CRESST REPORT 728

Joan L. Herman

ACCOUNTABILITY AND ASSESSMENT: IS PUBLIC INTEREST IN K–12 EDUCATION BEING SERVED?

OCTOBER 2007

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Accountability and Assessment: Is Public Interest in K–12 Education Being Served?

CRESST Report 728

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

October 2007

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
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the National Research and Development Centers, PR/Award Number R305A050004, as administered by the U.S. Department of Education's Institute of Education Sciences (IES).

ACCOUNTABILITY AND ASSESSMENT:

IS PUBLIC INTEREST IN K–12 EDUCATION BEING SERVED?

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

Abstract

The reauthorization of No Child Left Behind (NCLB) makes this a good time to consider whether and how current accountability serves the public interest and whether and how it can better do so. This report explores these issues in the context of the current literature on the effects of accountability in K–12 education. It considers the meaning of "public interest" and offers a model of how public interest may be served through accountability to benefit student learning. The report considers how well the model fits available evidence by examining whether and how accountability assessment influences students' learning opportunities and the relationship between accountability and learning.

The Meaning of Public Interest

What is the public interest? While policy debates, politicians, the media and public groups often evoke it, public interest is a slippery concept to define. Reich (1988) speaks of transcendent ideas and concerns for the good of society, rather than self-interest, that motivate political action. Moyers (2007) notes that the proposition that each of us has the right to "life, liberty, and the pursuit of happiness" is the foundation of this country and that this proposition carries with it the imperative that members of society have "obligations to each other, mutually and through their government, to ensure that conditions exist enabling every person to have the opportunity for success in life." But as Hochschild and Scovronick (2004) have observed, in the context of public schooling, the proposition blends both collective and individual responsibilities, and contains inherent conflicts between policies designed for the good of ALL students and those designed to enable individuals to succeed, particularly the privileged of society.

Different perspectives on what constitutes the public interest and the policies that can promote it grow out of differing ideals and the conflicts among them, varying definitions of basic societal goals such as liberty and equality, and different analyses of the sources of problems and obstacles (Stone, 1998). What constitutes the public interest is an interaction between the facts as one sees them and one's values. For example, some see the success of Department of Defense Schools as support for integration, high academic expectations, shared decision-making and investment in professional development for educators; others see it ratifying their ideas about the importance of home culture and discipline.

Whose and how many individuals' interests need to be served? How should a policy be designed to address the public interest? These remain to be open questions. Do all or nearly all need to be served? To what end? Is an action that serves some but hurts none in the public interest? What of policies that serve the many, but hurt the few? And how many are "many" and how "few" can few be? While these may be unanswerable questions, they reflect tensions that need to be balanced in any discussion of whether and how current accountability and assessment systems may serve the public interest to benefit (or not) the education and learning of K–12 students.

Do current accountability systems serve all students? Certainly if action in the public interest means serving the needs of those who otherwise would be left unsatisfied, then to be considered in the public interest, accountability must benefit students who traditionally have been under-served—economically disadvantaged students, English learners and diverse students of color. Yet if all students are to be served, then the system also needs to benefit—or at least not hurt—students who have traditionally been higher achieving, including our highest ability students. (As we shall see, however, it is difficult to design a single test and system that well serves students at different points of the distribution.) Furthermore, consideration of public interest must address long term and unanticipated side effects as well as immediate effects. Accountability that promotes attention to the short term, bottom line of student performance must yield long-term benefits for student learning and for public education.

The concept of public interest also brings with it a basic concern with social ends and goals. If we are an accountable society—as citizens, as a body politic responsible for others—what should we and education be held accountable for in terms of student learning? Recent commissions have raised questions *again* about whether schools are sufficiently preparing students for creative thinking and problem solving and in science and technology for this country to keep its competitive edge (Partnership for 21st Century Skills, 2004; Friedman, 2005). Furthermore, in the rush to reach consensus on the meaning of proficiency in reading and mathematics, we seem to have skipped over the dialog and potential disagreements on the goals of schooling (Ramaley, 2005) as well as having settled for standards that fall short of clearly articulating the academic knowledge and skills that students will need for future success (Wilson & Berenthal, 2005). Democracy carries with it the responsibility to help create citizens who will recognize and serve the public good, not only their own interests (Parker, 2003). The public too apparently wants schools that promote self-discipline and social responsibility (Mathews, 2006). But schools currently seem

overwhelmed by the need to raise test scores and meet academic mandates, and public interest goals seem to be beyond current, official standards and expectations for schooling.

The Role of Accountability in Serving the Public Interest: A General Model

Merriam-Webster's Dictionary defines accountability as "the quality or state of being accountable; *especially*: an obligation or willingness to accept responsibility or to account for one's actions." In current educational contexts, the concept carries with it the idea that individuals, organizations and the community not only are responsible for their actions, but must also answer for their performance to an outside authority that, in turn, may impose a penalty for failure. Schools and students are responsible for teaching and meeting learning goals—no excuses, no blame game, no victimhood, and under No Child Left Behind, there are serious sanctions for districts, schools and teachers failing to meet those goals. In the simplest sense, students come to school to learn, schools and the educators within them exist to teach and to promote student learning. Since tests show which students and what schools are meeting or exceeding standards and those that are not, students and teachers who are falling short should be held accountable for their failure (and less frequently, those who succeed beyond expectations should be rewarded for their success).

While this is a basic view of bureaucratic accountability, Darling-Hammond (2006) notes the importance of professional and capacity building forms of accountability as well. At its core the broader concept of accountability contains a strong ethical and internal orientation, a concern for the welfare of others, and a commitment to efficacy. Teaching has been called a "calling" as well as an occupation, and clearly most teachers are committed to their students' learning—and get satisfaction from their own efficacy—independent of external incentives. In fact, motivation researchers long have contrasted internal and external motivation and their research suggests that external rewards reduce internal motivation (Deci & Ryan, 2000).

Leaving aside the professional accountability and intrinsic motivational issues for the moment, the role that accountability is intended to play in today's standards-based reform seems relatively straightforward and well established. All states except Iowa have established standards for what students should know and be able to do. Spurred in part by No Child Left Behind Act (P.L. 107–110, 2001), and Goals 2000: Educate America Act (P.L. 103–227, 1994) that preceded it, states have created assessments that make explicit for schools and the students within them what the standards mean. Pressured by fear of sanctions—and less often by rewards—teachers and students are motivated to teach/learn the expected standards and to use the information from the assessment to improve their efforts, even as those same

assessment results reveal who has succeeded in meeting targets or expectations, and who has not. The assessment system thus serves both technically as a performance measurement system that provides feedback and as a motivational system that serves a number of socio-political or symbolic purposes in: establishing the target for reform efforts; communicating to educators, administrators and parent what is expected; insisting on high expectations for all students; providing incentives and/or sanctions; and thereby stimulating all levels of the education system to focus on achieving the NCLB goals for adequate yearly progress (AYP), ostensibly assuring that all children will be proficient by the year 2014.

