Accountability Policies and Measures

What We Know and What We Need

Susan M. Brookhart

National Education Association
Research Department
Ronald D. Henderson, Director

NEA Research
1201 16th Street, N.W.
Washington, D.C. 20036-3290
www.nea.org

89358 03/09 hls

Accountability Policies and Measures

What We Know and What We Need

Susan M. Brookhart

National Education Association
Research Department
Ronald D. Henderson, Director

The National Education Association is the nation's largest professional employee organization, representing 3.2 million elementary and secondary teachers, higher education faculty, education support professionals, school administrators, retired educators, and students preparing to become teachers.

Copies of this publication may be purchased from the NEA Professional Library Distribution Center, P.O. Box 404846, Atlanta, GA, 30384-4846. Telephone 1-800-229-4200 for price information or go to the NEA Professional Library Web site at http://www.nea.org/books.

The issue of accountability will no doubt play a prominent role in the debate shaping the reauthorization of the Elementary and Secondary Education Act.

At the National Education Association, we feel the need to reframe the old arguments about accountability by posing critical questions such as:

- What do we know from prior research and practice about accountability mechanisms and systems, especially as they relate to student achievement and narrowing the achievement gaps?
- What might an ideal accountability system look like?
- Are there ways to examine what we have learned over the past eight years and apply those lessons in a manner that supports student and teacher learning?

To that end, NEA commissioned a review of the research literature on accountability, particularly the way in which accountability systems have served their intended purpose as a mechanism for improving student achievement. This paper provides a substantive basis for discussing how well accountability systems are meeting this target and offers some alternative ways of thinking about accountability that might move us closer to the goal of enhancing student learning.

We hope this review is useful for revisiting ideas and generating new thoughts about accountability and learning. We also hope that this new approach to accountability will influence and inform the ongoing dialogue over new ESEA legislation—and ultimately help us ensure a great public school for every student.

John Wilson, Executive Director

Dennis Van Roekel, President

Contents

Executive Summary

This publication reviews the current literature on accountability systems. It reviews the research literature on how current accountability policies and measures affect students, teachers, and student achievement. It also reviews certain aspects of accountability policies and measures that result in criteria from which effective accountability systems may be built. It reviews alternative accountability policies and measures, and it provides conclusions and recommendations about future development of effective accountability systems.

The research shows that accountability is associated with gains in student achievement, at least in mathematics. This association is most apparent when the appropriate inputs and support are made available to students, teachers, and schools. External accountability appears more successful when both organizational capacity and a strong internal accountability are present.

While this review's findings support a recommendation that accountability policies continue rather than be abandoned, they also strongly recommend that current policies be refined. Furthermore, they suggest that any refinement of existing accountability policies should combine measures of school inputs and outputs with the school organization and instruction practices that connect them.

Measures of inputs and outputs themselves need refining. For example, an achievement metric other than percentile in an achievement-level category (*e.g.,* averages, effect sizes, and percentile-based graphical reporting) would have better statistical properties. Criteria for successful accountability systems are available from several experts, who agree on most points, especially that a successful accountability system requires and supports both external (*e.g.,* state, federal, or public) accountability and internal (*e.g.,* use of data for improvement) accountability.

Some of the studies reviewed here enumerate alternative accountability strategies. By addressing two important questions, those alternative strategies provide suggestions on what to do next.

Must Accountability Mean Federal Accountability? Two general suggestions are: 1) public reporting, with follow-up left to the public's actions; and 2) internal district reporting and planning for improvement, with state/external review. Two related suggestions are made for using the data generated by accountability reports: 1) use them as the basis for local school improvement plans; and 2) use them as the basis for generating questions about achievement, which would lead to local school improvement plans. Or, have stakeholders devise accountability reporting that includes other measures to help explain test results.

Must Accountability Measures Be Based on State Tests Alone? Suggestions include: multiple measures of achievement; student achievement-related accomplishments (*e.g.,* graduation, progression, and enrollment in advanced courses); inputs (*e.g.,* indicators of fiscal, human, and material resources); processes (*e.g.,* indicators of school organization) and instructional practices (*e.g.,* policy implementation, curricular/instructional coherence, and class size); and outcomes other than student achievement (*e.g.,* school safety).

Susan M. Brookhart, Ph.D.

Results of Current Accountability Policies and Measures

Hamilton (2003) noted that large-scale tests have been used as policy tools as far back as an 1840 test designed to monitor schools in Boston. The purposes of testing, however, have changed over time from serving a monitoring function to serving an accountability function. Until recently, educational accountability in the United States did not involve reporting educational or achievement *outcomes,* like test scores, but focused primarily on educational *inputs*. 'Inputs' refers to the human, financial, and material resources available to the school and students. For example, schools would be judged on the basis of library and other physical resources, teacher qualifications and other human resources, and so on. In the late 1980s and through the 1990s, a clear shift to reporting educational *outcomes*—most notably student achievement—had begun, and with it a concurrent interest in developing state standards for that achievement. This has come to be called the "standards movement" (Porter 1993) or the "standards-based reform movement" (Carnoy and Loeb 2002). During the 1990s there was much state variation in approaches to developing outcomes-based accountability policies.

The standards movement marked a shift from focusing on educational *inputs* to focusing on educational *outcomes.*

That changed in 2002. As the most recent reauthorization of the 1965 Elementary and Secondary Education Act, NCLB (passed in 2001 and signed into law in 2002) specifies a much larger accountability role for the federal government, including requiring schools and states to report student achievement and be subjected to federal sanctions. While this paper's purpose is not to evaluate NCLB, widespread dissatisfaction with some of NCLB's accountability policies and measures does form part of its context. Hope that NCLB's pending reauthorization will provide an opportunity to improve the existing accountability system is a significant reason for this paper's existence. Therefore, the following section reviews literature evaluating NCLB's effects on schools, teachers and teaching, and student achievement in order to provide context for subsequent sections' review of literature evaluating alternative accountability policies and measures.

NCLB's Effects on Schools

Jennings and Rentner (2006) summarized four years of research and evaluation results conducted by the Center on Education Policy about NCLB's early (2002–2006) effects. They found, among other things, that:

- student achievement as measured by state tests was rising, although it was unclear whether the number of proficient students was as large as the figures implied;
- there were curricular changes, including more time on tested subjects and less on non-tested subjects;
- the schools labeled "needing improvement" had leveled off to about 10 percent a year (although that may be changing as 2014 nears and the bar becomes higher), but not the same schools each year;
- good progress had been made at meeting teacher qualification requirements, but educators were not sure that would improve classroom instruction;
- attention has been directed to the achievement of subgroups of students; and
- schools that had been restructured as a result of low performance for the most part had not been radically changed.

NCLB's Effects on Teachers and Teaching

Au (2007) summarized 49 qualitative studies looking specifically at curricular change as a result of high-stakes testing, one (although not the only) feature of NCLB. His introduction offers a nice summary of, and citations for, the research debate on this matter. Some authors have found high-stakes testing to be one among many factors influencing teacher practices; others have found that high-stakes testing narrows the curriculum, limits the range of teachers' classroom practices, and corrupts the measurement of student achievement. Au concluded that the main effects reported in these studies are the narrowing of curriculum, increased fragmentation of content taught into pieces learned for the sake of the test, and increased use of lecture-based, teacher-centered pedagogy. However, he also concluded that, in a minority of cases, high-stakes tests led to increases in student-centered pedagogy and in content integration.

Some have found that high-stakes testing narrows the curriculum.

NCLB's Effects on Student Achievement

Fuller, Wright, Gesicki, and Kang (2007) studied the gaps in reported student achievement between state tests and National Assessment of Educational Progress (NAEP), pointing to numerous other studies that do the same. They report that meaningful conclusions about student achievement gains and student proficiency could not be drawn from state tests. In 4th-grade reading, average achievement measured by NAEP climbed one grade level between 1971 and 2004, about half that in the 1999–2002 period (after the institution of state accountability policies and pre-NCLB). The authors report that "some policy mix, rooted in state-led accountability efforts, appears to have worked by the late 1990s. But growth flattened out in 4th grade over the three years after the enactment of NCLB."

