J Educ Change (2009) 10:211–227
DOI 10.1007/s10833-009-9105-2

Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform

Jonathan Supovitz

Published online: 28 March 2009

Abstract This article examines major trends in testing and accountability reform in the United States over the past decade. The review covers the apex and decline of the national experimentation with a range of alternative assessments and the rise of test-based accountability as a central policy initiative. These trends signify that testing has become a widely utilized instrument for educational reform in America. Research on these trends indicates that high stakes testing does motivate teachers and administrators to change their practices, yet the changes they motivate tend to be more superficial adjustments in content coverage and test preparation activities rather than promoting deeper improvements in instructional practice. Further, the information provided by large scale assessments is primarily useful to measure school and system progress, but of more limited utility for instructional guidance. Most problematic is that the high stakes testing system in America has been repeatedly promoted as a substantive reform in itself. However, high stakes testing is a relatively weak intervention because, while it reveals shortcomings, it does not contain the guidance and expertise to inform response. The article concludes with suggestions on how to capitalize on the strengths of high stakes testing while minimizing its shortcomings.

Keywords High stakes testing · Assessment · Accountability · Education reform

The past decade has witnessed a tremendous escalation of high stakes testing in the United States. One barometer of this expansion is the vast increase in test sales and accompanying materials. In 1997 test sales from the major test publishers were estimated at $260 million annually. Today they have almost tripled to

J. Supovitz (✉)
Graduate School of Education, Consortium for Policy Research in Education,
University of Pennsylvania, 3700 Walnut Street #404, Philadelphia, PA 19104, USA
e-mail: jons@gse.upenn.edu

approximately $700 million (Frontline 2008). And this number understates rising test expenditures because it does not include the myriad of supporting and other test preparation materials and services offered by smaller vendors.

What have these expenditures bought? What influence has testing and accountability policy had on the behavior and performance of American educators? What has changed over the past ten years as a result? In this article I examine the trends and influences of test-based accountability on American teachers, school and district leaders, and education policymakers. This review covers the end and the beginning of two major cycles of testing reform in the United States. The first half cycle was the apex and decline of a national experimentation with a range of alternative assessments to replace standardized testing. The second half cycle is the rise of test-based accountability as a major policy initiative.

As I will show in this article, these two partial cycles of reform in testing and accountability—themselves extensions of more enduring trends—reveal several important lessons for educational administrators, policymakers, and reformers, not just in the United States, but in other nations as well. First, testing has indisputably become a widely utilized and relatively inexpensive American federal and state policy instrument to leverage change in districts, schools, and classrooms. Second, high stakes assessments do motivate teachers and administrators to change their instructional practices and align their efforts with the high stakes exams. However, third, and related, the changes they motivate tend to be superficial adjustments in content coverage and test preparation activities rather than promoting deeper improvements in instructional practice. Fourth, the information provided by large scale assessments is largely summative and primarily useful to measure school and system progress, but of more limited utility for instructional guidance because it is untimely, detached from local curricula, and does not provide much insight into student thinking or misconceptions. Fifth, and perhaps most important, the high stakes testing system in America, whose primary purposes are to motivate educators, measure system progress, align system components, and signal system accountability to stakeholders, has been repeatedly mistaken for a substantive reform in itself. As a reform, high stakes testing is a relatively weak intervention because, while it reveals shortcomings, it does not contain guidance and expertise to inform response. For this reason, test-based accountability is unlikely to alone bring about enduring educational improvement.

In this article I substantiate these conclusions through three major efforts. First, I lay out four major theories of how testing and accountability are intended to improve the education system. Second, I review the major trends and movements in large scale testing and accountability over past decade in the United States, from roughly the mid-1990s through 2008, contrasting them against the theories of improvement. Third, I conclude with an analysis of the lessons from our experiences with testing and accountability over the past decade and how test-based accountability systems can best be used to leverage meaningful improvements in the educational system.

Although this article focuses on assessment and accountability trends within the United States, it is useful to contextualize it amongst other movements across the world. In the last quarter of the 20th century, it appears that testing has been on

the rise internationally. Phelps (2000) chronicled large scale testing trends in 21 countries from 1974–1999. He found that testing increased in 18 of the countries he examined. The most typical expansions were large-scale, external national assessments, with secondary level exit exams and end-of-course exams also common. About half the nations studied by Phelps used their assessments solely to measure system progress, while half also incorporated stakes for students. Some nations, like Sweden, Spain, and Portugal, devolved authority from a central to a regional responsibility. Phelps concluded that overall, the data "reveals a clear trend towards adding, not dropping, testing programs" (p. 19). Although there are more recent stories of declines in high stakes testing internationally, I could find no more contemporary systematic comparative analysis.

Macro analyses such as Phelps' (2000) also gloss over the more complex sagas within nations as they seek to develop an appropriate role for large scale assessments and serve the needs of multiple constituencies through their assessment systems. Black (1994), for example, presents a cautionary tale of the tensions that arose in England in the late 1980s from efforts to enact a national testing system that contained both teacher assessments and external standardized performance assessments. The political jockeying and subsequent narrowing of the assessment system he described provide a useful backdrop for the American experience chronicled in this article.

