well.

This is a particularly important concept in empowering teachers and local schools to achieve high standards with their students. Educators must be able to work together to marshal their collective knowledge to reach common goals and make a difference for kids. The ways we work with professional staff must reflect this commitment.

Learn from Success

While the NCLB goals are ambitious, there are schools that are well on their way to achieving them. Hilda Borko's research (Borko, 2002), for instance, provides telling examples of principals and teachers working together to make a difference for student learning. What is most startling is the "can do" attitude of the leaders and their ability to inspire their staffs. They actively support their schools as learning communities and do everything they can to develop their staff's capacity to teach to standards, including bringing their staffs together to understand what the standards mean, how students are doing relative to them, and what the implications are for action. They constantly ask the question, what should we be doing differently in teaching and learning, then try to do it, while carefully monitoring the success of their strategies.

The careful attention to progress seems to be a hallmark of a number of successful schools. For example, schools showing unusual progress in one district instituted quarterly assessments of students' reading and mathematics progress, which were aligned with state grade level standards and assessments. Teachers came together around the results and decided on the next steps for each student, particularly students at risk of not meeting the grade level standards for their classes. Special interventions were mounted and resources applied to help teachers address students' problems and apply new strategies.

We should continue to learn from success. We know from research that there is a strong relationship between socio-economic status and student performance as well as student progress. That is, schools serving more advantaged students tend to start higher and to progress faster. Moreover, recent research by KC Choi shows that within these schools, students who start lower

tend to make less progress—in effect increasing the achievement gap. We now have sophisticated methodologies that can identify schools that are beating the odds—that is, schools serving predominantly poor students who are achieving well and making strong progress, and particularly schools that are adept at accelerating the progress of their low ability students. We should find these schools and validate and share their effective practices.

There is much to learn and much to do. Even if we may have questions about current accountability systems and about whether schools can meet all the goals of NCLB, we must agree that the goals are worthy and work toward achieving them. Schools can and must become better places for children and educators. We must use the mandates of NCLB to promote true improvement.

References

Achieve, Inc. (n.d.). *Testing policy tips.* Retrieved February 2003, from http://www.achieve.org/achieve.nsf/Testing?OpenForm

Borko, H. (2002, September). *The relationship between teaching and assessment.* Symposium presented at the annual CRESST conference, University of California, Los Angeles, CA.

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 (CRESST).

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

California Department of Education Website (n.d.). *Grade four: Mathematics content standards.* Retrieved February 2003, from http://www.cde.ca.gov/standards/math/grade4.html

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L. & Mayrowetz, D. (2000).State standards, socio-fiscal context and opportunity to learn in New Jersey. {On Line}. *Educational Policy Analysis Archives 8*(35). http://epaa.asu.edu/epaav8n35

Firestone, W. A., Mayrowetz, D. & Fairman, J. (Summer, 1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *EEPA, 20*(2): 95-114.

Forum for Youth Investment (2003) *Quality Counts.volume 1, issue 1.* July/August, *2003* www.forumforyouthinvestment.org

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*(4): 257-290.

Herman, J. L. & Golan, S. (1993). Effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practices, 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*(9): 246-256.

Herman, J. L. & Perry, M. (2002, June). California Student Achievement: Multiple views of K-12 progress. Menlo Park, CA: Ed Source.

Koretz, D., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. 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 Technical Report No. 355). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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 of Measurement in Education, New Orleans, LA.

Linn, R. L. (n.d.). *Standards-based accountability: Ten suggestions* (CRESST Policy Brief). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L. & Haug, C. (2002). *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 (CRESST).

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, Center for the Study of Evaluation.

Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.D., & Williams, R.L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66*, 628-634.

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.

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

Stecher, B., & Barron, S. (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 (CRESST).

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 (CRESST).

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 (CRESST).

Stecher, B. & Borko, H. (2001). *Combining surveys and case studies to examine standards-based educational reform* (CSE Tech. Rep. 565). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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.Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.D., & Williams, R.L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66*, 628-634.