In 4th-grade math, Fuller *et al.* report results that are more encouraging: achievement rose about two grade levels between 1973 and 2004, about half that in the 1999–2004 period (including the beginnings of the NCLB era). Achievement gaps for subgroups grew smaller from 1971 through 1992, widened in 1994, grew smaller again until 2002, and have not closed further since then. The one exception to this pattern is the Latino/White gap, which continues to narrow.

Results Comparing Aspects of Accountability Policies and Measures

The research literature documents that NCLB accountability policies and measures have affected schools, teachers, and student achievement. A few of these effects have been positive (*e.g.,* increased attention to subgroups), but many have been arguably negative (*e.g.,* curricular changes). Thinking about improving accountability policies and measures as NCLB is up for reauthorization is reasonable. To this end, the following section compares the effects of types of accountability systems—or at least compares the differential effects of some aspects of accountability systems—by reviewing studies that used pre-NCLB United States data.

In the decade before NCLB, several states geared up accountability systems that collected data on multiple indicators, not just test scores, and required schools to use their own data for improvement. Because there was a climate of accountability in the 1990s, and yet there was not the federal mandate that came with NCLB, research comparing the effects of different aspects of accountability systems was possible. After NCLB, accountability requirements and consequences became much more homogeneous across the country, so there was less variation to study. Thus, after NCLB studies tend to be about the effects of NCLB as opposed to being about the nature and effects of various accountability system designs.

Effects on Schools

Newmann, King, and Rigdon (1996) studied accountability in 24 "restructuring" elementary, middle, and high schools in 16 states. They found that strong accountability was rare; organizational capacity was not related to accountability; schools with strong *external* accountability tended to have low organizational capacity; and strong *internal* accountability tended to reinforce a school's organizational capacity. The policy implication is that, if an external agent like the federal government in a reauthorization of NCLB wants to increase school accountability, it would do well to design systems that strengthen internal accountability, which is linked to organizational capacity.

Those wanting to increase school accountability would do well to design systems that strengthen internal accountability, which is linked to organizational capacity.

Watts (2000) reported on what states in the Southern Regional Education Board (SREB) had developed for school report card and accountability systems and what they were doing to support low-performing schools. The SREB's work led them to define what they called "the basic elements of a sound accountability system" that were key to accountability systems that result in improved

student achievement (Watts 2000). These elements included the following: the existence of content and student achievement standards; testing; professional development; accountability reporting; and rewards, sanctions, and targeted assistance.

Hamilton's (2003) review of assessment as a policy tool also noted that test-based accountability has been reported to affect school climate, including teachers' satisfaction with working conditions. The literature she reviewed also found that testing can improve teachers' motivation if goals are personally significant to the teachers, not competitive, and accompanied by professional development opportunities. In some studies, principals reported providing professional development opportunities as a result of accountability pressure, although sometimes the focus of the opportunities was narrowly about raising test scores.

Effects on Teachers and Teaching

Herman (2004) reviewed a large number of studies done in individual states reporting the effects on instruction of standards-based accountability systems that relied on tests. The studies she reviewed relied on data from early reform movements (like Kentucky's KIRIS) through only pre-NCLB times. She found that the results of these state-by-state studies clearly indicated teachers paid attention to the signal coming from the test, adjusting content and pedagogy to model what students will have to do on the test. This included both more direct test-preparation activities as well as more emphasis on writing and problem solving if these were part of the (pre-NCLB) state test. More instructional time was paid to tested rather than non-tested subjects and skills. Perhaps not surprisingly—but contrary to the intent of the standards-based reform movement—the tests themselves, and not the standards they were meant to measure, received more emphasis in reported instructional and curricular changes.

More instructional time was paid to tested rather than non-tested subjects and skills.

Hamilton's (2003) review reported similar effects on teachers and curriculum, noting however that the literature indicated that effects of testing on teachers and teaching was one influence among many, including teachers' own belief systems, their subject matter and pedagogical knowledge, their experience, and their access to resources.

Diamond and Spillane (2004) investigated the effects of one high-stakes accountability policy and specifically looked at the effects of sanctions. The 1995 Chicago School Reform Amendatory Act gave a mayoral appointee power to place on probation schools with poor performance on the Iowa Test of Basic Skills. Interviews and observations in four elementary schools—two high-performing and two on probation—led the researchers to conclude that accountability status (on probation or not) affected teachers' responses to accountability policies and pressures. In all schools teachers paid attention to test scores, but their responses to them were different. In probation schools, teachers focused narrowly on actions that might help raise scores (*i.e.,* targeting particular students and particular content). In high-performing schools, teachers focused more broadly on practices aimed at raising the performance of all students. In effect, this was the opposite of what the school reform legislation intended.

Effects on Student Achievement

Lee (2006) studied the predictive effects of states' accountability policies on NAEP performance (4th-grade reading and math 1992–2003, 8th-grade reading 1998–2003,

and 8th-grade math 1990–2003). The 2005 NAEP scores were not included because they would have been confounded with NCLB effects, if any. Accountability policies were measured according to the amount of state support for inputs (per-pupil expenditures, class size, and in-field teaching, coded from NCES data) and the amount of press for outputs (uses and consequences of state test results, coded from national policy surveys of three different national organizations). The correlation between these two measures was close to zero, indicating that states' activism in test-driven external accountability was not related to their support for school resources.

At baseline, states' achievement scores in both math and reading were negatively related to accountability press (output guarantee) and positively related to state support (input guarantees); that is, states with more resources and support for schools had higher scores while states with lower performance were more active in test-driven accountability reform. In general, the effects of both accountability press and state support were associated with student performance gains in math but not in reading. The author wondered if this meant the 1990s reform efforts were more concentrated in math than in reading. For 8th-grade math, there was an interaction between input- and output-guarantee policies. When state support for school resources (inputs) was low, state press for student achievement (outputs) made little difference in the size of math achievement gains. When state support for school resources (inputs) was high, state press for student achievement (outputs) was associated with relatively large math achievement gains. The results of this carefully done study suggest that successful accountability system design should pay attention to both inputs and outputs.

Successful accountability system design should pay attention to both inputs *and* outputs.

Using data from four different international student achievement tests, Wössmann (2007) also reports that inputs and external accountability, together, support student achievement. His literature review cited economic arguments that included studies supporting school choice. Wössmann concluded that students performed better in countries where there was not only more competition from privately managed schools (and other choice-related factors) and where there were external accountability tests, but also in countries where teachers had the freedom and incentives to select appropriate teaching methods and where external exams and school autonomy were combined. In other words, both good inputs and processes (institutional autonomy and attention to good instruction) and press for outputs (external accountability press) were required for maximum achievement.

Carnoy and Loeb (2002, 2004) used NAEP math data to compare states' changes in proportion of students who scored at basic or proficient levels as a result of the standards-based reform movement that began in the late 1980s. Their research question asked whether stronger accountability systems lead to increased student achievement, and their main predictor variable was a 0 to 5 scale indicating accountability strength, which they constructed on the basis of information from the Consortium for Policy Research in Education. They also investigated the effects of accountability strength on 9th-grade retention rates and high-school progression rates. Their models used several other control variables: for example, percentage of special education or limited English proficiency students who took the NAEP exam. Investigating effects for different ethnic groups, they found a positive and significant effect of the strength of accountability systems on 8th-grade math for both White and African-American students. Effects at 4th grade were smaller and not statistically significant. Despite positive associations

between accountability and gains in tested achievement, accountability was not associated with changes in retention or progression rates of students through high school.

Three cautions about both metrics and measures are in order for interpreting studies of effects of accountability on student achievement outcomes. First, statistical problems with percent proficient scores (Ho 2008) lead this reviewer to place more weight on the Lee (2006) and Wössmann (2007) studies than on Carnoy and Loeb (2002, 2004). Second, studies of effects of accountability policies on achievement as measured by NAEP typically yield less dramatic results than studies using the state tests as outcome measures, both overall and with respect to smaller reductions in achievement gaps between ethnic groups (Hamilton 2003). Haney's (2000) re-analysis of data about reforms in Texas in the 1990s (concluding that "the Texas 'miracle' is more hat than cattle") has become well-known as an example of how NAEP gains do not always affirm claims made on the basis of state test data. This is why the review of studies for this report has focused on NAEP-based and other large-scale analyses. Third, readers should also remember that predictive associations, even with statistical controls, do not unequivocally establish cause or rule out other (non-causal) explanations.