What is the theory underlying test-based accountability?

Tests in education are used for many purposes. If a test is used to hold individuals or institutions responsible for their performance and has stakes attached to it, it constitutes what is commonly called a test-based accountability system. Test-based accountability systems are based on the belief that attaching incentives (either positive or negative) to standardized achievement tests will improve student performance (Hamilton et al. 2002). Such systems are broadly supported in the United States as a strategy for improving public education (Hart and Teeter 2004). They are in place in nearly every American state and the cornerstone of federal education policy in the U.S. (Quality Counts 2008).

Within this broad support, different constituencies operate under different notions of the theory underlying high stakes testing systems and its conceptual links to improvements in the educational system. What exactly are the theoretical connections between the imposition of a test-based accountability system and broad scale improvements in student performance?

Upon reviewing the literature surrounding high stakes testing and accountability theory more generally, there appear to be four major underpinning theories that encourage the employment of test-based accountability. The first is *motivational theory*, or the idea that test-based accountability can motivate improvement. The second theory is *the theory of alignment*, or the concept that test-based accountability can play a major role in spurring the alignment of major components of the educational system. The third precept is *informational theory*, or the idea that such systems carry important information that educators can use to guide

improvement. The fourth notion is the idea of *symbolism*, or the notion that an accountability system signals important values to stakeholders. Each of these will be reviewed briefly below.

Perhaps the predominant theory underlying test-based accountability is motivational theory. This theory holds that the extrinsic rewards and sanctions associated with the high stakes test can be used to motivate school faculty members to improve performance. Such "exogenous" approaches are commonly used as "action levers" by policy makers to change the motivation of workers (Katzell and Thompson 1990). As noted by McDonnell (2005), some policymakers hold that extrinsic motivations are not just helpful, but *necessary* to change behavior. An important assumption underlying motivational theory in this context is that educators have, or can acquire, the knowledge to improve performance if incented to do so. This deficit model assumes that educators lack the motivation to improve, and that incentives will address this.

Yet, many teachers and educators have strong internal senses of responsibility to their profession. Since intrinsic and extrinsic motivations are complexly related, one question underlying the use of external motivation is what effect it would have on employees' intrinsic motivations? Research on this is conflicting. Some have found that extrinsic motivations do not reduce intrinsic motivation. For example, Cameron and Pierce (1994) conducted a meta-analysis of 94 experimental studies of the effects of reinforcement/reward on intrinsic motivation. They found that that, overall, reward does not decrease intrinsic motivation. Another meta-analysis by Deci et al. (1999) contradicted the earlier findings and concluded that tangible rewards often do undermine internal motivations.

A second theory underlying the use of high stakes testing is that of *alignment*. The theory of alignment holds that system-wide improvements can best come about if educators align major efforts and initiatives so that they are reinforcing of each other (Smith and O'Day 1991). Since the late 1980s, policymakers have sought to align the major parts of the educational system, including standards, curriculum, and assessments (Chatterji 2002). Policymakers have used high stakes accountability systems as a major component of alignment to produce coherence in the educational system (Fuhrman 1993).

Alignment is usually thought of in terms of synchronizing components of the external system, but Ablemann et al. (2004) identified another important dimension of alignment related to accountability. They posited that schools were not blank slates onto to which external accountability systems sought to influence behavior. They sought to distinguish between the internal accountability cultures within schools and external accountability systems. They defined internal accountability as the relative strength of the norms, values and expectations—both individually and collectively—that shaped school faculties' work. They found that the strength of various dimensions of internal accountability, including who people felt responsible to and for what, had a strong predictive influence on how faculties responded to external accountability systems. Thus, internal accountability influenced response to external accountability.

A third potential theory underlying the use of test-based accountability is that by providing student performance information to local educators and giving them

incentives to improve it, the data will guide improvements in both classroom and organizational decision making. In general, data are seen as an essential ingredient for decision-making, problem solving and inquiry-based learning (Preskill and Torres 1999). More specifically, test data are believed to be important information for teachers and school leaders to make decisions about both students and programs (Boudett et al. 2005; Holcomb 1999). The facile manipulation of data has been made more feasible by the rapid proliferation over the last decade of technology based tools by which to collect, aggregate, and organize a range of quantitative data (Stringfield et al. 2005; Wayman et al. 2004).

A fourth theory underlying the growth and prevalence of high stakes testing is that of public answerability. According to this theory, the public has a right to expect that its resources are being used responsibly and that public institutions are accountable for caretaking the public trust. Along these lines McDonnell (2005) argues that "assessment policies have become a way to show that public education can be rigorous, responsive, and accountable" (p. 45). A slightly different perspective within this theory of public answerability is that high stakes assessments are a way of symbolically validating the authority of the prevailing education system. From this perspective Airasian (1988) observes that high stakes testing programs "symbolize order and control, desired educational outcomes, and traditional moral values" (p. 301). He contended that the political benefit for politicians who enacted such programs was "sufficient motivation despite unexamined implementation" (p. 302).