Figure 1 shows one view of how accountability is supposed to work: Accountability sets the context and creates incentives for educational action to enable all students to attain standards. State standards thus are the foundation on which the whole system sits, and the theory of action assumes that these standards establish clear and important goals for student learning.

For students to attain standards, educators must take action to improve students' learning opportunities (termed OTL in Figure 1, and also known as opportunity to learn) in what and how well students are taught in classrooms, through supplemental services and programs, and through specially targeted in- and out-of-school activities and interventions. And these improvements in OTL, in turn, are necessary precursors to improvements in students' learning, as indicated by performance on state tests and other indicators of students' progress toward or attainment of standards.

Feedback from the assessments is used to improve learning opportunities for students in terms of targeting instruction on areas of need and evaluating and refining educational programs, materials and strategies to increase students' attainment of standards. Because NCLB requires that every subgroup of students within the school attain established adequate yearly progress targets, all students must be provided with effective learning opportunities, including whatever augmented programs and special services that traditionally low achieving students (children of poverty, English learners and student with disabilities [SWD]) may need to attain success.

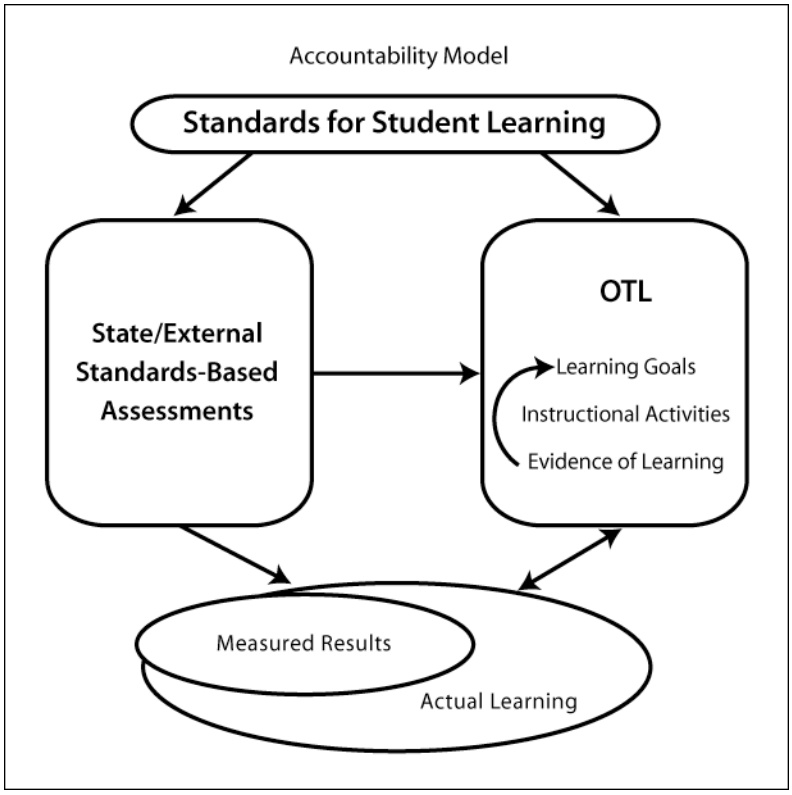

Figure 1. Accountability Model.
Note. OTL = Opportunity to learn.

Surely, however, improving student learning is not only a bureaucratic and management problem. Darling-Hammond (2006) notes the importance of professional and capacity-building forms of accountability as well. Recent research on the power of formative classroom assessment underscores these sources of accountability in that it shows not only the value of on-going assessment relative to accountability assessment but also supports the benefits of reflective, professional practice (Black & Wiliam, 1998).

While Figure 1 focuses on the impact of accountability on students' opportunity to learn, the underlying theory of action assumes that the federal government, states, districts, and schools will be accountable for assuring the that there are sufficient talent, resources, policies and practices at all levels of the educational system and that these will be coordinated and integrated to support teaching and learning, Moreover, policymakers and actors and these levels are expected to use the feedback from state assessments for management and improvement purposes: to gauge their strengths and weaknesses: to identify students, schools, and classrooms that may need special help; and to be strategic in taking action and coordinating available resources to improve student performance, e.g. through professional development, instructional materials, mentoring, and technical assistance.

This simplified theory provides a starting point for examining whether and how accountability is serving the public interest. Despite intractable problems in attributing causality and the innumerable other factors in state and local policy and practice that have an influence, I ask: Do standards provide a sound foundation for accountability? What is the evidence accountability and assessment are improving students' opportunities to learn? Is there evidence that accountability is promoting student learning and attainment of standards? For whom?

Quality of Standards

As noted above, high quality standards are the foundation on which the whole enterprise of standards-based accountability and reform rests, as standards supposedly provide the reference point for all action. Recent reviews, however, raise questions about the quality of that foundation. For example, Finn, Petrilli and Julian (2006) reviewed state standards in English-Language Arts, Mathematics, History and Science, based on their clarity in communicating what students ought to know and be able to do, their academic rigor, and their attention to the most important knowledge in the discipline. Summarizing ratings across subjects and review criteria using an A–F grading scale, the researchers report that states on average score a C-minus, although a few states were regarded as exemplary.

Similarly, the National Research Council's (NRC) review of science standards across the states (Wilson & Berenthal, 2005) found great variety in the kinds of knowledge privileged—states tended to focus on lower level declarative and procedural knowledge (define, know, describe), while some also attended to higher level schematic and strategic knowledge (predict, justify, compare, analyze, explain). NRC also found great variety in the scope of content addressed, how broadly or specifically content was defined, and found most states unrealistic in the number of learning goals that feasibly were possible to attain over the course of a year(s), highly variable in attention to the most important science content, and vague in defining performance expectations. No state's standards meet the committee's criteria of:

- Clear, detailed and complete
- Reasonable in scope
- Rigorous and scientifically accurate
- Based on sound models of learning
- Describe performance expectations and identify proficiency levels

Without a clear and realistic target to aim for, educators tend to rely on tests to define expectations. Absent rigorous, important and accurate content, standards provide a faulty foundation for assessment and instruction and can focus the educational system on trivial and superficial learning. With these caveats in mind, I turn next to evidence on the effects of accountability on students' learning opportunities.