Effects of Policy Attributes

Desimone, Smith, Hayes, and Frisvold (2005) conducted an interesting study using NAEP 8th-grade math data from 2000 and 2003—NCLB's beginning period. They studied the effects of specific attributes of state accountability policies, based on a theoretical framework that specifies five attributes contributing to successful implementation of a policy: consistency, specificity, authority, power, and stability. They constructed indicators of four (all but stability) for each state and used them to predict 2003 state-average NAEP-based scores, controlling for 2000 scores, in three domains of mathematics achievement: procedural knowledge, problem-solving ability, and conceptual understanding. They found all four policy attributes were related (power and authority most strongly) to one another, but not statistically significantly related to achievement gains when the initial (2000) level was controlled. What relationships were found between policy attributes and achievement were mostly in the procedural knowledge domain, not in conceptual understanding or problem solving. States that were initially lower-scoring gained more in conceptual understanding and problem solving, with stronger influence from policy power and authority. Specificity was the only policy attribute associated with gains in initially higher-achieving states. The causal direction is not clear, however; it could be that higher-achieving states were quicker to adopt specific and detailed standards.

Either the existence (yes/no) or strength (on a short rating scale) of an accountability policy has been the variable studied in most research to date about the effectiveness of accountability policies. The Desimone *et al.* study contributes to the discussion of the concept that accountability policy considerations are not just a matter of strength or pressure (power and authority). The coherence of the system—especially alignment between what is tested and taught and the specificity of standards—also matters, providing implications for NCLB's reauthorization. The alignment of tests and standards is already required in the current NCLB implementation. Less attention has been paid to the quality and specificity of standards and the coherence of the system. While Desimone *et al.* looked at the consistency between standards and tests, the consistency of these with instruction is equally important (Pellegrino, Chudowsky, and Glaser 2001, Porter 2002).

Criteria for Effective Accountability Systems

Several authors have addressed the issue of criteria for effective accountability systems, drawing on professional standards in testing and evaluation, on empirical evidence—including some of the studies reported here and studies of the effectiveness of NCLB—and on principles of research design.

Porter, Chester, and Schlesinger (2004) proposed criteria for effective assessment and accountability systems based on research, on their professional experience, and on published professional standards for testing. An assessment and accountability system should:

- *Provide a good target for student and school effort.* The authors discussed "targets" in several senses: in terms of content standards (in a broad range of subjects and grades), of performance standards (for individuals and groups, and shared responsibility for meeting performance standards), and of rewards/sanctions (for students and schools, sufficient to focus attention on desired content and to increase effort).
- *Be symmetrical.* Both schools and students should be accountable, because schools cannot succeed without motivated students and students cannot succeed without good opportunities to learn.
- *Be fair.* All students should have adequate opportunities to learn, and schools should have adequate resources to provide such opportunities. Accountability measures should be valid and reliable for the manner in which they are used.

Similar themes are echoed in the results of an interview study by the Education Commission of the States. Interviewing members of the National Forum on Accountability and other experts, Armstrong (2002) distilled seven principles, which read much like the above standards or criteria for effective accountability systems with the exception that these principles would hold adults accountable for student performance (Porter *et al.* and Baker *et al.* would hold both adults *and* students accountable).

1. The purpose of accountability systems should be to improve teaching and learning so all students reach challenging standards.
2. Accountability should be expanded to include all levels of the education system—state, district, and classroom.

3. Adults in the system should be held accountable for student performance.
4. Better—and multiple—measures should be used to hold the system accountable. Measures should meet the technical requirements for their purpose.
5. Accountability results need to be both more timely and useful.
6. System capacity should be built to support better data use and improve teaching and learning.
7. Accountability systems should be evaluated on an ongoing basis.

Baker and Linn (2004, also citing Herman and Koretz) proposed criteria for effective accountability systems based on research, on their professional experience, and on published professional standards for testing. These standards (see box on page 9) agree in principle with the standards above, but are written at a smaller grain size.

Criteria for effective accountability systems can be based on research, professional experience, and professional standards for testing

Research-based Accountability Systems

Standards on System Components

- Accountability systems should employ different types of data from multiple sources.
- The weighting of elements in the system, different test content, and different information sources should be made explicit.
- Accountability systems should include data elements that allow for interpretation of student, institution, and administrative performance.
- Accountability expectations should be made public and understandable for all participants in the system.
- Accountability systems should include the performance of all students, including subgroups that historically have been difficult to assess.

Testing Standards

- Decisions about individual students should not be made on the basis of a single test.
- Multiple test forms should be used when there are repeated administrations of an assessment.
- The validity of measures that have been administered as part of an accountability system should be documented for the various purposes of the system.
- If tests are to help improve system performance, data should be provided illustrating that the results are modifiable by quality instruction and student effort.
- If test data are used as a basis of rewards or sanctions, evidence of technical quality of the measures and error rates associated with misclassification of individuals or institutions should be published.
- Evidence of test validity for students with different language backgrounds should be made publicly available.
- Evidence of test validity for children with disabilities should be made publicly available.
- If tests are claimed to measure content and performance standards, evidence of the relationship to particular standards or sets of standards should be provided.

Stakes

- Stakes for accountability systems should apply to adults and students.
- Incentives and sanctions should be coordinated for adults and students to support system goals.
- Appeal procedures should be available to contest rewards and sanctions.
- Stakes for results and their phase-in schedule should be made explicit at the outset of the implementation of the system.
- Accountability systems should begin with broad, diffuse stakes and move to specific consequences for individuals and institutions as the system aligns.

Public Reporting Formats

- System results should be made broadly available to the media, with sufficient time for reasonable analysis and with clear explanations of legitimate and potential illegitimate interpretations of results.
- Reports to districts and schools should promote appropriate interpretation and use of results by including multiple indicators of performance, error estimates, and performance by subgroup.

Evaluation

- Longitudinal studies should be planned, implemented, and reported, evaluating effects of the accountability program.
- The validity of test-based inferences should be subject to ongoing evaluation.

—Baker and Linn (2004)

Alternative Accountability Policies and Measures

Darling-Hammond (2004) pointed out that a test(s) should be seen as information for an accountability system; tests are not the system itself. It has been known for decades that high-stakes use of any social indicator, including educational tests, will tend to shape behavior toward it (Campbell 1969, Lindquist 1951) in both desirable and undesirable ways. Therefore, it is important that an accountability *system* be constructed that does not rely too heavily on one measure, a system that has the effect of teachers and schools working toward high-quality instruction for all instead of avoiding low-quality instruction for some. For many schools, current NCLB policy truly is about avoiding low-quality instruction for some. Balfanz, Legters, West, and Weber (2007) found that one of the strongest predictors of adequate yearly progress (AYP) status in low-performing high schools was the number of student subgroups for which the school was required to report.

An accountability system should not rely too heavily on one measure.

Must Accountability Mean Federal Accountability?

The accountability system enacted in NCLB is 1) federal, 2) based largely on a particular kind of measure of student outcomes, and 3) can result in sanctions. To support thinking about changes to that accountability system in the next reauthorization of the law, this section reviews discussion, examples, and/or empirical evidence:

- of systems where accountability is to other entities than the federal government;
- the use of different kinds of student outcomes measures and the use of different kinds of accountability measures besides student outcomes; and
- discussion, examples, and/or empirical evidence where sanctions are not the primary response to accountability results.

The literature in this section is descriptive. No experimental or quasi-experimental studies were found testing the short- or long-term effects of the policies and measures described here. Current U.S. policy requires states to have approved, peer-reviewed plans in place that result in annual reporting of school proficiency levels in reading, math, and science (based on the testing of all students, including 95% of students with disabilities) and to meet AYP goals charting growth to standard by 2014 for students overall and for subgroups of students.

States and districts have not always been accountable to the federal government in this way. It is reasonable to ask: What other accountability models exist that are not so heavily federal in nature and yet are equitable and effective? The Education Commission of the States (2008) described three different approaches used by states for "results-based accountability systems."

1. Some states developed public reporting systems. Schools gave an account of their programs and performance, and then parents and others could use the publicly available information to, for example, advocate for improvements and/or choose alternative schools for their children.
2. Some states let local districts establish their own standards and criteria. Districts were to set standards for performance, collect data, and use data to create improvement plans. Then states held districts accountable for their own school improvement plans.
3. And, of course, there are high-stakes federal systems, such as NCLB.