These four theories of accountability are important, because they provide a set of frameworks against which we can examine past and current trends in high stakes accountability. In the following section I describe trends in testing and accountability that have occurred over the past decade or more, pointing out connections to these accountability theories.

Trends in high stakes testing over the past decade

Movements in the 1980s and early 1990s set the stage for the particular formulations of test-based accountability of the present. Contextually, the 1980s saw a national shift from the tracking of educational inputs (e.g., per-pupil expenditures, teachers salaries, class size, required courses, seat time) as indicators of educational performance to an increased emphasis on testing as a means to hold schools accountable for educational outcomes. The "new accountability" focused on standardized test results as indicators of school and student performance. It contained consequences for excellent and poor performance including public reporting, monetary or nonmonetary rewards, and a range of interventions for low performing schools, including technical assistance, intervention, and even reconstitution (Elmore et al. 1996; Fuhrman and Elmore 2004).

The period from the early 1990s through the mid 2000s saw two major streams of reform that were focused on testing and accountability. The first set of reforms comprised efforts to widen the forms of assessments beyond standardized multiple choice tests by incorporating a range of more "authentic" formats into the testing

system. The second set of reforms was focused on ratcheting up the frequency and stakes of testing systems.

The rise and fall of alternative assessment as a reform movement

By the early 1990s the United States seemed poised for an assessment revolution. Standardized multiple choice high stakes testing was under siege from many corners of the education community for containing gender bias (National Center for Fair and Open Testing 1992), ethnic prejudice (Gardner 1992; Willie 1985) and socioeconomic favoritism (Haney and Madaus 1989). Critics bemoaned the narrowing of curriculum and instruction (Darling-Hammond et al. 1995; Seligman 1989) and the perverse incentives inherent in high stakes testing to retain and reclassify students (McGill-Franzen and Allington 1993). Many felt that the decontextualized nature of multiple choice testing, with its emphasis on the recall of isolated bits of knowledge, represented an outdated behaviorist view of learning, rather than more contemporary constructivist and socio-cognitive perspectives (Resnick and Hall 1998; Wiggins 1992).

Research confirmed many of these critiques. Researchers documented how poor and minority students performed worse on standardized tests than their wealthier, culturally dominant counterparts (Garcia and Pearson 1994; Mullis and Jenkins 1990). Girls tended to perform worse than boys on many standardized tests (Jovanovic et al. 1994; Moore 1989). Efforts to control for ethnic/racial and gender disparities in standardized test performance did not eliminate them altogether, suggesting something inherent in the tests themselves (Pennock-Roman 1992; Rodriguez 1992).

A bevy of alternative forms of assessment, including portfolios, performance assessments, and open-ended tasks seemed the wave of the future. Alternative assessments were viewed not only as a way to 'fix' the problems with standardized multiple choice assessments, but they were also perceived to be a potential catalyst for school reform (Simmons and Resnick 1993). Advocates saw them as a way to gain a more valid measure of student performance, incorporate a richer array of more authentic samples of student capability into assessment, model contemporary curricular reforms, contribute to instructional improvement, and engage students in more meaningful assessment activities (Linn and Baker 1996; Simmons and Resnick 1993).

Several states and national organizations began incorporating alternative forms of assessment into their test-based accountability systems. In the early 1990s, Vermont adopted a statewide portfolio assessment system (Koretz et al. 1994). In the mid-1990s, Kentucky incorporated into its accountability system several different assessment formats including open-response essays, performance tasks in reading, math, science, writing, and social studies, and writing portfolios (Stecher and Barron 1999). The Maryland School Performance Assessment Program (MSPAP) consisted of a set of criterion referenced constructed response performance tasks (Hambleton et al. 2000). The New Standards Project offered performance tasks and portfolios as part of its nationally available assessment system.

Research on the influence of these assessments on teacher and administrator behavior indicated both the motivational and alignment effects of such systems.

Borko and Elliott (1999) found that teachers organized their instruction around the timing of the high stakes assessments. Lane et al. (1999) surveyed Maryland teachers using the Maryland State Performance Assessment Program (MSPAP) and teachers reported that the performance assessments were influencing their choice of curricular activities and their local assessment practices to make them more in line with the form of the state assessment. Stecher and Barron (1999) conducted mixed method research on the impacts of Kentucky's high stakes testing reforms on teachers. They found that the testing system helped to shape teachers' choices for professional development and focus instruction on relevant content areas, and improve students' abilities to demonstrate relevant skills. However, they also found that teachers tended to narrow their emphasis on the curriculum to that which they thought would be tested and to prepare students for the test rather than the larger learning goals in the curriculum.

While the buzz for replacing standardized tests with other formats increased, a bevy of research on the growing use of alternative forms of assessment examined their potential as richer, less biased, more informative ways of assessing student impact. Evidence on the psychometric quality of alternative assessments was mixed. Baxter et al. (1992), for example, found that science performance assessments could be scored with high reliability. However, Koretz et al. (1994) found that Vermont's statewide portfolio system was hampered by unreliable scoring.