Effects on Learning Opportunities

Relying largely on survey, interviews, and observation data from teachers and school administrators, substantial research over the last decade has shown a consistent picture of the effects of state-level accountability testing on curriculum and teaching, see for example: Arizona (Smith & Rottenberg, 1991), California (Hamilton, et al., 2007; Gross & Goertz, 2005; McDonnell & Choisser, 1997), Florida (Gross & Goertz, 2005), Georgia (Hamilton, et al., 2007) Kentucky (Koretz, Barron, Mitchell, & Stecher, 1996; Stecher, Barron, Kaganoff, & Goodwin, 1998; Borko & Elliott, 1998; Wolf & McIver, 1999), Maine (Firestone, Mayrowetz, & Fairman, 1998), Maryland (Lane, Stone, Parke, Hansen, & Cerrillo, 2000; Firestone et al., 1998; Goldberg & Rosewell, 2000), Michigan (Gross & Goertz, 2005), New Jersey (Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000), North Carolina (Gross & Goertz, 2005; McDonnell & Choisser, 1997), New York (Gross & Goertz, 2005), Pennsylvania (Hamilton, et al., 2007; Gross & Goertz, 2005), Vermont (Koretz, McCaffrey, Klein, Bell, & Stecher, 1993), and Washington (Stecher, Barron, Chun, & Ross, 2000; Borko & Stecher, 2001).

State assessments focus instruction. Research and practical experience show that 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. Research shows that almost all principals also take action to assure that teachers engage their students in direct test preparation. Teachers consistently report that state tests have a substantial effect on the content they teach and how they assess student learning.

Teachers model what is assessed. When many states developed performance assessments in the early 1990s, many classroom teachers revised their instruction and classroom assessment accordingly. Teachers in fact, scrambled to replace their own multiple-choice tests with the same types of open-ended items and/or extended writing questions that state tests had begun to use. In the middle 1990's, when states largely moved back to multiple-choice and short-answer formats, teachers' practice and assessment also reverted to

multiple choice, vocabulary lists, and the like. More recently, Hamilton, et al. (2007), in examining the effects of NCLB in California, Georgia, and Pennsylvania, found more mathematics teachers in Pennsylvania, as compared to the other two states, reporting open-ended tests in their classrooms as a result of their state assessment. The researchers attributed the difference to the use of open-ended items on Pennsylvania's math assessment. In other words, change the test and instruction follows.

Schools focus on the test rather than the standards. At least initially, educators in self-defense pay attention to what is tested and how it is tested, rather than to the underlying standards that the tests are supposed to represent. Teachers in Washington, for example, reported that their instruction tended to be more like Washington's state assessment than the Washington state standards (Stecher & Borko, 2002). When states like Washington and Kentucky tested different topics in different years, researchers found that teachers provided more time on particular subjects in the years they were tested than in those they were not (Stecher & Barron, 1999; Stecher, Barron, Chun, et al., 2000). If mathematics was tested in fifth grade but not language arts, teachers taught more mathematics and reduced instruction in other subjects such as language arts. These changes were not motivated by any coherent sense of curriculum nor were they driven by the need to continuously develop students' learning.

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 example, if extended math problems are not included on the math test, instructional time may go to computation or other problem types that are on the test. Similarly, as more time goes to the tested subjects—typically reading, language arts and mathematics—this time must come from other areas of the curriculum.

Curriculum and instruction are aligned. Across the board, districts and schools have made efforts to align curriculum and instruction with standards. This is particularly true for schools failing to meet their targets and identified as underperforming. For example, in the national *Evaluation of Title I Accountability Systems and School Improvement Efforts*, Shields et al. (2004) found that that 80% of the sampled schools were actively working to align their curricula with standards and assessment, and many were also implementing new curricula in reading/language arts and mathematics. Similarly, the Consortium for Policy Research in Education's (CPRE) study of designated, underperforming high schools uniformly shows schools concentrating on aligning curriculum and instruction with

assessments—through revisions in their regular curriculum, through the addition of new courses, test preparation and remedial and extra-school tutoring (Gross & Goertz, 2005).

More attention to assessment and student data. Shields et al. (2004) also found that 85% of schools reported using student achievement data to target their instruction, echoing other studies which also report schools using data to identify students who need special help (Center on Education Policy [CEP], 2006; Hamilton, et al., 2007). These studies highlight that districts and schools increasingly are mandating interim or benchmark assessments throughout the year to monitor student progress on expected standards and assessments. These assessments tend to mimic the content and format of state assessments and their technical quality is moot (Herman & Baker, 2005). As a result, the use of these assessments tends to encourage teachers to keep their eyes firmly on student progress, especially the knowledge and skills that will be tested, and correspondingly may heighten curriculum narrowing to focus on what is tested, rather than underlying standards.

At the same time, districts have become more prescriptive about how and what teachers are supposed to teach, have moved to common instructional materials and have created pacing guides detailing what is to be covered when—even as far as prescribing what text book pages teachers should be covering on any particular day. Interestingly, required, rigid adherence to pacing guides leaves little time for going back and re-teaching knowledge and skills which interim tests reveal as weak and can create a discouraging environment for teachers' professional practice.

Growing attention to formative assessment. Even so, formative assessment, the use of classroom assessment to inform ongoing teaching (Wiliam, 2006) shows growing popularity. Black and Wiliam's (1998) landmark review showed the potential power of formative assessment, and educators have increasingly recognized that they need ongoing information about student learning if they are to be accountable for results. Yet available evidence suggests that the rhetoric surpasses the reality of formative assessment use: all teachers are increasingly talking the talk (Herman, Yamashiro, Lefkowitz, & Trusela, 2001; Hamilton et al., 2007) but the studies looking at practice closer up suggest the challenges of developing teachers' capacities to engage in valid, formative practice (Gearhart et al., 2006).

At-risk students face curricular distortions. There also is growing evidence that curriculum options are grossly narrowing for low scoring students and underperforming schools. In the context of No Child Left Behind's requirements for annual yearly progress, schools are increasingly focusing on reading and mathematics, to the exclusion of science, social studies, and the arts; and at the secondary level, low performing students are being

pulled from academic courses to concentrate on literacy development (Gross and Goertz, 2005; Greenleaf, Jimenez, & Roller, 2002; Mintrop & Trujillo, 2007). Indeed one recent national study shows that 71% of school districts indicated that they have reduced instructional time in at least one other subject to make more time for reading and mathematics, and in some districts, struggling students receive double periods of reading and/or math, missing electives or other subjects (CEP, 2006). Moreover, there is anecdotal evidence that literacy intervention programs for the lowest performing students are devoid of actual book reading. Instead, students uniformly read excerpts and seek out answers to the information-type questions that are expected to be on the state test. At the extreme, there are anecdotes as well about schools using "triage" strategies to focus on what they consider "pushables" and "slippables" (relative to reaching proficiency) and virtually ignoring both students in greatest educational need and overriding issues of improving instructional quality (Booher-Jennings, 2006). Such distortions provide counter evidence to the claim that current accountability is improving instruction for low performing students and are worrisome as well in the context of a Gates Foundation survey indicating that students do not drop out of school primarily because they cannot do the work or pass their courses, but more often because they are bored by school (Bridgeland, DiIulio, & Morison, 2006).