If, in the current climate, it is impossible to retreat completely from federally mandated accountability, it may well be possible to incorporate at least elements of the first two kinds of systems into a revised NCLB act. The examples (Milwaukee, WestED, and Rhode Island) described below fit generally into the second category. The system Louisiana initiated, also described below, just before NCLB seems to be a mixture of high-stakes accountability for all schools and additional local work (but accountable to the states) for systems in special need. Another system (in Nevada, Purvis 1997) is described in a subsequent section because the report gives more details about the indicators than about the accountability system (it too fits in the second category of using data to create improvement plans).

Webb (2000), studying reform efforts in the Milwaukee Public Schools, found that local context mattered a great deal. In 1998, the Milwaukee school board had approved new curriculum standards and grade-level expectations, and the requirement that beginning in 1999–2000 8th graders would be required to demonstrate acceptable accomplishment in communication, science, math, and research in order to be promoted to 9th grade. Multiple measures were in use, including both standardized tests and performance assessments. Webb's method of "embedded research" demonstrated that district context had a huge impact on what the district was able to do. Contextual factors that proved crucial in this reform and accountability effort might well be important to other localized accountability models. These included lack of stability of administrative leadership, lack of budget to support the development of psychometrically sound performance assessment, and the alignment and coherence of standards, instruction, and assessments, which can vary from teacher to teacher and class to class.

WestED (Jamentz 2001) designed and implemented "Accountability Dialogues," forums involving both school and community members in discussions to establish agreement on what students should know and be able to do (standards), to build assessment systems to indicate achievement of those standards, to increase understanding of what contributes to student achievement, and to increase a sense of urgency and efficacy among all stakeholders to improve student achievement. Increasing efficacy is perhaps the most original contribution of this sort of system, and it is based in the active participation

of all stakeholders. The WestED document recounts the experiences of five school districts that used Accountability Dialogues and includes materials for other districts to use to conduct Accountability Dialogues themselves.

Rhode Island (1999) designed SALT (School Accountability for Learning and Teaching), a school-centered cycle of activities to improve school and student performance in Rhode Island public schools. Schools formed school improvement teams, which looked at data including the results of state tests and a bi-annual survey of parents, teachers, and students. The team then developed a school improvement plan for improving student performance at a school report night, open to all members of the school and its community. Once every five years, the school was to host a SALT visit, a peer review process by a visiting team. In some cases, RIDE and the districts took on a more directive role, called Progressive Support and Intervention.

Louisiana designed the School Effectiveness and Assistance Pilot (SEAP) process in three phases, for monitoring (all schools), intensive assessment (of some schools identified as needing assistance), and assistance (Teddlie 1998). The plan was based on Act 478 of the 1997 Louisiana legislative session. SEAP-I, the statewide monitoring system, experimented with several different models to arrive at measures of school effectiveness. One used regression to build relative performance indicators identifying schools whose achievement was more than or less than predicted from school context indicators like percentage of free lunches, percentage in special education, percentage of gifted and talented, percentage limited English proficient, and urbanicity. Another was a school performance score calculated as a weighted average of norm- and criterion-referenced test scores, attendance, and dropout rates.

This work was in the pilot stage at the time of the report cited. All schools were monitored. Schools whose performance was identified as being in need of assistance collected a great deal more data, including an on-site visit (SEAP-II), and used those data to formulate and implement a school improvement plan (SEAP-III). This two-tiered model, with monitoring of some data for all schools and additional data and improvement plan requirements for some schools, offers a compromise position between accountability focused heavily on student achievement (for all schools) and a more in-depth collection of indicators of all sorts targeted to just those schools that need to plan for improvement.

This set of literature resonates with the decades-long and well established finding that teacher involvement in decision making is associated with better implementation of school and classroom change (for a classic in this research agenda see Berman and McLauglin 1975). While the intervening decades of research on teacher involvement are not reviewed here, it is worth noting that recent studies of the effects of policy sanctions on teachers (*e.g.,* Finnegan and Gross 2007) converge on this by finding the negative: lack of teacher involvement in decision making is associated with loss of morale for implementing school and classroom change.

Teacher involvement in decision making is associated with better implementation of school and classroom change.

Coupling the need to involve teachers in accountability decisions with the public's desire for measures of achievement that focus on improvement and include examples of student work and teacher observations (Bushaw and Gallup 2008), it follows that accountability systems should pay attention to the stakeholders in the education process. It seems reasonable that a larger role could be found for community members, educators, and even

It seems reasonable that a larger role could be found for community members, educators, and even students themselves, whether or not accountability remains heavily federal.

students themselves, whether or not accountability remains heavily federal. The federal government could, theoretically, even hold states and, through states, local districts accountable for consulting stakeholders. All the literature reviewed in this section focused on consulting stakeholders by presenting student achievement data of some kind and conducting discussions about the data, including seeking perceived causes or reasons for achievement patterns and suggestions and plans for what to do next.

Must Accountability Measures Be Based on State Tests Alone?

From the beginning of the NCLB period up until very recently, schools and districts reported to states the percentage of students measured as proficient in reading and math, usually on one state test. The nature of using this kind of status measure as opposed to a growth measure has been debated, especially among measurement specialists. There are statistical issues to be resolved in using growth models and measures, and this is very much a growing direction of research in the measurement field. The status versus growth debate can, however, remain a debate about how to analyze scores on the one state test. Do accountability measures have to be based on state test data alone?

Status vs. Change (and Variations) in Test Scores. Before moving beyond state tests, it is important to say that using state test data as the main achievement outcome is itself not one single method. Gong (2002) presented four different ways to define 'good' quality student achievement: as achievement status, as achievement change (improvement from one test year to the next), as effectiveness status (students' improvement as they progress through the grades), or as effectiveness change (increase in student improvement rates from year to year). It is important for an accountability system to define what it means by good achievement. At the time of his report (just pre-NCLB), Gong found examples of states using one or more of the first three models. He also pointed out that variations on these four definitions might include a comparable group requirement, either subgroup comparisons or comparable schools comparisons. He noted that it is also possible to combine status and growth: an upper bar might be set, above which students would not be expected to improve; a lower bar might be set, below which—even had there been improvement—schools would be identified as not performing properly.

Some states are currently using or considering growth models with student achievement data. Perhaps the most well known of these is the Sanders model, originally the Tennessee Value-Added Achievement Model (Sanders and Horn 1994). Goldschmidt, Roschewski, Choi, Auty, Hebbler, Blank, and Williams (2005) presented a summary of types of growth models for a policy audience. Technical studies are still addressing questions about the meaning, reliability, usefulness, and relationship to other variables of the various kinds of growth models.

Technical studies are still addressing questions about the meaning, reliability, usefulness, and relationship to other variables of the various kinds of growth models.

Whether measuring status or growth, the percent (or proportion) proficient scores is a poor choice of metric. Ho (2008) demonstrated the dependence of percent proficient scores on where in the distribution the cut score is set for any given proficiency level. The cut score's position in the distribution affects 1) the size of trends

(gains and losses in percentage proficient), 2) the size of gaps between groups, and 3) the size of changes in gaps. The weight of the students near the cut score (more students score in the middle of a distribution than at the tails) can cause differences in percent proficient trends that are so dramatic that gap trends ("closing" or "widening") can not only change size, but they can even switch signs from positive to negative without corresponding real changes in meaning. Ho (2008) recommended the use of scale score averages, effect sizes, and percentile-based graphical reporting of trends. He also recommended paying attention to the percent proficient score problem and to other distributional properties as growth models are developed.

Instructional Sensitivity of Achievement Measures. Another kind of criticism suggests that student achievement outcomes should be measured by tests that are more instructionally sensitive than the current crop of state tests. Popham (2007) summarized the position. In order to draw conclusions about instruction from student test results the tests must be sensitive to changes in instruction. However, most of the current crop of state tests draws heavily from item banks developed in norm-referenced contexts, performance on many of which is related to socio-economic status, and measures many skills with a few items each. Scores on such tests do not reflect changes in instruction well.

In order to draw conclusions about instruction from student test results the tests must be sensitive to changes in instruction.