Further, the promise of greater equity associated with alternative assessments was disappointing. Studies of the effects on gender and ethnicity of alternative assessments found that the particular tasks and forms of assessment contributed to gender and ethnic differences as much as did the format of the assessment. Supovitz and Brennan (1997), for example, examined the equitability of portfolio assessments relative to standardized tests and found them to reduce racial/ethnic gaps in performance, but exacerbate gender differences. Jovanovic et al. (1994) found that the content of performance tasks produced gender related biases. These findings suggested that while the form of tests played some small part in inequitable performance, racial and gender differences had more to do with deeply rooted social inequities than biases in the tests themselves.

Alternative assessments were also found to be cost prohibitive. Stecher and Klein (1997), for example, examined the cost of large-scale science performance assessments in California and found that they were 20–60 times more expensive than standardized multiple choice assessments for an equally reliable score.

The collective research dampened hopes that alternative assessments could become a viable means of addressing the problems inherent in standardized multiple choice high stakes assessments. However, some elements of alternative assessments subsequently were incorporated into high stakes tests. Open-ended writing tasks and even performance tasks have become regular components of high stakes exams.

The ratcheting up of test-based accountability

The ebbing of the promise of alternative assessment as a means of improving testing roughly coincided with the election of Texas Governor George Bush as President of the United States in November, 2000. One of Bush's signature reforms as Governor

was the institution of a statewide accountability system that called for annual testing and the reporting of subgroup performance (McNeil 2000). Bush brought his form of test-based accountability to Washington, and it was incorporated into the No Child Left Behind (NCLB) Act of 2001. NCLB was a major reform initiative intended to bring about widespread improvements in student performance and reduce inequities between ethnic groups and other traditionally under-served populations.

NCLB required that states adopt test-based statewide accountability systems, whereas in the past many states treated Title I students and others differently. NCLB stipulated that states test annually in reading, mathematics, and eventually science from grades 3 through 8 and one year of high schools. States were to define both proficiency and adequate yearly progress to get all students to proficiency in 12 years. NCLB further required measurable objectives for subgroups including economically disadvantaged students, students from major racial groups, students with disabilities, and limited English proficiency students. Schools that failed to make adequate yearly progress for two consecutive years would be identified for improvement and students would have the right to transfer to another public school (Education Funding Research Council 2002). The law also required states to certify teachers as highly qualified.

Studies and analyses are starting to emerge on the multiple aspects of NCLB. Sunderman (2008) conducted a thorough examination of the implications of the law on the education policy system. They viewed NCLB as a revolutionary expansion of Federal authority because it imposed a single test-based accountability model on all states. Others have noted the specific attention of NCLB to disaggregating subgroup performance and believe this has important potential to call attention to the reasons underlying inequities in performance (Ladson-Billings and Tate 2006; Sunderman 2008). Berry et al. (2004) focused on the law's attention to teacher quality and argued that the NCLB has narrowed the definition of good teaching to mean conveying content at the expense of richer teaching experience, development of the whole child, and the fostering of social skills.

Researchers from the Center for Education Policy (Rentner et al. 2006) conducted a thorough four year analysis of the effects of NCLB on schools and districts based on surveys of state policymakers, district administrators and case studies of schools and districts. They found impacts on alignment as schools are trying to align curriculum and instruction with state academic standards and assessments. They also found a narrowing of the curriculum, as over 70% of school districts reported a focus on reading and mathematics that reduced instructional time for other subjects. They also reported that over 90% of the schools sanctioned under NCLB as in need of improvement were in urban districts, with 54% being Title I schools. Finally, many survey respondents attributed NCLB with credit for rising student performance.

Herman (2004) conducted a synthesis of the pre- and post-NCLB literature on the impact of accountability on instruction. She concluded that accountability gains the attention of teachers, that they model test content and pedagogy, and that test preparation merges with instruction in practice. She also found that teachers were more influenced by testing than by standards, and that non-tested content gets de-emphasized in instruction. Ingram et al. (2004) examined the effects of

accountability on teachers in nine high schools and found that teachers had significant concerns about the kind of information provided by external high stakes assessments and how it was used to judge performance.

In this era, many researchers have explored ways to use the data from high stakes tests to improve instruction (Black et al. 2003; Boudett et al. 2005) While these data provide general information to teachers about students' initial starting points, they lack the nuance necessary for finer instructional guidance (Supovitz and Klein 2003). This has spurred many districts to move towards more frequent quarterly or benchmark assessments (Herman and Baker 2005).

Over time the distinction between the use of assessment for accountability purposes and use of assessment data for instructional improvement purposes has increasingly emerged (Supovitz and Brennan 1997). Citing a number of meta-analyses on what makes up effective performance feedback, based upon a number of experimental and quasi experimental studies (Kluger and DeNisi 1996; Natriello 1987), Black and Wiliam noted that there is substantial evidence that short cycle formative assessments are a potentially powerful means for informing instructional practice to improve student understanding. Several researchers have noted that assessments for accountability purposes reflect a longer feedback cycle than what is desirable to inform teachers. Supovitz and Klein (2003) described state tests as a national map, district tests as a compass, and classroom assessments as a global positioning system to inform teachers about students. Wiliam and Leahy (2006) developed the useful distinction between short-, medium-, and long-cycle assessments. They defined short-cycle assessments as those that give the teacher feedback within the course of a single lesson; medium-cycle assessments that generate feedback from lesson to lesson; and long-cycle assessments that give feedback beyond the instructional unit, from more than four weeks to a year after the data are gathered. While the most powerful evidence of the effectiveness of test data to provide feedback to teachers points to short cycle assessments that are closely linked to the specific curriculum of the classes, the promise of using longer cycle test data for instructional feedback is embodied in a growing set of improvement literature (Bernhardt 1998; Earl and Katz 2006).