So the rhetoric and dominant stories are changing, as are some aspects of practice. Across school levels and types of schools, and regardless of the specifics or strength of states' accountability systems or the intensity of their incentives or sanctions, research suggests that accountability testing does serve to motivate attention and action and that the action so motivated serves to change the alignment of curriculum and instruction with standards and assessment and to change students' opportunities to learn. In some cases, changes in curriculum are school wide, and in other cases they are specialized courses or services for students identified as at risk.

But do these changes in instruction and opportunities to learn represent real *improvements* that actually benefit learning for students, and particularly students who are most at risk, or do they really impoverish learning opportunities, as critics have charged, relegating students to a narrow curriculum of test preparation that is devoid of complex thinking and problem solving and devoid of learning in the arts and sciences, as previous evidence suggests (National Research Council, 2003; Pelligrino, 2006)? Koretz (2005) has conceptualized a number of ways in which schools and teachers respond to the alignment challenge, ways which differ dramatically in terms of their potential to improve student learning (in contrast to inflating their test scores): from changes in the allocation of time (do more of the same), to meaningful alignment of instruction (do something different in

curriculum and instruction), to substantive and non-substantive coaching or test preparation, to cheating. Available evidence shows more attention to changes in the allocation of time and attention to test preparation than to changes in the quality or effectiveness of instruction.

Ultimately, the proof of the pudding in whether accountability actually improves students' opportunities to learn may lie in student performance. That is, if learning opportunities are improving, should not such improvements be reflected in student performance? We turn now to this body of evidence, including studies conducted prior to 2001 and prior to NCLB when there was more diversity in state accountability systems, and post 2001.

Effects on Performance Prior 2001

Several studies have used data prior to NCLB from the National Assessment of Educational Progress (NAEP) to study the effects of accountability on student performance. Generally their results have been positive. For example, Grissmer, Flanagan, Kawata and Williamson (2000) hypothesized that accountability reforms might be responsible for the rapid growth, relative to other states, found for North Carolina and Texas for the period 1990–1996. Similarly, Carnoy and Loeb (2004) observed a relationship between the strength of a state's accountability system (its consequences for schools) and gains in the percentage of students scoring at least "basic" on NAEP mathematics assessments 1996–2000, but saw no relationship in retention or survival rates. A similar relationship was observed in percentage of White and African American students scoring at least "proficient" at eighth grade. Fourth grade effects were less clear in that only African American students showed significant gains at the basic level. Using slightly different methodologies and different strategies for dealing with changes in exclusion rates, Braun (2004), Hanushek and Raymond (2004), and Rosenshine (2003) came to similar conclusions about the relationship between NAEP gains and high stakes testing systems: results favored high stakes versus no stakes states.

Student accountability: Ending social promotion. In terms of effects of consequences for students, research on the Chicago Public Schools "end to social promotion" policies represents one of the most extended and thorough studies. The authors were careful to note that the social promotion reform occurred simultaneously with a new accountability program for the lowest achieving schools in the district and that the effects of the two programs could not be disentagled. Overall, Roderick, Nagaoka, and Allensworth (2005) concluded that the 1996 reforms were related to improvements in middle grades performance in the middle grades that extended into high school. However, there were no effects in the

early grades—perhaps an area where student motivation would not otherwise be problematic. The details of their findings showed that some students near the cut-off worked harder and escaped retention, so the threat of retention helped them. More problematic, however, the study found that low achieving students who were retained because of the reform did not benefit educationally during the retained year, experienced lower achievement gains in the sixth grade than students with similar test scores who were promoted, and based on existing research on retention, were at increased risk of dropping out. In short, the most vulnerable students did not benefit from this student-level accountability reform.

Student accountability: High school exit exams. Today's high school exit exams revisit some of these same patterns and echo issues that emerged in response to the minimum competency exams of the late 1970s and 1980s. Then, as now, results have shown initial high failure rates that decline over time, with large disparities in performance for poor and minority students, students with disabilities and English learners (CEP, 2005, 2006; Heubert, 2004). California represents a current example: one year before the graduation test requirement went into effect, an estimated 78% of the class of 2006 had passed the state's High School Exit examination, leaving nearly 100,000 who had not. By the time of graduation in June 2006, an estimated 40,173 students still did not meet the requirement (California Department of Education [CDE], July 2006).

As with the effects of retention, a number of studies have suggested a troubling relationship between high school exit exams (or their precursors) and students' dropping out of school. For example, Catterall (1989) early found that students who had failed to pass a minimum competency test on their first try, relative to their similar ability peers, were more likely to doubt their chances of graduating and to report the possibility of dropping out. Subsequently, researchers examining patterns of performance by state found that high school enrollment and completion rates generally were lower for economically disadvantaged and/or low ability students in states that had such tests compared to states without such tests (Reardon 1996; Bishop, Mane, Moriarty and Bishop 2001, Bishop and Mane 2004). At the same time, studies found positive effects on subsequent educational success: eighth graders in states with high school exit exams were more likely to go to college and equally likely to graduate from college, and controlling for high school graduation, likely to get higher-paying jobs than their peers in other states (Bishop, Mane, and Moriarty, 2001; Bishop and Mane 2005).

Different effects for different kinds of tests. Moreover, there also is evidence that the effects of high school exit exams may be different for different types of tests. Bishop (2005) shared empirical data from a variety of sources to argue that rigorous, course-based exit

examinations, such as those used Europe and introduced in North Carolina and New York, benefit student achievement (National Assessment of Educational Progress [NAEP], 2005) substantially more so than more typical minimum competency tests, and these positive effects are achieved without any increase in drop out rates. Similarly, Darling-Hammond, Rustique-Forrester, and Pecheone (2005) used evidence from NAEP to show that while states that use a single exit exam for high graduation show higher dropout rates, particularly for African American students, Latino students, students with disabilities, and English learners, those states that use a multiple measures approach and consider a variety of student work in making graduation decisions have tended to maintain high achievement and high graduation rates. The nature of the accountability and assessment system apparently matters, and there are existence proofs for engendering higher performance without the unacceptable costs associated with higher dropout rates.

Performance Effects Since 2001

What of more recent effects of accountability? *Education Week* released 10 years of *Quality Counts* data in 2006, which has monitored state progress in adopting core elements of standards-based reform, including establishing academic standards, aligning assessment with those standards, implementing accountability measures, and providing supports for improving teacher quality. *Quality Counts* indicators show increases since 1997 in the implementation of policies in all of these areas across states, although the trajectories of individual states vary considerably (Swanson, 2006).