For one state test (the 2003 Arizona Instrument to Measure Standards 5th-grade math test), D'agostino, Welsh, and Corson (2007) found that there was instructional sensitivity in the narrow sense; that is, the best predictors of achievement were both 1) a teacher-reported match between how the standards were taught and tested and the interaction of reported emphasis on the content and 2) the match of teaching and testing. It could be argued that this kind of narrow instructional sensitivity, where state tests are related to both amount of instruction and whether the instruction was carried out in a style that matched the test, is not what Popham (2007) is aiming for. The next generation of tests should be sensitive to, and able to measure, what students learned regardless of the specific instructional methods used.

Multiple Measures of Student Achievement. Some have called for better tests of student achievement, and others would broaden that to include a collection of indicators of student achievement, both from tests and from other indicators. This is sometimes called a "multiple measures" approach. The multiple measures approach has been used to mean several different things, but two things are basically at issue. First, which of the following count as multiple measures?

- measures of different constructs;
- different measures of the same constructs; or
- multiple opportunities to pass the same test.

Second, which of the following methods are applied?

- conjunctive—using *and* logic, must pass all measures;
- compensatory—where higher performance on one measure can compensate for lower performance on another;

- complementary—using *or* logic, where achieving standard on one of the multiple measures suffices; and (some would add)
- confirmatory—where additional measures are used only to validate primary measures (Chester 2005).

An argument for multiple measures can be made on the basis of providing multiple modes for performance: tests, writing, projects, and so on. Some students may be better able to show what they know by using one modality than by using another. This argument makes the most sense at the individual student level, for uses such as high school graduation or placement decisions. The argument for multiple measures has been made for Native Americans (ERIC 2001) and for English language learners (Garcia 2000). Marwick (2004) found that when multiple measures of academic preparedness were considered in placing students into community college mathematics courses, Hispanic students were initially placed into higher-level courses—where they achieved equal or greater academic success—than when only standardized test score or only high school preparation was considered.

An argument for multiple measures can be made on the basis of providing multiple modes for performance.

Chester (2005) described Ohio's system of using four measures of school effectiveness to classify schools and districts into one of five categories:

- Excellent
- Effective
- Continuous Improvement
- Academic Watch
- Academic Emergency

The four measures themselves were actually sets of measures:

- Performance Indicators
- Performance Index
- Growth Calculation
- Adequate Yearly Progress

Chester illustrated how all four measures contributed to classification decisions. The Performance Index was a composite score based on a weighted index of scores from all tested grades and subjects. Since higher scores were weighted more heavily, the index was sensitive to changes in the distribution of scores.

Performance indices in general are another way to consider multiple measures of student achievement at the school level (Schaefer 2003), and they do not have to be just weighted averages of performance on state test results across grades and subjects. Schaefer described Maryland's School Performance Index (SPI), used for state accountability before NCLB. For a Maryland elementary school, the SPI included both grade 3 and grade 5 performance (the two grades tested at the time) as the average of 13 ratios, six at each grade level (percent satisfactory or better on each of 6 tested content areas) plus the school's overall attendance rate (in percent) divided by the criterion of 94.

Decisions about schools were based on the overall SPI. Other decisions, for example about school improvement plans or curriculum, were based on results for particular tests and subtests. In this way the performance index was an average of quality for judgment purposes, and more fine-grained results could be used for improvement. The SPI was a compensatory index; that is, good performance in one content area could compensate for poor performance in another.

Darling-Hammond, Rustique-Forrester and Pecheone (2005) studied the use of multiple measures of achievement for a different purpose—high school graduation. They reported that graduation rates stayed the same (within 2 percentage points, 2 states) or declined slightly (3 states) from 1998 to 2001 for five states (Indiana, North Carolina, New York, Florida, and South Carolina) that required exit exams only. Results were somewhat better for multiple measures. Graduation rates stayed the same (within 2 percentage points, 3 states) or rose (1 state) from 1998 to 2001 for four states (New Jersey, Wisconsin, Pennsylvania, and Connecticut) that used a multiple measures approach to graduation during that time. In addition, the graduation rates for these four states were higher overall (73% to 86% in 2001) than for the five exam-only states (51% to 67% in 2001). Note, however, that state graduation rates vary widely and are affected by other economic, political, and cultural factors in addition to state graduation policies. The authors concluded that multiple measures of student performance, appropriate alternatives for all students, local performance assessments developed with state support, and a process for review and improvement of local assessment systems can encourage students to stay in school, stimulate more thoughtful teaching, and drive professional development.

Before 2007, the state of Nebraska used a locally-based accountability system (Roschewski 2004) called STARS (School-based, Teacher-led Assessment and Reporting System). STARS relied on locally developed assessments to report student performance on state content standards. STARS was begun in 2000, before NCLB, and never did receive full federal approval. One of its greatest benefits was instructional. Teachers used results for instructional planning and gained in assessment literacy (Bandalos 2004). Bandalos also concluded that there were technical issues still to be resolved within STARS, the most salient being lack of review of the local assessments themselves (district procedures were reviewed, not the resulting assessments) and that compatibility with NCLB was in question.

As proof of concept that multiple measures can be truly multiple and varied, consider the National Educational Monitoring Project (NEMP) in New Zealand (Guskey, Smith, Smith, Crooks, and Flockton 2006). The project tests a sample of, not all, students. The project's assessments are mostly performance based, but each student only works on about a third of the tasks. Nearly all subjects (not just language, mathematics, and science) are assessed. Assessment tasks can be one-to-one interviews, stations (where 4 students work independently on tasks, some using computers), teams (where 4 students work together on collaborative tasks, supervised by the teacher and videotaped), and independent (including tests, but also including art or physical tasks, with videotapes). Each task is scored and reported separately, so there are no 'total reading' or 'total math' scores or the like. Year 4 and Year 8 students perform the same tasks, so teachers can look at overall growth on specifics. No individual or school scores are provided, just national scores (overall and disaggregated by gender, ethnicity, and socioeconomic status). In addition to the tasks, NEMP collects student data on student enjoyment of the task, used partly

to consider the validity of results and for future task design, and student survey data about their perceptions of school subjects and activities. As the name suggests, NEMP is more about monitoring than about accountability. However, as evidence that this kind of assessment does not make students, teachers, or schools soft, so to speak, consider that New Zealand scores are among the best in the world in reading, mathematics, and science (OECD 2007).

Finally, Achieve and Jobs for the Future (2007) has suggested that 'achievement' be broadened to include not only how students score, but 'what students do.' Because this could be considered either multiple measures of achievement broadly construed or measures beyond student achievement, this paragraph comes right between these two sections in this report. In a report prepared for the state of North Carolina, Achieve and Jobs for the Future suggested that next-generation achievement measures in that state should include indicators that high school students:

- stay in school and graduate on time, including 4- and 5-year cohort graduate rates and percentage of 9th graders promoted to 10th grade;
- successfully complete the NC Future-ready Core Course of Study, including percentage of students who earn this diploma, percentage of students proficient or better on end-of-course exams in English, mathematics, science, and social studies in this course of study;
- earn career-ready industry-recognized credentials and/or college credit, including percentage of recent graduates who earn industry-recognized credentials, percentage of graduates earning college credit before graduation (through AP, IB, dual enrollment, and other programs), and percentage of graduates who earn an AA degree within one year of high school graduation; and
- succeed in post-secondary education and careers, including percentage of graduates who score well enough on college placement exams not to need remediation, percentage of graduates who persist in post-secondary education, percentage of graduates who earn degrees, and percentage of graduates who enter the military or find 'family-supporting' employment within three years of graduation.

All of these things are 'achievements,' broadly construed. For some, logic dictates that the school is not the only factor in students' accomplishment.

Accountability Using Measures in Addition to Student Achievement Outcomes. For accountability at the school level—the subject of this literature review—student achievement is the most important outcome but not the only one. Linn (2006) called for accountability systems to gather and report information on student outcomes, prior achievement, student background, school organization, and institutional practices. This list encompasses inputs (prior achievement and student background), processes (school organization and institutional practices), and outcomes (student achievement). As such, it fits in with the research that suggests that inputs and resources as well as strong expectations for student achievement outputs are required for effective accountability systems (Lee 2006, Wössmann 2007). And it fits in with evaluation theory that says the quality of the processes that connect inputs and outputs should be measured and reported.