Despite some critical press as an unfunded mandate, the overwhelming majority of Americans (80%) have supported the notion of test-based accountability throughout the NCLB era (Hart and Teeter 2004). However, this support may have reached a tipping point in terms of perceived value to improve education. The 2006 Phi Delta Kappan/Gallup public opinion poll (Rose and Gallup 2006), conducted annually since 2000, showed that for the first time, more respondents felt that there was too much testing in public schools (39%) than those who felt there was the right amount of testing (33%). Two thirds of respondents felt that testing encouraged teachers to teach to the test, with 75% of those respondents reporting this was a bad thing.

Two other trends in efforts to use data from high stakes assessments over the past decade are worth noting. The first involve the development of value added models to assess teacher effectiveness. Sanders and Rivers (1996) and Mendro et al. (1998) generated enormous interest in value added models by showing that the cumulative effects of a series of above average teachers produced tremendous growth for

students relative to below average or uneven quality teachers. However, McCaffrey et al. (2003) caution against using these techniques of producing overall effects for evaluating individual classrooms. Analyzing these and other studies using a series of careful simulations, they found that the results of value added models are sensitive to the statistical approach, missing data, confounding effects, omitted variables, and the tests themselves.

Second, some systems have tried to move away from general achievement testing to end of unit exams. End of unit exams have the advantage of tying the high stakes exam more tightly to the curriculum. Many states are using end of course exams in high school rather than generic subject matter tests (Chudowsky et al. 2002). There is very little evidence of the impact of end of course exams relative to less curriculum specific high stakes exams. Bishop et al. (2001) examined the issue from an international perspective and found that countries that use end of course exams outperform those that do not and states, such as New York, which use end of course exams perform relatively well to comparable states.

Finally, over the past five years there has been a healthy debate about whether NCLB has produced gains in student performance. There is some evidence to suggest that there have been improvements in national performance associated with test-based accountability. Hanushek and Raymond (2004) analyzed the National Assessment of Educational Progress (NAEP) and found student performance rose over the late 1990s and early 2000s during the time that states expanded their accountability systems. Changes in performance gaps between majority and minority students were mixed over this time period, as black-white achievement did not narrow, while the Hispanic-white gap did contract. A report by the Center on Education Policy (Kober et al. 2008) found that reading and mathematics performance, particularly in the elementary school grades, has gone up in most states since 2002 and that these results have also been reflected in trends in NAEP. Further, that racial/ethnic gaps in performance on NAEP have also generally narrowed in that time period. Qualitative work by the Center (Rentner et al. 2006) found that educators tended to credit school district policies and programs as important contributors to test score gains. In an earlier study, Grissmer and Flanagan (1998) found that North Carolina and Texas posted the largest gains on NAEP between 1990 and 1997 and posited that these state's attention to test-based accountability was associated with their gains. Thus performance seems to be improving, although the contribution of high stakes testing to these gains is unclear.

What can be learned from America's past experiences with test-based accountability systems?

The last decade or so of the nation's policy experimentation with reforms in testing and accountability has witnessed parts of two substantive reform cycles. First, we saw wide-spread explorations into a variety of alternative assessment forms, including performance tasks, portfolios, and open-ended writing prompts. Second, we experienced an increase in annual testing and greater emphasis on state test performance as the authoritative indicator of the quality of schools and districts.

Based upon these experiences, and in light of the theories of the influence of accountability, what have we learned about the prospects of high stakes accountability to be a constructive force to improve education across America?

High stakes testing does motivate educators, but responses are often superficial

There can be no doubt that state tests have the full attention of most public educators, particularly those in urban and other lower performing schools and districts. Confirmed in both research and lore, the high profile attention to state test performance and the attached stakes are influencing the behavior and practices of literacy and mathematics (and some science) teachers as well as school and district leaders (Borko and Elliot 1999; Herman 2004; McGill-Franzen and Allington 1993; Stecher and Barron 1999). In the best of cases, high stakes testing has focused instruction towards important and developmentally appropriate literacy and numeracy skills. But this has also resulted in a narrower curricular experience for children and a steadier diet of test preparation activities that distract from the larger goals of educating students with the more complex skills and habits to compete in the global economy and a more sophisticated democratic society.

Test-based accountability fosters alignment of the central components of the educational system

The evidence from the last decade suggests that high stakes testing in the United States is encouraging educators to align curriculum, standards and assessments (Lane et al. 1999; Rentner et al. 2006). As the highest profile member of this triumvirate, there is some question of whether, in practice, curriculum and standards are being aligned to the tests, or whether (as is more appropriate) the tests are being aligned to the standards. But the research seems to suggest that the alignment theory of accountability is producing a more coherent education system.

