The Gordon Commission
on the Future of Assessment in Education

Test-Based Accountability

Robert L. Linn
University of Colorado at Boulder
Center for Research on Evaluation, Standards and Student Testing

Test-based accountability has played an important role in education in the United States for over half a century. The continued emphasis on test-based accountability as an educational reform policy over the last 50 years led Elmore (2004) to conclude that it is more persistent than any other policy and that there is no indication that the emphasis on test-based accountability will decrease. The roles that tests play in educational accountability have changed from one decade to the next, but student testing has been a key component of accountability. There are a variety of reasons for the widespread interest in test-based accountability.

Demands for Accountability

Policy makers and the general public have been dissatisfied with student achievement for several decades. In the mid 1960s when the Elementary and Secondary Act (ESEA) of 1965 was enacted, the focus of concern was largely on the generally low achievement of economically disadvantaged students. Over time, however, dissatisfaction has expanded to include a broader range of students. The belief that students are underperforming has been bolstered by the lack luster performance of American students on international assessments, by the increases in the number of students who need to take non-credit-bearing remedial courses in college, and by complaints of employers about the lack of preparedness of high school graduates for work and job training programs.

Another consideration that has fueled the interest in test-based educational accountability is the persistent finding of sizeable gaps in achievement between black and white students, between Hispanic and white students, and between economically-disadvantaged and economically-advantaged students. Substantial gaps in achievement have been documented on a wide range of tests including state assessments, college admissions tests, and the National Assessment of Educational Progress (NAEP). The magnitude of the gaps has remained relatively constant over a number of years. The requirement of the No Child Left Behind (NCLB) Act of 2001 for disaggregated reporting of achievement test results for subgroups of students reflects the imperative of reducing these persistent achievement gaps.

The widespread belief that teachers are not doing an adequate job or working as hard as they should has also led to demands for accountability. Consequently there is a desire on the part of policy makers and the public to do more to hold teachers and other educators responsible for

student learning. Student achievement tests are seen as a relatively inexpensive and objective way of holding educators responsible for student achievement.

Theory of Action

The primary goals of test-based educational accountability systems are (1) to increase student achievement and (2) to increase equity in performance among racial-ethnic subpopulations and between students who are poor and their more affluent peers. The belief that test-based accountability will lead to these goals rests on a number of assumptions.

Assumptions

Proponents of test-based accountability often assume that teachers know what to do to improve student achievement, but aren't putting forth sufficient effort. They believe that teachers and other educational personnel know what needs to be done, but need incentives to put forth the effort needed to improve achievement and reduce gaps. Thus, if accountability mechanisms, involving sanctions and/or incentives, are put in place, it is assumed that teachers and other educators will work harder and student achievement will improve while achievement gaps will diminish.

The use of test-based accountability systems also requires assumptions about the tests that are used. It is assumed that the tests are adequate measures of the important goals, or at least the important academic goals, of education. This requires that the tests provide an adequate representation of adopted content standards, not only when they are first used, but after they have been in place for several years when teachers and schools are being held accountable for results. In other words, the tests must be resistant to various forms of potential corruption such as narrowly teaching to the items that are on the test rather than the broader content domains they are intended to represent as well as direct attempts to cheat.

It is also assumed that test results along with sanctions or rewards attached to the results will increase student and teacher motivation. The increased motivation will lead to changes in teaching practices and that the changes in teacher practice will, in turn, lead to improved student achievement that will be indicated by higher student test scores.

Who is Accountable?

Accountability systems sometimes have targeted individual students. At other times, schools have been the target and only indirectly teachers and other educators in the school. Other systems have targeted individual teachers. Yet other systems have targeted students and schools, schools and individual teachers, or all three simultaneously. The key distinction among the various possibilities is the assignment of responsibility for performance and the associated rewards or sanctions. Porter, Chester and Schleinger (2004) suggested that accountability systems should have symmetry of responsibility. They argued that it is unfair to hold schools responsible if students do not share in the responsibility and vice versa.