Reporting inputs, processes, and outcomes requires using multiple measures of more than student achievement. This section reviews literature that reports what some of those constellations of indicators have looked like. A good place to start would be the three input measures from the Lee (2006) study that were related to math achievement: per-pupil expenditures, class size, and in-field teaching. Various indicators of these could be developed; for example, average class size could be reported, or other kinds of student/teacher ratios could be constructed.

Reporting inputs, processes, and outcomes requires using multiple measures of more than student achievement.

Before NCLB, educators and researchers in many states had begun to discuss the importance of educational indicators. As the standards-based reform movement ramped up, indicators were conceived broadly as "any statistic that casts light on the conditions and performance of schools" (Lashway 2001). Porter (1993) detailed the beginnings of the standards movement and wrote that holding schools accountable for student performance outcomes was more tractable from a measurement perspective than holding schools accountable for delivery (input and process). However, he envisioned "a system of school input and process indicators [that] would serve purposes of description, monitoring reform, and suggesting explanations when student performance is not acceptable." He envisioned using input and process indicators at the system, not the school, level.

One process indicator known to be important is system coherence. In a research agenda with colleagues, Porter (2002) developed an indicator of the content of instruction and the alignment of standards, instruction, and assessment, now called the Survey of the Enacted Curriculum. He envisioned alignment indices as useful, among other things, as "a descriptive variable in assessing the coherence of a state's or district's curriculum policy system" (Porter 2002). There are other alignment indicators; for example, indicators associated with procedures for reporting the alignment of state tests and standards developed by Webb (2002) and by Achieve, Inc. (Rothman, Slattery, Vranek, and Resnick 2002). System coherence is a broader concept than alignment (DeSimone *et al*. 2005, Webb 2000), including at least adding the content of instruction (Porter 2002). Other indicators of system coherence might be the fit between standards and curriculum materials, teacher background and expertise, or administrative policies.

Darling-Hammond and Ascher (1991) produced a thoughtful paper about various kinds of accountability and accountability systems as the standards movement was beginning. At the time, they reported there was little agreement on exactly what indicators should be used, and pointed out that for real decision making, locally developed indicators would be most useful. The statistical properties of indicators also must be carefully documented. Averaging and other analytical methods are affected by an indicator's distributional properties, its reliability, and other qualities that can make real differences during analysis and interpretation. Despite the age of the paper, a contribution that seems still useful is their conclusion that indicators should offer at least one of the following types of information (Darling-Hammond and Ascher 1991):

- Problem-oriented information
- Policy relevant information
- Information on educational outcomes

- Information on students' background and placements
- Information about school context factors

Nevada pre-NCLB (Purvis 1997) had a system of school accountability requiring reports from individual schools and from districts. Data were requested and summarized on the following subjects: progress toward goals; enrollment; attendance and truancy rates; dropout rates for grades 9–12; incidents involving violence, weapons, or distribution of controlled substances; counselor-student ratios; teacher-student ratios; average class size; teacher qualifications; academic achievement; standardized test results; parent participation; and expenditures per student. The indicators reported were very specific; for example, counselor-student ratios were requested as well as teacher-student ratios. Average daily attendance, truancy rates, and transiency rates were reported. It should be noted that not all schools provided every kind of data. The purpose of reporting these data was twofold. Schools and districts were accountable to the state for providing information. Schools were also accountable for using their data to inform goal setting and improvement planning. Schools were evaluated in part on the quality of their use of data and improvement plans.

Working with groups of teachers and administrators in Alabama, McLean, Snyder, and Lawrence (1998) identified school inputs, processes, and outcomes that educators thought would be relevant for accountability. The plan was to pilot an accountability model collecting these data in two large districts in Alabama, analyze the data via structural equation modeling (which was part of the model), and then customize the model for other districts. No such report was found, so it is not known how the model worked. However, the educational inputs, processes, and outcomes are of interest for this report because they were the results of committees of educators and because they are at a larger grain size than some of the other examples reported here. Many different indexes could be constructed for each suggestion. McLean, Snyder, and Lawrence's (1998) educators identified the important elements (see the box on page 21) for an accountability system.

Jones (2004) offered what he called a "balanced school accountability model" with four components. Schools would be measured on indicators of these four things, and be accountable to both the state and to parents and community members. The role of the federal government would be to set standards, to provide resources, and to have processes for quality review of systems. This is not the role of the federal government currently; however, the set of indicators is interesting and fits well with the other literature in this review. Jones's (2004) components were: 1) student learning; 2) opportunity to learn; 3) responsiveness to students, parents, and community; and 4) organizational capacity.

A high-quality data storage and retrieval system is necessary for keeping accurate records of a large amount of information. The organization, Data Quality Campaign, (2008) has identified 10 essential elements for a longitudinal student data system.

1. A unique statewide student identifier that connects student data across key databases across years
2. Student-level enrollment, demographic and program participation information
3. The ability to match individual students' test records from year to year to measure academic growth
4. Information on untested students and the reasons they were not tested

5. A teacher identifier system with the ability to match teachers to students
6. Student-level transcript information, including information on courses completed and grades earned
7. Student-level college readiness test scores
8. Student-level graduation and dropout data
9. The ability to match student records between the P–12 and higher education systems
10. A state data audit system assessing data quality, validity, and reliability

Of course records of school organizational and institutional characteristics would likewise require high-quality data storage and retrieval capabilities. One of the reported complaints regarding indicator systems was the amount of time required to prepare reports (Purvis 1997). Indicator systems that are not supported with adequate data storage and retrieval systems for school- as well as student-level data are likely to fail.

Important Accountability System Elements

Inputs

- Financial (*e.g.*, expenditures per average daily attendance; local, state, and federal revenue per average daily attendance)
- Personnel (*e.g.,* pupil-teacher ratio, average teacher salary)
- Facilities
- Equipment
- Materials
- School policy/law
- Student attributes (*e.g.*, SES, average ability)

Processes

- Curriculum and instruction
- Implementation of policies *(e.g.,* admission, grading, promotion)
- Diverse educational opportunities
- Parent involvement
- Leadership (*e.g.,* planning, style, efficiency)

Outcomes

- Academic achievement (*e.g.,* test scores)
- Accomplishments (*e.g.,* graduation rates, college enrollment and completion rates)
- Attitudes
- Retention/dropout rates
- School safety
- Discipline

—McLean, Snyder, and Lawrence (1998)

Do Accountability Consequences Have to Be Mostly About Sanctions?

In the current accountability system, schools that do not make AYP for two years in a row are subject to sanctions that can include parents relocating students to other schools and various restructuring and reorganizing procedures. Do accountability consequences have to be mostly about sanctions? If not, what other consequences might make sense in an accountability system?

Darling-Hammond (2004) characterized the two alternative views about standards-based reform as being based on two different "change theories." The theories are based on two different diagnoses of what the problem is when achievement does not reach desired levels. A diagnosis that lack of effort is the problem, she reported, results in the theory that using standards to apply sanctions to those who fail to meet them will cause change. A diagnosis that the problem is the fact that teachers, schools, and school systems need to learn more effective practices and have better resource allocation to support implementing these practices, she reported, results in the theory that standards should be used to inform "investments and curricular changes that will strengthen schools."

In fact, the discussions, solutions, and evidence found in the literature reviewed for this section of this report suggest that Darling-Hammond's characterization is a good one. All of the alternatives to a system based on sanctions operate on the theory that accountability measures should provide evidence schools can use to improve.

Accountability measures should provide evidence schools can use to improve.

One of the most appealing answers to the question in this consultant's opinion is Linn's (2006) suggestion that accountability system results be used as descriptive data and used to identify schools for more intensive investigation of organizational and instructional processes. Linn's main point was that, in order to support imposing sanctions directly from accountability results one must infer that the schools caused the results. This inference is not justifiable with most accountability system results. Treating results as descriptive, using them to generate hypotheses for further investigation, and *then* collecting more data would be ideal. Linn also points out that more proper attention to when one can infer cause from accountability results would be consistent with NCLB's own call for "scientifically based research."

However, an accountability system that treats its results as the first step in a process that would require full-blown evaluation studies in approximately 10 percent of schools every year—each of which would require its own hypotheses and data collection and

analysis—could be very cumbersome. Some of the alternative data collection suggestions in the previous section amount to a way to standardize collecting and evaluating some of the organizational and instructional process data so that schools could regularly and uniformly evaluate hypotheses about their own school functioning.