High stakes testing regimes provide system level data, but not useful classroom level information

American experiences with high stakes testing over the last decade and longer has revealed the limitations of these measures as information tools. Research shows that the data that come from high stakes testing are useful for school and system level performance, but are problematic for individual level accountability (either student or teacher) (McCaffrey et al. 2003; Rogosa 2005) or instructional guidance (Supovitz and Klein 2003; Ingram et al. 2004). Recent investigations into the promise of value-added testing show that this technique can produce reliable estimates when aggregated to the school-level, but are sensitive to assumptions and contain considerable error for high stakes decision-making on smaller units (McCaffrey et al. 2003). Annual tests also provide limited instructional guidance for teachers due to their long feedback cycle, lack of close connection to the curriculum, and silence on the nature of student misunderstanding.

Test-based accountability is an appealing political strategy that effectively conveys public accountability

High stakes testing is one of the most visible means for politicians to demonstrate that they can influence what goes on in classrooms within short 2–4 year electoral cycles (McDonnell 2005). Research by Linn (2000) has shown that the introduction of a new test can cause a predictable fall and rise in test score performance within a few years. His results indicate a test effect independent of student knowledge exhibited on the test. Further, the continuing popularity of high stakes tests, despite persistent stories of their misuse and shortcomings (Hart and Teeter 2004), suggest that there is a real need for the education system to openly demonstrate that it is spending public dollars judiciously. Thus test-based accountability serves a useful purpose in demonstrating public accountability. However, this use of high stakes testing is largely symbolic and says nothing about fostering real improvements in our system of education.

Where does the United States go from here?

Despite different emphases, the two partial cycles of reform that have been experienced over the past decade show a repeated desire to solve the problems of American education through changes in the testing and accountability system. However, the over-riding lesson from both the experimentation with alternative forms of assessment and NCLB is that change in the testing system itself cannot resolve the deeper problems of the education system. The tests themselves are not the root cause of the problems of the education system. Reform itself has become confused with the instrument used to measure it. While the testing system can reveal serious educational problems, these problems cannot be fixed by reforming the assessment system alone.

The low levels and inequities in performance revealed through broad-scale testing are real. The results reveal true disparities in our education system driven by our social priorities. While the tests are flawed to be sure, their imperfections are minor compared to the more intransigent problems they reflect. High stakes tests are not circus mirrors, distorting the reality of American society, but rather they show a largely accurate reflection of a societal ambivalence with education.

Rather than investing in more substantial ways to improve teaching and learning, and putting high stakes testing in its appropriate place within a broader improvement system, we are over-relying on a summative testing system to be both a treatment and a monitoring system. We have learned most emphatically over the last decade that the test is not the cure to what ails us. Rather, the inflated role of test-based accountability in our current system is a symptom of our lack of will and capacity to enact deeper reforms.

As we enter into the second part of the NCLB era in America, where do we go from here? It is clear that test-based accountability will not get us anywhere near all schools meeting proficiency by 2014. A more substantive set of reforms are needed so that we can relegate test-based accountability to its appropriate place as a measurement and incentive companion.

Past experience has demonstrated that test-based accountability does a pretty good job of motivating educators to change, but thus far we have failed to incorporate into our assessment system an answer to the crucial question: what do we want to motivate teachers and school and district leaders to do? We now know that in the absence of clear direction, high stakes tests will produce shallower, not deeper, instructional experiences for students.

Developments in two areas will likely determine how much progress American education is able to make in the next ten years. The first is the extent to which reformers are able to *develop regularized responses to patterns found in assessment data*. Well constructed assessments can do a good job of identifying students' strengths and weaknesses, but are silent on the crucial question of what instructional response can best improve student understanding and how to best deliver that instructional action. This suggests that we need more regularized and readily available repertoires of responses to patterns found in assessment data. This would also require a deeper understanding of *why* students don't understand particular concepts. Better information about student misconceptions can be incorporated into assessments, but student misunderstanding is also related to other aspects of their lives that cannot be captured in medium and long cycle assessments. Thus teachers need complementary and less formal measures and other techniques that they can use to hone in on learning problems in addition to a range of responses to increase student understanding. Beyond the development of these tools and technologies, teachers will also need substantial training in their use. In essence, our misguided view of our testing systems as the solution to our educational problems has obscured the simple fact that a good assessment can point out areas of strength and weakness, but much of the guidance for how to act on that knowledge must be found elsewhere.

The second area ripe for more development is the integration of different assessments into a more comprehensive *system of assessment*. We must find a way to assimilate short, medium, and long cycle assessments together into a more coherent system that takes advantage of the strengths of each and ameliorates the undue influence that a single high stakes assessment carries. A system of assessments can provide a series of counterbalances to the influence of state tests to moderate and channel behavior towards desired practices. A more robust assessment system might begin in the schools with more formative assessments, continue with a set of curriculum related interim assessments that act like guideposts, and culminate in a summative annual assessment. Technological advances make the integration and standardization of such concepts more feasible than ever. Developing such a system would no doubt be politically difficult to negotiate, but it would help to more evenly distribute the incentives so they fall across the entire assessment system.