Types of Accountability Systems

Low-Stakes Monitoring Systems

The use of standardized tests in the public schools prior to the 1960s was largely for the purpose of providing information about individual student achievement to teachers, students, and parents. At the school, district, and state levels the summary results were used primarily to monitor progress. The stakes were generally low. What stakes there were came mainly from the public reporting of results in newspapers and in school board meetings. Admittedly, the results influenced where parents choose to live and the price of real estate. There were few, if any, sanctions or rewards for schools,, teachers, or students, however. For most students, it was not until they wanted to apply to college and were required to take tests such as the SAT or ACT that there were any real stakes involved.

Stakes began to increase for schools with the enactment of ESEA in 1965. The schools that were affected, however, were only those that received Title I funds and even then there were no rewards for good results and only limited sanctions for poor results. In the 1970s the consequences of poor performance began to increase for students when some states introduced minimum competency requirements.

Minimum Competency Testing

Concerns about the poor performance of high school graduates led to the rapid introduction of additional minimum-competency testing (MCT) requirements by a number of

states. In 1973, only two states had MCT requirements for high school graduation. By 1983, 34 states had adopted some form of MCT requirement for graduation from high school or grade-to-grade promotion. The tests generally tested only low-level basic skills judged to assess the bare minimums needed for the next grade or the award of a high school diploma. It was not long, however, before it was recognized that there was a need to more than minimum levels of performance. This recognition led to calls for tests that went beyond minimum basic skills and measured higher-order thinking and problem-solving skills.

A Nation at Risk

About the time that the MCT movement began to lose steam, several reports were published that expressed dissatisfaction with student achievement and called for tested-based accountability. The reports not only called for increased accountability, but they encouraged the development of systems that went beyond minimum levels of achievement that was needed to meet MCT requirements. A round of reform efforts stressed school-level accountability and attempted to push beyond minimums needed to meet MCT requirements. *A Nation at Risk: The Imperative for Educational Reform*, issued by the National Commission on Excellence in Education (1983), was particularly notable in this regard. *A Nation at Risk* relied heavily on tests both to document shortcomings in student achievement and recommend that tests be used as a mechanism of reform.

The frequently cited conclusion of the report used hyperbolic language to characterize the problem**.** "The educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people" (National Commission on Excellence in Education, 1983, p. 5). *A Nation at Risk* along with several other reports that appeared in 1983 had a major impact. All 50 states introduced some type of educational reform in response to *A Nation at Risk*. Test-based accountability systems were central to most of these state-initiated reforms. Indeed, in many cases, externally mandated tests were relied on as the major instrument of reform.

Building and district test results were used to make educators more accountable for student achievement. Test preparation materials were used heavily by a number of schools and districts to try to increase test scores. In response to the pressure many teachers focused their instruction on the skills tested at the expense of other course content (see, for

example, Haladyna, Nolan & Haas, 1991; Nolan, Haladyna, & Hass, 1992; Shepard, 2000).

Lake Wobegon

Most state and district accountability systems of the 1980s relied heavily on published norm-referenced standardized tests. In the vast majority of cases, states and districts reported results that reflected substantial increases in average scores relative to the national norms for the first few years the test was used. A physician, John Cannell (1987), published a paper in which he noted that almost all states and most districts were reporting that their students were above the national norm. This finding came to be known as the "Lake Wobegon effect," in reference to Garrison Keillor's mythical town where all the children are above average.

Cannell's report attracted a great deal of attention. Follow-up studies (e.g., Linn, Graue & Sanders, 1990) noted that that although states and districts had substantial gains in student test scores in the first few years following the adoption of a test leading to the Lake Wobegon effect, there was almost always a large drop in the average test score the first year that a new norm-referenced test was adopted. This pattern provided evidence that the gains were spurious and due to familiarity of teachers with the specific test. The test score gains did not generalize to the broader achievement domains the tests were supposed to represent. An important implication of the Lake Wobegon experience is that the same test cannot be used several years in a row in an accountability system without having a serious score inflation problem.