National trends on NAEP document a similarly positive trajectory over a similar time period, showing definite if not modest increases from 2000 to 2005 in mathematics at Grades 4 and 8 and some improvement for Grade 4 reading; Grade 8 reading, however, shows a decline. Furthermore, while there has been some reduction in the achievement gap during the 10-year period, substantial differences persist, and there has been little or no reduction since NCLB (Lee, 2006).

Looking at the relationship between states' changes in standards-based policy implementation and their progress on NAEP, *Quality Counts* shows a consistently positive—though again, modest—relationship for policies related to academic content standards, aligned assessment, and accountability measures, particularly for mathematics. Oddly, however, the implementation of policies related to teacher quality negatively correlated with performance.

In an effort to take a closer look at what assessment and accountability system characteristics might be related to state performance, we tried to use available data to validate

existing quality indicators and to identify states that were over- and under-performing based on NAEP results. While recognizing the multitude of variables and system levels that could influence student learning, we speculated that if accountability was supposed to be a strong intervention, the quality of standards and assessments and the nature of the accountability system might make a difference for student performance. States that communicated strong expectations with clear standards and backed them up with rigorous assessments of high technical quality that included multiple measures and in turn could provide accurate feedback, we thought, might do better than states whose systems lacked these features. At the same time, we thought that we might be able to find differences in the assessment and accountability systems of states that showed exceptional performance on NAEP, compared to those that did not. Beyond the obvious caveats related to any analysis, we encountered conundrums in exploring with avenue.

What states are achieving well? A first challenge was the identification of over- and under-performing states, using NAEP as the common, comparable measure across states. We reasoned that such classifications should be based on both the status of student performance and progress in student performance. We thus identified states whose performance, controlling for socioeconomic status (SES), was better than expected across the two recent NAEP administrations, 2001 and 2003, in reading and/or mathematics across both Grades 4 and 8 (regression analysis using concentrations of economically disadvantaged students and state NAEP results, (drawn from School Matters, 2005). Massachusetts and New York stand out uniquely from this analysis as outperforming states with similar SES in reading at Grades 4 and 8 and mathematics at Grade 8—that is, for three of the four NAEP assessments. South Carolina, Kansas and Minnesota show better than expected performance in mathematics, for both Grades 4 and 8, and Kentucky for reading, in Grades 4 and 8.

However, in moving to the identification of states whose improvements in performance were outstanding relative to other states, defined as states showing at least 1 standard deviation above the mean state gain from 2001 to 2003, the consistent performers generally are different. Massachusetts was the exception, showing exceptional improvement[1] for three of four possible assessments—reading at Grade 4 and mathematics at Grades 4 and 8. Six additional states achieve better than average on three of four assessments. Of these, Pennsylvania and Washington show consistent improvement in reading (i.e., at both Grades 4

[1] Exceptional improvement was defined as a *z* score equal to or greater than 1.0, compared to the 50 state, plus DC sample.

and 8); and Arkansas, New Jersey, Ohio, and Texas in math. Idaho shows consistent improvement at Grade 4 in both reading and mathematics.

Using the Trial Urban District Assessment as another proxy for how traditionally lower achieving, poor and minority students are doing, results show greater improvement in mathematics than in reading, with 8 of the 10 participating districts showing a statistically significant increase, ranging from scale points 4–9, from 2003 to 2005 in fourth grade mathematics (NAEP, 2005). Boston, Houston, Los Angeles and San Diego also showed a significant increase in 8th-grade mathematics, with Los Angeles showing significant increases across all four assessments. Mirroring weaker performance trends in reading nationally, only Atlanta showed consistent increases from 2002 to 2003 and 2005 in reading, while New York showed consistent increases at Grade 4. It is interesting to note that the two districts that performed highest relative to their peers, Charlotte and Austin, are the only two that did not show any significant increase over the period. Because both Charlotte and Austin are in states that early on implemented strong accountability systems, one might suspect that this lack of improvement may show a topping out of what can be accomplished with traditional impacts of accountability, without dramatic changes in teaching and learning practices, but then what explains Houston? And as noted above, Boston has been operating under Massachusetts' longstanding and stable accountability system.

The two sets of NAEP analyses—status relative to SES and improvement in scores—uniquely identify Massachusetts as a high performer but their state assessment results for the same period show more modest improvement relative to other states, and as one tries to compare state assessment and NAEP results in other states, the patterns (or lack thereof) are puzzling. Moreover, Quality Counts identifies Massachusetts as no. 49 in terms of the achievement gap on NAEP between students who do and do not qualify for the federal free lunch program, even as results also show that the state is making progress in closing the gap. It is also interesting to note that Massachusetts started in 1997 as the highest amongst the 50 states in their implementation of standards-based reform and has maintained its position over the years. Could it be that consistency in policy is a contributing factor, even as we know that relative wealth and early childhood indicators among many other variables also contribute?

Furthermore, whereas there is limited consistency in what states are identified as high performers on NAEP across SES and improvement analyses, there is more consistency in the under-performers. However, as we see in the next section, it was difficult to differentiate these states based on features or qualities of their assessment or accountability systems.

Differentiating system characteristics using available indicators. If we believe that assessment and accountability ought to have benefits for student learning, then it stands to reason that the quality of the assessment and accountability system ought to matter. However, it is difficult to get a handle on quality, given the depth and validity of existing indicators. *Quality Counts* data show more surface similarities than deep differences in current state systems. For example, virtually all states combine multiple-choice testing with an extended assessment of language arts (writing). Two thirds of the states also include open-ended items on their assessments; and half of the states include extended responses in subjects in addition to language arts. Most states claim they have developed customized tests relative to their standards, and almost all claim that they have done alignment studies. Of course, these are required under NCLB. But evidence from these alignment studies shows uneven quality. For some states these show major imbalances between standards and assessment and across states generally reveal a disproportionate representation of lower level skills relative to thinking and problem solving (e.g., Webb, 1999).

Studies commissioned by the Fordham Foundation show a more varied picture of the quality of standards and assessment systems across states (Cross, Rebarber, Torres, & Finn, 2004; Finn et al., 2006; Klein et al., 2005; Stotsky & Finn, 2005) and, as they use somewhat different criteria, it perhaps is not surprising that we found only modest relationships between the Fordham and *Quality Counts* ratings (.49 for ratings of standards in mathematics and .39 for ratings in English, based on analyses of states standards in 2005). As a source of convergent validity evidence for existing indicators, the evidence then is scanty.

To determine the quality of state tests, the Fordham study reviewed available documents and technical reports rather than relying on survey data, which was the source for *Quality Counts*. Rating state assessments in terms of their content, alignment of standards and assessment, academic rigor, and technical trustworthiness, the Fordham study found significant room for improvement, starting with the availability of materials from which state tests could be described and evaluated (Cross et al., 2004). However, three states were distinguished in receiving high marks in three of six categories: Massachusetts, Pennsylvania, and Virginia.