In 1993, the state of Nevada designed an accountability system where both data (discussed above) and plans about what to do about the data were to be included in accountability reports. The districts reported data at the school and district level (Purvis 1997) and also submitted a report analyzing the accountability process, describing any exemplary or problematic programs at the district or school level, and enumerating the school district's efforts to address any deficiencies noted. The evaluation and improvement plans were to be based on conclusions from the data. Of course some districts accomplished these plans better than others, and there was concern expressed about the cost of preparing the accountability reports and plans (Purvis 1997). But the design of schools and districts being accountable to the state to use their own data for improvement does represent a significant alternative to having standard sanctions applied to schools and districts.

Conclusions and Recommendations

Readers of the literature reviewed in this paper might form conclusions other than those of the author, who is entirely responsible for the material in this section. However, it does seem reasonable to contend that the literature supports these conclusions—

- NCLB has had both positive and negative effects.
- Accountability is associated with gains in student achievement, most notably in conjunction with appropriate inputs and support, at least in mathematics.
- Both external accountability and internal organizational capacity are needed. Strong internal accountability tends to reinforce a school's organizational capacity.
- Criteria for successful accountability systems are available from several experts, and agree on most points. One point on which they don't agree is about how students should be accountable for their work.

These conclusions in turn seem to support the recommendations that accountability policies should—

- Exist, as opposed to be abandoned.
- Be based on measures of school inputs and outputs, and probably the school organization and instructional practices that connect the two.
- Require and support both external (*e.g.,* state or federal or public) accountability and strong internal (*e.g.,* use of data for improvement) accountability.
- Use a metric for achievement other than the percent in various achievement level categories: averages, effect sizes, and percentile-based graphical reporting have better statistical properties.

This is as much as the research can be said to *recommend*, in the sense that these recommendations are based on conclusions from rigorous studies. While not true experiments, the studies used rigorous statistical controls. As such, they stop short of demonstrating cause, but they do support the reported associations with student achievement and make these recommendations the "best available hypotheses" for what to try next in accountability.

These recommendations beg some important questions, however, the most immediate probably being: What could be done besides federally mandated sanctions?

and What potential accountability measures, besides test scores, have been suggested? Descriptive information was found that enumerated some things that were in the process of being tried when NCLB was signed. No studies were found that evaluated the effects of these specific things, although several authors made the general recommendation that any accountability system should be evaluated. The points (see the box below), then, are more like a summary than conclusions and recommendations. However, they do answer questions about what else has been tried and may save some 'reinventing the wheel' as the reauthorization of NCLB goes forward.

What could be done besides federally mandated sanctions?

- Other than federal high-stakes accountability, two other general models have been identified: public reporting (with follow-up left to the public's actions) and internal district reporting and planning for improvement (with state review). Examples of the second have been found. Follow-through—actual preparation and implementing plans for improvement—is at once the challenge and the promise of such a system.
- Other than to be the basis for punitive sanctions, two related suggestions have been made for using the data in accountability reports: as the basis for local school improvement plans, and as the basis for generating hypotheses to be investigated (which ultimately would lead to local school improvement plans). A proxy for that might be to devise accountability reporting that includes potential explanatory variables (input and process measures).

What potential accountability measures, besides test scores, have been suggested?

- Achievement measures should include multiple measures of the same constructs (at least reading, mathematics, and so on, or perhaps a few important standards like reading comprehension or math problem solving).
- Student achievement-related accomplishments could be reported (things like graduation, progression, enrollment in advanced courses).
- Input measures could include indicators of fiscal, human, and material resources.
- Process measures could include indicators of school organization and instructional practices (*e.g.,* policy implementation, curricular-instructional coherence, class size).
- Outcomes other than student achievement could be included (*e.g.,* school safety).

References

Achieve and Jobs for the Future. July 2007. *Moving North Carolina Forward: High Standards and High Graduation Rates: A Framework for Next-generation Assessment and Accountability Indicators.* ERIC Document No. ED500443.

Armstrong, J. May 2002. *"Next-generation" Accountability Models: Principles from Interviews.* Education Commission of the States Briefing Paper 4029. Retrieved 1/23/09 from http://www.ecs.org/clearinghouse/40/29/4029.htm.

Au, W. 2007. "High Stakes Testing and Curricular Control: A Qualitative Metasynthesis." *Educational Researcher* 36(5): 258–267.

Baker, E.L., and Linn, R.L. 2004. "Validity Issues for Accountability Systems." In S.H. Fuhrmann and R.F. Elmore, eds., *Redesigning Accountability Systems for Education.* New York: Teachers College Press.

Balfanz, R., Legters, N., West, T.C., and Weber, L.M. 2007. "Are NCLB's Measures, Incentives, and Improvement Strategies the Right Ones for the Nation's Low-performing Schools?" *American Educational Research Journal* 44(3): 559–593.

Bandalos, D.L. 2004. "Can a Teacher-led State Assessment System Work?" *Educational Measurement: Issues and Practice* 23(2): 33–40.

Berman, P., and McLaughlin, M.W. 1975. *Federal Programs Supporting Educational Change (vol. 4): The Findings in Review.* Santa Monica, CA: RAND. ERIC Document No. ED108330.

Bushaw, W.J., and Gallup, A.M. 2008. "Americans Speak Out—Are Educators and Policy Makers Listening? The 40th Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes toward the Public Schools." *Phi Delta Kappan* 90(1): 9–20.

Campbell, D.T. 1969. "Reforms as Experiments." *American Psychologist,* 24: 409–429.

Carnoy, M., and Loeb, S. 2002. "Does External Accountability Affect Student Outcomes? A Cross-state Analysis." *Educational Evaluation and Policy Analysis* 24: 305–331.

Carnoy, M., and Loeb, S. 2004. "Does External Accountability Affect Student Outcomes?" In S.H. Fuhrmann and R.F. Elmore, eds., *Redesigning Accountability Systems for Education.* New York: Teachers College Press.

Chester, M.D. 2005. "Making Valid and Consistent Inferences about School Effectiveness from Multiple Measures." *Educational Measurement: Issues and Practice* 24(4): 40–52.

D'Agostino, J.V., Welsh, M.E., and Corson, N.M. 2007. "Instructional Sensitivity of a State's Standards-based Assessment." *Educational Assessment* 12(1): 1–22.

Data Quality Campaign. 2008. "Essential Elements and Fundamentals of a P–12 Longitudinal Student Data System." Retrieved 1/23/09 from http://www.dataqualitycampaign.org/.

Darling-Hammond, L., and Ascher, C. March 1991. *Creating Accountability in Big City School Systems.* Urban Diversity Series #102. New York: ERIC Clearinghouse on Urban Education. ERIC Document No. ED334339.

Darling-Hammond, L. 2004. "Standards, Accountability, and School Reform." *Teachers College Record* 106(6): 1047–1085.

Darling-Hammond, L., Rustique-Forrester, E., and Pecheone, R.L. April 2005. *Multiple Measures Approaches to High School Graduation*. Stanford, CA: Stanford University School Redesign Network. Retrieved 1/23/09 from http://www.srnleads.org/data/pdfs/multiple_measures.pdf.

Desimone, L.M., Smith, T.M., Hayes, S.M., and Frisvold, D. 2005. "Beyond Accountability and Average Mathematics Scores: Relating State Education Policy Attributes to Cognitive Achievement Domains." *Educational Measurement: Issues and Practice* 24(4): 5–18.

Diamond, J.B., and Spillane, J.P. 2004. "High-stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality?" *Teachers College Record* 106(6): 1145–1176.

Education Commission of the States. 2008. *ECS Education Policy Issue Site: Accountability—Current Designs/Models.* Retrieved 1/23/09 from http://www.ecs.org/html/issue.asp?issueid=68&subissueid=92.

ERIC Development Team. 2001. *American Indian/Alaska Native Education and Standards-based Reform.* ERIC Digest. ERIC Document No. ED459039.

Finnegan, K.S., and Gross, B. 2007. "Do Accountability Policy Sanctions Influence Teacher Motivation? Lessons from Chicago's Low-performing Schools." *American Educational Research Journal* 44(3): 594–629.