References

Ablemann, C. H., Elmore, R. F., Even, J., Kenyon, S., & Marshall, J. (2004). When accountability knocks, will anyone answer? In R. F. Elmore (Ed.), *School reform from the inside out* (pp. 133–199). Cambridge, MA: Harvard Education Press.

Airasian, P. W. (1988). Symbolic validation: The case of state-mandated, high stakes testing. *Educational Evaluation and Policy Analysis, 10*(4), 301–313.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring on hands-on science assessment. *Journal of Educational Measurement, 29*(1), 1–17.

Bernhardt, V. (1998). *Data analysis for comprehensive school wide improvement*. Larchmont, NY: Eye on Education.

Berry, B., Hoke, M., & Hirsch, E. (2004). The search for highly qualified teachers. *Phi Delta Kappan, 85*(9), 684–689.

Bishop, J. H., Mane, F., Bishop, M., & Moriarty, J. (2001). The role of end-of-course exams and minimum competency exams in standards-based reforms. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 267–345). Washington, DC: Brookings Institute.

Black, P. J. (1994). Performance assessment and accountability: The experience in England and Wales. *Educational Evaluation and Policy Analysis, 16*(2), 191–203.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning*. London: Open University Press.

Borko, H., & Elliott, R. (1999). Hands-on pedagogy versus hands-off accountability. Tensions between competing commitments for exemplary mathematics teachers in Kentucky. *Phi Delta Kappan, 80*(5), 394–400.

Boudett, K. P., City, E. A., & Murnane, R. J. (2005). *Data wise*. Cambridge MA: Harvard Education Press.

Cameron, J., & Pierce, D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research, 64*(3), 363–423.

Chatterji, M. (2002). Models and methods for examining standards-based reforms and accountability initiatives: Have the tools of inquiry answered pressing questions on improving schools? *Review of Educational Research, 72*(3), 345–386.

Chudowsky, N., Kober, N., Gayler, K. S., & Hamilton, M. (2002). *State high school exit exams: A baseline report*. Washington, DC: Center on Education Policy.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action*. New York: Teachers College Press.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*, 627–668.

Earl, L., & Katz, S. (2006). *Leading schools in a data-rich world*. Thousand Oaks, CA: Corwin Press.

Education Funding Research Council. (2002). *Title I monitor*. Washington, DC: Author.

Elmore, R. F., Ablemann, C. H., & Fuhrman, S. H. (1996). The new accountability in state education reform: From process to performance. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 65–98). Washington, DC: Brookings Institution.

Frontline. (2008). *The testing industry's big four*. Retrieved December 26, 2008, from http://www.pbs.org/wgbh/pages/frontline/shows/schools/testing/companies.html.

Fuhrman, S. (1993). *Designing coherent education policy*. San Francisco: Jossey-Bass.

Fuhrman, S. H., & Elmore, R. F. (2004). Introduction. In Susan H. Fuhrman & Richard F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 3–14). New York: Teachers College Press.

Garcia, G. E., & Pearson, P. D. (1994). Assessment and diversity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 20, pp. 337–391). Washington DC: American Educational Research Association.

Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In B. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 77–119). Boston: Kluwer.

Gloria Ladson-Billings, G., & Tate, W. F. (2006). *Education research in the public interest: Social justice, action, and policy*. New York: Teachers College Press.

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas*. Washington, DC: National Education Goals Panel.

Hambleton, R. K., Impara, J., Mehrens, W., & Plake, B. S. (2000). *Psychometric review of the Maryland school performance assessment program (MSPAP)*. Retrieved December 26, 2008, from http://www.abell.org/pubsitems/ed_psychometric_review_1000.pdf.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: RAND.

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan, 70*(9), 683–687.

Hanushek, E. A., & Raymond, M. E. (2004). *Does school accountability lead to improved student performance?* National Bureau of Economic Research Working Paper No. W10591. Cambridge, MA: Author.

Hart, P. D., & Teeter, R. M. (2004). *Equity and adequacy: Americans speak on public school funding: A national opinion survey conducted for the educational testing service*. Retrieved January 5, 2009, from http://www.ets.org/portal/site/ets/menuitem.22f30af61d34e9c39a77b13bc3921509/?vgnextoid=8b5a253b164f4010VgnVCM10000022f95190RCRD.

Herman, J. L. (2004). The effects of testing on instruction. In Susan H. Fuhrman & Richard F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 141–166). New York: Teachers College Press.

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

Holcomb, E. (1999). *Getting excited about data: How to combine people, passion, and proof*. Thousand Oaks, CA: Corwin Press.

Ingram, D., Seashore Louis, K., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*(6), 1258–1287.

Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science be eliminated? *Education and Urban Society*, *26*(4), 352–366.