Effects on Special Populations

We have noted above glimmerings that accountability is having an impact on low SES students—nationally and within states, students who qualify for free lunch and in general students served by large urban school districts. What of other special populations? The research base examining effects on students with disabilities and on English language learner

students is scanty. What we do know is that the validity and comparability of the state assessment results for these groups is suspect and thus it is hard to get a handle on the status and progress of their performance. The logic of using state assessment results to support the improvement of learning for these groups also is weak, given that the assessment results may well lack integrity. Nonetheless, advocates for these groups generally see the inclusion of special populations in state accountability systems as a plus, because it has made the educational needs of these students visible and mandated expectations and plans for progress in mainstream contexts that were too often without them, even if the prior contexts were deeply "caring." At the same time proponents worry about accountability targets that are unrealistic and fear backlash when the performance of EL and/or SWD subgroups deter schools and districts from meeting AYP targets.

What of the impact of accountability on other segments of the student population—traditionally higher performing students? On the gifted? The average student? From a measurement perspective, we know that it is difficult for a single test of limited duration to differentiate and/or motivate students at all points on the achievement spectrum. High ability students may be engaged with advanced placement (AP) and college entry exams, as well as in honors classes and gifted programs, which serves to motivate attention to their learning needs. But there is no obvious accountability mechanism for the "average student, " who may have made it just over the proficient level. There is little research on this issue, but one might speculate that current federal accountability requirements need to do more to spur attention to the learning of "average" students, who represent the majority of students.

Effects on Teachers

While a thorough treatment of the effects on teachers is also beyond the scope of this report, it is worth noting a growing literature that is cause for concern. Research shows the strong relationship between student learning and the quality of teachers (Carey, 2004; Haycock, 1998; Sanders & Rivers, 1996) and the quality of interactions between teachers and students. Yet any number of survey studies have suggested that teachers believe that current accountability models are causing schools to focus too much on state tests and not pushing schools in educationally productive directions, which includes concerns that have been raised earlier about curricular distortion, neglect of complex thinking, and a focus on test format and test preparation rather than on effective pedagogy etc. (Hoffman, Assaf, & Paris, 2001; Jones & Egley 2004; Pedulla et al., 2003). These studies similarly raise questions about whether accountability is increasing student learning (rather than simply inflating scores on state tests) and about the potential negative effects of accountability on teacher morale and motivation.

Concerns about accountability effects on teacher morale and motivation in the current NCLB context are bolstered by both theory and empirical evidence. Expectancy theory (Vroom, 1964) suggests that motivation is a function of one's perceived probability of success (expectancy), connection of success and reward (instrumentality) and value of obtaining goals (valence). In other words, people are motivated by things that are desirable, that they know how to do and that they feel capable of achieving. Yet much been written about the feasibility of even the most effective schools achieving NCLB annual yearly progress goals and closing the achievement gap (Linn, 2003; Rothstein, 2006). Research further shows that schools serving low performing students and students of color are least likely to be able to achieve these goals and have been the first to be subject to increasingly severe sanctions (Kim & Sunderman, 2005). Expectation theory would anticipate serious negative effects on motivation (and subsequent retention) in these settings where there is a low expectation of success. Indeed, the theory is supported by empirical evidence from a North Carolina study documenting the negative effects of strong accountability systems of low performing schools' ability to retain teachers in general, and quality teachers in particular (Clotfelter, Ladd, Vigdor, & Diaz, 2004).

If ultimately it is professional accountability—teachers' day-to-day commitment to effective practice and their ongoing motivation and sense of pedagogical responsibility—that is most important in advancing student learning, then one must question the relationship between professional accountability and unrealistic bureaucratic accountability requirements.

Summary and Conclusions

So returning to the question of whether accountability is serving the public interest: Trite but true, the answer is complicated. Available evidence suggests that the theory of action underlying accountability is generally working.

Support for theory of action. Accountability systems make public expectations and motivate educators and students to pay attention to learning and performance: Schools are changing what they are doing, they are focusing on teaching and learning and aligning curriculum and instruction with standards—or at least those that are tested. They are working to better use data to refine their programs and to identify students who are falling behind. Districts and schools are trying to expand available opportunities so that students will get the extra help they need to catch up—or at least be proficient in reading and mathematics tests. Administrators and teachers are paying better attention to and making plans to respond to and engage *all* their students, particularly traditionally low achieving, identifiable subgroups. Moreover, as we look to NAEP results as an external indicator of performance effects, we

find a modest relationship between strength and duration of accountability systems and improvement in student performance, and we find small improvements in NAEP performance for economically poor students and perhaps some small movement in closing of the achievement gap.

Danger spots. Admittedly these effects are quite modest compared to the challenge of helping all children reach their potential, and it must be acknowledged that there are dangers here for our most vulnerable children—for example, those neediest students who got left behind in Chicago's promotion program and who were at greater risk of dropping out; students who do not pass high school exit exams required for diplomas and are more likely to drop out; the lowest ability students who may be ignored as schools work to move students closer to the proficiency level over the line.

It is clear from the research that accountability is changing what gets taught, but whether the change represents real improvement in students' opportunity to learn is moot. Research suggests a narrowing of the curriculum to focus on what is tested on state assessments, and what gets tested as well as what is included in standards, tends to over-represent lower level skills and give scant attention to higher level thinking and complex applications. In responding to current accountability systems, then, teachers may be less likely to engage students—particularly low performing and minority children—in meaningful problem-solving and reasoning activities, or to build the skills that students will need for success in the 21st century. Teachers echo this concern, believing that accountability is moving education in the wrong direction. At the same time, unrealistic accountability targets may be discouraging the best teachers from teaching in schools with high proportions of academically needy students. The potential combination of meager curriculum and lesser quality teachers in the long run could increase rather than decrease the real achievement gap.

Yet even against these dangers, perhaps modest positive effects for most students should be viewed as an important accomplishment, even as we work to make the system work better for the most vulnerable and work to guard against unintended curriculum effects. It is sad but true that expecting something of all students, even if it is only what is tested, may in fact be an improvement for some. We can and should do better.