Fuller, B., Wright, J., Gesicki, K., and Kang, E. 2007. "Gauging Growth: How to Judge No Child Left Behind?" *Educational Researcher* 36(5): 268–278.

Garcia, P.A. 2000. "Increasing the Rigor of Evaluation Studies of Programs for English Learner Students." Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans. ERIC Document No. ED441838.

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R., and Williams, A. October 2005. *Policymakers' Guide to Growth Models for School Accountability: How Do Accountability Models Differ?* Washington, DC: Council of Chief State School Officers.

Guskey, T.R., Smith, J.K., Smith, L.F., Crooks, T., and Flockton, L. 2006. "Literacy Assessment, New Zealand Style." *Educational Leadership* 64(2): 74–79.

Hamilton, L. 2003. "Assessment as a Policy Tool." *Review of Research in Education* 27: 25–68.

Haney, W. 2000. "The Myth of the Texas Miracle in Education." *Education Policy Analysis Archives* 8(41). Retrieved 1/23/09 from http://epaa.asu.edu/epaa/v8n41/.

Herman, J.L. 2004. "The Effects of Testing on Instruction." In S.H. Fuhrmann and R.F. Elmore, eds., *Redesigning Accountability Systems for Education.* New York: Teachers College Press.

Ho, A.D. 2008. "The Problem with 'Proficiency': Limitations of Statistics and Policy under NCLB." *Educational Researcher* 37(6): 351–360.

Jamentz, K. 2001. *Accountability Dialogues: School Communities Creating Demands from Within.* San Francisco: WestED. ERIC Document No. ED456147.

Jennings, J., and Rentner, D.S. 2006. "Ten Big Effects of the No Child Left Behind Act on Public Schools." *Phi Delta Kappan* 88: 110–113.

Jones, K. 2004. "A Balanced School Accountability Model: An Alternative to High-stakes Testing." *Phi Delta Kappan* 85(8): 584–590.

Lashway, L. August 2001. *Educational Indicators.* ERIC Digest. ERIC Document No. ED457536.

Lee, J. 2006. "Input-guarantee versus Performance-guarantee Approaches to School Accountability: Cross-state Comparisons of Policies, Resources, and Outcomes." *Peabody Journal of Education* 81(4): 43–64.

Lindquist, E.F. 1951. "Preliminary Considerations in Objective Test Construction." In E.F. Lindquist, ed., *Educational Measurement*. Washington, DC: American Council on Education.

Linn, R.L. June 2006. *Educational Accountability Systems*. CRESST Technical Report 687. Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Marwick, J.D. 2004. "Charting a Path to Success: The Association between Institutional Placement Policies and the Academic Success of Latino Students." *Community College Journal of Research and Practice* 28: 263–280.

McLean, J.E., Snyder, S.W., and Lawrence, F.R. November 1998. "A School Accountability Model." Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans. ERIC Document No. ED428440.

Newmann, F.M., King, M.B., and Rigdon, M. 1996. *Accountability and School Performance: Implications from Restructuring Schools*. Madison: Wisconsin Center for Educational Research. ERIC Document No. ED412631.

Organisation for Economic Co-operation and Development. 2007. *Executive Summary PISA 2006: Science Competencies for Tomorrow's World.* Paris: OECD. Retrieved 1/23/09 from http://www.oecd.org/dataoecd/15/13/39725224.pdf.

Pellegrino, J.W., Chudowsky, N., and Glaser, R., eds. *Knowing What Students Know.* Washington, DC: National Academy Press.

Popham, W.J. 2007. "Instructional Insensitivity of Tests: Accountability's Dire Drawback." *Phi Delta Kappan* 89(2): 146–155.

Porter, A.C. 1993. "School Delivery Standards." *Educational Researcher* 22(5): 24–30.

Porter, A.C. 2002. "Measuring the Content of Instruction: Uses in Research and Practice." *Educational Researcher* 31(7): 3–14.

Purvis, C.S. January 1997. *Analysis of Nevada School Accountability System, School Year 1994–95*. Carson City: Nevada Department of Education. ERIC Document No. ED461146.

Rhode Island Department of Education. 1999. *SALT WORKS, School by School: A School-Centered Plan To Improve Teaching and Learning. Book One*. Providence: Author. ERIC Document No. ED444960.

Roschewski, P. 2004. "History and Background of Nebraska's School-based Teacher-led Assessment and Reporting System (STARS)." *Educational Measurement: Issues and Practice* 23(2): 9–11.

Rothman, R., Slattery, J.B., Vranek, J.L., and Resnick, L.B. 2002. *Benchmarking and Alignment of Standards and Testing.* CSE Report 566. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing. Retrieved 1/23/09 from http://www.cse.ucla.edu/products/Reports/TR566.pdf.

Sanders, W.L., and Horn, S.P. 1994. "The Tennessee Value-Added Assessment System (TVAAS): Mixed-model Methodology in Educational Assessment." *Journal of Personnel Evaluation in Education* 8: 299–311.

Schaefer, S.D. 2003. "A State Perspective on Multiple Measures in School Accountability." *Educational Measurement: Issues and Practice* 22(2): 27–31.

Teddlie, C., ed. April 1998. "Integrating School Indicators, School Effectiveness, and School Improvement Research: The Louisiana School Effectiveness Pilot (SEAP)." Symposium presented at the Annual Meeting of the American Educational Research Association, San Diego. ERIC Document No. ED422373.

Watts, J. 2000. *Getting Results with Accountability: Rating Schools, Assisting Schools, Improving Schools*. Atlanta: Southern Regional Education Board. ERIC Document No. ED457585.

Webb, N.L. April 2000. "Embedded Research in Practice: A Study of Systemic Reform in Milwaukee Public Schools." Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans. ERIC Document No. ED448227.

Webb, N.L. 2002. *Alignment Study of Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessments in Four States.* Washington, DC: Council of Chief State School Officers.

Wössmann, L. 2007. "International Evidence on School Competition, Autonomy, and Accountability: A Review." *Peabody Journal of Education* 82(2–3): 473–497.

Further Reading

It is interesting to see how many journals have recently published special issues on accountability. The special issues below are recent and varied, were produced by mainstream scholarship (in the opinion of the consultant), are readily available. They present research and theory about the current accountability system, its historical antecedents, and suggestions for improvement.

Chester, M.D., ed. 2005. "Special Issue: Test Scores and State Accountability." *Educational Measurement: Issues and Practice* 24(4).

Floden, R.E., ed. 2003. "Special Issue on Policy Tools for Improving Education." *Review of Research in Education* 27.

Herman, J.L., and Haertel, E.H., eds. 2005. *Uses and Misuses of Data for Educational Accountability and Improvement*. 104th Yearbook of the National Society for the Study of Education, Part II. Malden, MA: Blackwell.

Hollingsworth, S., and Gallego, M.A., eds. 2007. "Special Issue on No Child Left Behind. Section on Social and Institutional Analysis." *American Educational Research Journal* 44(3): 454–629.

Natriello, G., ed. 2004. "Special Issue on Testing, Teaching, and Learning." *Teachers College Record* 106(6): 1045–1400.

Parker, L., ed. 2005. "Special Issue on the Elementary and Secondary Education Act at 40: Reviews of Research, Policy Implementation, Critical Perspectives, and Reflections." *Review of Research in Education* 29.

Susan M. Brookhart, Ph.D., is Coordinator of Assessment and Evaluation for the School of Education at Duquesne University and an independent educational consultant. She is a former Professor and Chair of the Department of Educational Foundations and Leadership at Duquesne University. Previous to her higher education experience, she taught both elementary and middle school. She earned her Ph.D. in Educational Research and Evaluation from Ohio State University. She is a past president of the American Educational Research Association's Special Interest Group on Classroom Assessment. She has been the education columnist for *National Forum*, the journal of Phi Kappa Phi. She is currently newsletter editor for the National Council on Measurement in Education and was program co-chair for the 2004 NCME Annual Meeting. She is the author of two books, *The Art and Science of Classroom Assessment* and *Grading,* and the co-author of *Educational Assessment of Students, Fifth Edition*. She has written or co-authored over 40 articles on classroom assessment, educational measurement, program evaluation, and professional development and serves on the editorial boards of several journals.