Katzell, R. A., & Thompson, D. E. (1990). Work motivation: Theory and practice. *American Psychologist, 45*(2), 144–153.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Kober, N., Chudowsky, N., & Chudowsky, V. (2008). *Has student achievement increased since 2002? State test score trends through 2006–07*. Washington, DC: Center on Education Policy.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Lane, S., Ventrice, J., Cerillo, T. L., Parke, C. S., & Stone, C. A. (1999). *Impact of the Maryland school performance assessment program (MSPAP): Evidence from the principal, teacher and student questionnaires (reading, writing, and science).* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Linn, R. L. (2000). Assessment and accountability. *Educational Researcher, 29*(2), 4–16.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 84–103). Ninety-fifth yearbook of the National Society for the Study of Education. Illinois: University of Chicago Press.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

McDonnell, L. M. (2005). Assessment and accountability from the policymaker's perspective. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 35–54). One hundred and fourth yearbook of the National Society for the Study of Education. Malden, MA: Blackwell Publishing.

McGill-Franzen, A., & Allington, R. L. (1993). Flunk'em or get them classified: The contamination of primary grade accountability data. *Educational Researcher, 22*(1), 19–22.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge/Falmer.

Mendro, R., Jordan, H., Gomez, E., Anderson, M., & Bembry, K. (1998). *An application of multiple linear regression in determining longitudinal teacher effectiveness*. Paper presented at the 1998 Annual Meeting of the American Educational Research Association, San Diego, CA.

Moore, E. G. J. (1989). Ethnic group differences in the armed services vocational aptitude battery (ASVAB) performance of American youth: Implications for career prospects. In B. Gifford (Ed.), *Test policy and test performance: Education, language, culture* (pp. 183–229). Boston: Kluwer.

Mullis, I. V. S., & Jenkins, L. B. (1990). *The reading report card 1971–1988: Trends from the nation's report card*. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

National Center for Fair and Open Testing. (1992). *K-12 testing fact sheet*. Cambridge, MA: Author.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*, 155–175.

Pennock-Roman, M. (1992). Interpreting test performance in selective admissions for Hispanic students. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 99–135). Washington, DC: American Psychological Association.

Phelps, R. P. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues and Practice, 19*(1), 11–21.

Preskill, H., & Torres, R. T. (1999). *Evaluative inquiry for learning in organizations*. Thousand Oaks, CA: Sage.

Quality Counts. (2008). Retrieved December 29, 2008, from http://www.pewcenteronthestates.org/uploadedFiles/National%20Highlights%20Report.pdf.

Rentner, D. S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S., et al. (2006). *From the capital to the classroom: Year 4 of the no child left behind act*. Washington, DC: Center on Education Policy.

Resnick, L. B., & Hall, M. W. (1998). Learning organizations for sustainable education reform. *Daedalus, 27*(4), 89–118.

Rodriguez, O. (1992). Introduction to technical and societal issues in the psychological testing of Hispanics. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 11–15). Washington, DC: American Psychological Association.

Rogosa, D. (2005). Statistical misunderstandings of the properties of school scores and school accountability. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 147–174). One hundred and fourth yearbook of the National Society for the Study of Education. Malden, MA: Blackwell Publishing.

Rose, L. C., & Gallup, A. M. (2006). The 38th annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 88*(1), 41–50.

Sanders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future academic achievement*. Technical report, University of Tennessee Value-Added Research and Assessment Center.

Seligman, D. (1989). *A look at student achievement from the school dimension: Demythologizing standardized tests*. Austin, TX: Southwest Educational Development Laboratory (ERIC Document Reproduction Service No. ED 317 562).

Simmons, W., & Resnick, L. (1993). Assessment as the catalyst for school reform. *Educational Leadership, 50*(5), 11–15.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233–267). New York: Falmer.

Stecher, B., & Barron, S. (1999). *Test based accountability: The perverse consequences of milepost testing*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Stecher, B. M., & Barron, S. I. (1999). *Quadrennial milepost accountability testing in Kentucky*. CSE Technical Report 505. Los Angeles, CA: National Center for Research on Evaluation Standards, and Student Testing.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*(1), 1–14.

Stringfield, S., Wayman, J. C., & Yakimowski-Srebnick, M. E. (2005). Scaling up data use in classrooms, schools, and districts. In C. Dede, J. P. Honan, & L. C. Peters (Eds.), *Scaling up success: Lessons learned from technology-based educational improvement* (pp. 133–152). San Francisco: Jossey-Bass.

Sunderman, G. (2008). *Holding NCLB accountable: Achieving accountability, equity, & school reform*. Thousand Oaks, CA: Corwin Press.

Supovitz, J. A., & Brennan, R. T. (1997). Mirror, mirror on the wall, which is the fairest test of all? An examination of the equitability of portfolio assessment relative to standardized tests. *Harvard Educational Review, 67*(3), 472–506.

Supovitz, J. A., & Klein, V. (2003). *Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement*. Philadelphia, PA: Consortium for Policy Research in Education.

Wayman, J. C., Stringfield, S., & Yakimowski, M. (2004). *Software enabling school improvement through the analysis of student data (Report No. 67)*. Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*(8), 26–33.
Wiliam, D., & Leahy, S. (2006, April). *A theoretical foundation for formative assessment*. Paper presented at the National Council on Measurement in Education, San Francisco.
Willie, C. (1985). The problem of standardized testing in a free and pluralistic society. *Phi Delta Kappan, 66*, 626–628.