Toward better accountability systems for the public interest. Accountability systems are most apt to serve the public interest when they are designed to maximize benefits and minimize negative effects. To maximize the benefits, this report suggests the importance of assuring that:

- Standards clearly communicate realistic expectations and represent the knowledge and skills that students will need for future success. One of the reasons that schools may teach to the test is that it is the only concrete guidance they have about learning expectations. Standards must provide a solid foundation for assessment, instruction and accountability systems and focus on meaningful learning.
- Accountability systems reflect the full depth and breadth of standards and encourage good educational practice. Clearly, on demand, annual state tests of limited duration cannot measure all that is important for students to know and be able to do. Measurement theory and policy analysis suggests the value of multiple measures (AERA, APA, NCME, 1999; Darling-Hammond et al., 2005). Assessments or simply accountability requirements that students be engaged in meaningful problem-solving and reasoning tasks could help to ameliorate current imbalances. Rather than simply mimicking state tests, benchmark assessments could be used to expand the depth and breadth of standards coverage and could be embedded in meaningful curriculum activities.
- Performance expectations are suitably high yet attainable. Grossly unrealistic performance expectations that carry sanctions are counterproductive to good teaching and learning. They discourage teacher motivation and encourage testing to the test. Bob Linn (2003) has suggested using the trajectories of the fastest improving schools as a starting point for setting reasonable targets. Accountability models that credit schools for improvement in student performance at each levels of the proficiency continuum, (e.g., from *below basic* to *basic*, from *proficient* to *advanced*), could help assure that schools do not ignore their lowest achieving or average students.

Yet even with the most optimal system, there are limits to what accountability alone can accomplish. Accountability systems can provide motivation, evidence, and a target for action, but effective action depends on educators' capacity. If educators already knew how to respond to the needs of their most challenging or all their students, they would be doing so. Available evidence cited here suggests that educators are trying, but without dramatic success. Without continued investment in capacity building and resources to improve teaching and learning, there can be little closing of the achievement gap.

Even so, there is only so much that public schools can do to close an achievement gap that grows out of greater social and historical inequities. As Richard Rothstein (2006, p. 1) has observed:

> If as a society we choose to preserve big social class differences, we must necessarily also accept substantial gaps between the achievement of lower-class and middle-class children. Closing those gaps requires not only better schools, although those are certainly needed, but also reform in the social and economic institutions that prepare children to learn in different ways. It will not be cheap.

So, is accountability serving the public interest? I say yes—although it is not a resounding success and clearly we can do better. We need more safeguards in the system to guard against the potentially deadening effects of accountability and to stimulate the empowerment and efficacy it can bring when we as educators make a difference. We need to continue to ask: Are our systems in the public interest? For whom are they working, for whom not? How do we know? How can we optimize? Research and development (R&D) and capacity building must continue. But at the same time, we need to be honest about what accountability systems can and cannot accomplishing in helping all children to succeed.

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bishop, J. (2005). High school exit exams: When do learning effects generalize? In J. L. Herman, & E. Haertel (Eds.), Uses and misuses of data for educational accountability and improvement. *Yearbook of the National Society for the Study of Education, 104(2),* 99-118. Malden, MA: Blackwell Publishing.

Bishop, J. H., & Mane, F. (2004). The impacts of career-technical education on high school labor market success. *Economics of Education Review. 23*(4), 381-402.

Bishop, J. H., & Mane, F. (2005). Raising academic standards and vocational concentrators: are they better off or worse off? *Education Economics, 13*(2), 171-187.

Bishop, J. H., Mane, F., & Moriarty, J. Y. (2001). Diplomas for learning, not seat time: The impacts of New York regents examinations. *Economics of Education Review*, *19*(4), 333-349.

Black, P., & Wiliam, D. (1998, March). Assessment and classroom learning. *Assessment in Education, 4,* 7-74.

Booher-Jennings, J. (2006). Rationing education in an era of accountability. *Phi Delta Kappen, 87*(10), 756-761.

Borko, H., & Elliott, R. (1998). *Tensions between competing pedagogical and accountability commitments for exemplary teachers of mathematics in Kentucky* (CRESST 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 a symposium at the annual meeting of the American Educational Research Association, Seattle, WA.

Braun, H. (2004). Reconsidering the impact of high stakes testing. *Educational Policy Analysis Archives. 112*(1). Retrieved January 5, 2004, from http://epaa.asu.edu/epaa/v12n1/v12n1.pdf

Bridgeland, J., DiIulio, J., & Morison, K. (2006). *The silent epidemic.* Seattle, WA: Gates Foundation.

California Department of Education (CDE), (July, 2006). *California high school exit exam reports.* Retrieved July 21, 2006, from http://www.cde.ca.gov/nr/ne/yr06/yr06rel82.asp

Carey, K (2004). The real value of teachers. Washington, DC: Education *The Trust, 8*(1), 3-32.

Carnoy, M., & Loeb, S. (2004). Does external accountability affect student outcomes? A cross-state analysis. In S. H. Fuhrman, & R. F. Elmore (Eds.). *Redesigning accountability systems for education.* New York: Teachers College Press.

Catterall, J. S. (1989). Standards and school dropouts: A national study of tests required for high school graduation. *American Journal of Education, 98,* 1-34.

Center on Education Policy (2005). *State high school exit exams: States try harder but gaps persist.* Washington, DC: Author.

Center on Education Policy (2006). *Year 4 of the No Child Left Behind Act.* Washington, DC: Author.

Clotfelter, C., Ladd, H., Vigdor, J., & Diaz, R. (2004, Spring). Do school accountability systems make it more difficult for low performing schools to attract and retain high quality teachers? *Journal of Policy Analysis and Management 23,* 251-271.

Cross, R., Rebarber, T., Torres, J., & Finn, C. (2004). *Grading the system: The guide to state standards, tests and accountability policies.* Washington, DC: Thomas B. Fordham Foundation.

Darling-Hammond, L., (2006). Securing the right to learn: Policy and practice for powerful teaching and learning. *Educational Researcher. 35*(7), 13-24.

Darling-Hammond, L., Rustique-Forrester, E., & Pecheone, R. (2005). *Multiple measures approaches to high school graduation*. Palo Alto, CA: School Redesign Network at Stanford University.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11,* 227-268.

Finn, C., Petrilli, M., & Julian, L. (2006). *The state of state standards.* Washington, DC: Fordham Foundation.

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000). State standards, socio-fiscal context and opportunity to learn in New Jersey. *Educational Policy Analysis Archives, 8*(35). Retrieved July 26, 2000, from http://epaa.asu.edu/epaa/v8n35

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

Friedman, T. (2005). *The world is flat.* New York: Farrar, Straus & Giroux.

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, S., Vendlinski, T., et al. (2006). Developing expertise with classroom assessment in K–12 science. *Educational Assessment 11*(3&4), 237-263.

Goals 2000: Educate America Act*,* Pub. L. No. 103-227, 108 Stat. 125 (1994).

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.

Greenleaf, C. L., Jimenez, R. T., & Roller, C. M. (2002). Conversations: Reclaiming secondary reading interventions: From limited to rich conceptions, from narrow to broad conversations. *Reading Research Quarterly, 37*(4), 484-496.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us* (RAND Publication No. MR-924-EDU). Santa Monica, CA: RAND.

Gross, B., & Goertz, M. E., (2005). *Holding high hopes: How high schools respond to state accountability policies* (CPRE Research Report Series, RR-056). Consortium for Policy Research in Education.

Hamilton, S., Stecher, B., Marsh, J. A., Sloan McCombs, J., Robyn, A., Russell, J., et al. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states.* Santa Monica, CA: RAND.

Hanushek, E., & Raymond, M. (2004). *Does school accountability lead to improved student performance?* (NBER Working Paper 59). Cambridge, MA: National Bureau of Economic Research.

Haycock, K. (1998). *Good teaching matters.* Washington, DC: Education Trust.

Herman, J. L., Yamashiro, K., Lefkowitz, S., & Trusela, L. (in press). *Exploring data use and school performance in urban public schools.* (CRESST Tech. Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., & Baker, E. L. (2005, November). Making benchmark testing work. *Educational Leadership, 63*(3), 48-54.

Heubert, J. P. (2004) High stakes testing in a changing environment: Disparate impact, opportunity to learn and current legal protections. In S. Fuhrman & E. Elmore, (Eds.), *Redesigning accountability systems for education.* New York: Teachers College Press.

Hochschild, J., & Scovronick, N. (2004). *The American dream and the public schools.* Oxford: Oxford University Press.

Hoffman, J. V., Assaf, L. C., & Paris, S. G. (2001). High-stakes testing in reading: Today in Texas, tomorrow? *The Reading Teacher, 54,* 482-492.

Jones, B. D., & Egley, R. J. (2004). Voices from the frontlines: Teachers' perceptions of high- stakes testing. *Educational Policy Analysis Archives*. Retrieved August 9, 2004, from http://epaa.asu.edu/epaa/v12n39

Kim, J., & Sunderman, G. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*(8), 3-13.

Klein, D., Braams, B. J., Parker, T., Quirk, W., Schmid, W., & Wilson, W. S. (2005). *The state of math standards 2005.* Washington, DC: Thomas B. Fordham Foundation.

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. M. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (CRESST Tech. Rep. 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.

Lee, J. (2006). Tracking achievement gaps and assessing the impact of No Child Left Behind on the gaps: An in-depth look into national and state reading and math outcome trends. Cambridge, MA: Harvard Civil Rights Project.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3-13.

Mathews, D. (2006). *Reclaiming public education by reclaiming our democracy.* New York: Kettering Foundation Press.

McDonnell, L. M., & Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments* (CRESST Tech. Rep. No. 355). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Merriam-Webster's collegiate dictionary (10th ed.). (1993). Springfield, MA: Merriam-Webster

Mintrop, H., & Trujillo, T. (2007). *The practical relevance of accountability systems for school improvement: A descriptive analysis of California schools* (CRESST Tech. Rep. No. 713). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Moyers, B. (2007, January 22) *For America's sake. The Nation.* Retrieved January 22, 2007, from http://www.thenation.com/doc/20070122/moyers

National Assessment of Educational Progress (2005). *The nation's report card: Trial urban district assessment. Mathematics 2005.* Washington, DC: Author. Retrieved February 1, 2006, from http://nces.ed.gov/nationsreportcard/pdf/dst2005/2006457r.pdf

National Research Council (2003). *Assessment in support of learning.* Washington, DC: National Academies Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Parker, W. C. (2003). *Teaching democracy: Unity and diversity in public life.* New York: Teachers College Press.

Partnership for 21st Century Learning (2004). *Framework for 21st century learning.* Retrieved July 27, 2007, from http://www.21stcenturyskills.org/index.php?option=com_content&task=view&id=254&Itemid=120

Pedulla, J., Abrams, G., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state mandated testing programs on teaching and learning.* Boston: National Board on Educational Testing and Public Policy, Boston College.

Pelligrino, J. (2006) *Rethinking and redesigning curriculum, instruction, and assessment: What contemporary research and theory suggest.* Paper commissions by the National Center on Education and the Economy.

Ramaley, J. D., 2005. Goals for Learning and assessment. In J. L. Herman, & E. Haertel (Eds.), Uses and misuses of data in accountability testing. *Yearbook of the National Society for the Study of Education, 104*(2), 55-77. Malden, MA: Blackwell Publishing.

Reardon, S. F. (1996, April 8-12). *Eighth grade minimum competency testing and early high school dropout patterns.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Reich, R. (1988). *Introduction in the power of public ideas,* Cambridge, MA: Ballinger.

Roderick, M., Nagaoka, J., & Allensworth, E. (2005). Is the glass half full or mostly empty? Ending social promotion in Chicago. In J. L. Herman, & E. H. Haertel (Eds.), Uses and misuses of data in accountability testing. *Yearbook* of the *National Society for the Study of Education, 104*(2), 223-259. Malden, MA: Blackwell Publishing.

Rosenshine, B. (2003). High stakes testing: Another analysis. *Educational Policy Analysis Archives, 11*(24). Retrieved August 4, 2003, from http://epaa.asu.edu/epaa/v11n24

Rothstein, R. (2006). Reforms that could help narrow the achievement gap. *Policy Perspectives.* San Francisco, CA: WESTED.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement. Research progress report*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

School Matters (2005). *Leveling the playing field. Identifying outperforming and underperforming states on the NAEP in demographic context.* Retrieved from http://www.schoolmatters.com/pdf/naep_comparative_state_performance_2005_schoolmatters.pdf

Shields, P., Esch, C., Lash, A., Padilla, C., Woodworth, K., LaGuardia, K., et al. (2004). *Evaluation of Title I accountability systems and school improvement: First year findings.* Washington, DC: US Department of Education.

Smith, M. L., & Rottenberg, C. (1991). 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* (CRESST Tech. Rep. No. 505). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

Stecher, B. M., Barron, S. L., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-1997 RAND survey of Kentucky teachers of mathematics and writing* (CRESST Tech. Rep. No. 482). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stone, D., (1988). *Policy Paradox and Political Reason*. Glenview, IL: Scott, Foresman and Little, Brown.

Stotsky, S., & Finn, C. E. (2005). *The state of English standards 2005.* Washington, DC: The Thomas Fordham Foundation.

Swanson, C. B. (2006, August). *Making the connection: A decade of standards-based reform and achievement.* Editorial Projects in Education Research Center. Retrieved August 1, 2006, from http://www.edweek.org/ew/toc/2006/01/05/

Vroom, V. H. (1964). *Work and motivation.* New York: John Wiley & Sons.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessment in four states* (NISE Research Monograph No.18). Madison, WI: University of Wisconsin–Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment, 11*(3&4). 283-287.

Wilson, M., & Berenthal, M. (2005). *Systems for state science assessment.* Washington, DC: National Academy.

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.