Standards-Based Systems

Demands for test-based accountability increased in the 1990s, but there was a major change in the types of tests that were called for. Rather than relying on off-the-shelf norm-referenced tests, states began to demand a new type of test that was based on a set of content standards that were developed and adopted by states. An article by Smith and O'Day (1991, see also O'Day and Smith, 1993) was influential in promoting the idea that systemic reform should start with identifying the subject-matter content that is important for students to learn. Systemic reform was to be built around the identified content standards and would include support for teachers to teach the identified content.

A number of states developed content standards and associated standards-based tests were tailored to the content standards that a state adopted. The test results were reported in

relation to a set of performance standards that were meant to define the level of achievement that students are expected to meet. The reference for the tests was to the content standards, and scores were used to locate students in one of several categories of performance (e.g., below basic, basic, proficient, and advanced, or fails to meet standard, meets standard, and exceeds standard) instead of being reported in terms of how a student's score compared to national norms. Summary results for schools, districts and states were most commonly reported in terms of the percentage of students who scored above the minimum cut score for a selected performance standard. These percent-above-cut statistics were commonly used in place of the normative level of a school or district mean.

Sanctions and Rewards

During the standards-based testing era, states and districts introduced a variety of sanctions and rewards for teachers and schools based on performance on mandated state or district tests. The rewards ranged from commendations to monetary payments. On the sanction side, schools and districts were graded and sometimes placed on probation. Principals were occasionally replaced as the result of continuing poor student test performance. Uniform sanctions across the nation did not begin, however, until the No Child Left Behind (NCLB) law was enacted in 2001.

NCLB

NCLB, President George W. Bush's signature educational legislation, reauthorized the Elementary and Secondary Education ESEA) of 1965. The preceding reauthorization of ESEA under President Clinton, known as the Improving America's Schools Act IASA) of 1994, encouraged the use of content standards as the foundation for tests and performance standards, but did not have any clear consequences for low performing schools. In contrast, the testing requirements and associated accountability mechanisms were quite explicit in the NCLB Act.

NCLB required states to adopt grade-specific content standards in mathematics and reading or English language arts. States also had to develop tests that were aligned with those standards and to set performance standards on those tests. The key performance standard was one that states identified as "proficient" or meets standards. The standards-

based tests adopted by states had to be administered each year to students in grades 3 through 8 and in one grade in high school.

NCLB has very explicit rules that are used to hold schools and districts accountable for student achievement. Schools that receive NCLB funds are required to make adequate yearly progress (AYP) or be subject to a series of corrective actions. Schools that fail to make AYP are designated "needs improvement." Sanctions for schools that continue to fail to make AYP in subsequent years after being identified as in need of improvement become increasingly severe for each additional year that the school fails to make AYP.

Making AYP requires that students in the school score above set targets each year in both mathematics and reading or English language arts and these targets must be set so the state is on progress to have all students (100%) score at the proficient level or above in 2014. In addition, the percent of students meeting proficiency targets must hold for a variety of subgroups of students defined by race/ethnicity, socieo-economic status, disability, and English language status as well as the total of all students. Moreover,at least 95% of the eligible students had to be assessed in each subject.

NCLB required states to adopt student performance standards that would identify at least three levels of achievement (usually called basic, proficient, and advanced). States had to set intermediate performance targets (called annual measurable objectives) each year that would lead to all students performing at the proficient level or above by 2014. The intermediate performance targets are used to determine AYP for schools and districts each year.

One of the notable features of NCLB is the emphasis that it has on assessing all students. Prior to the enactment of NCLB, it was not unusual for state testing systems to exclude many low performing students, students with disabilities, and English language learners. NCLB has changed that demanding that all students be included in the assessments. This has led to wider use of accommodations, such as extra time, having the tests read to the student, providing alternate modes for responding, large print, or brail. It has led to the development of alternate assessments for students with severe cognitive disabilities. Tests are also required to assess student proficiency in English for English language learners.

Growth

As 2014 came closer and closer it has become increasingly clear, as some (e.g., Linn, 2003) predicted early on, that the 100% proficiency target is unobtainable and as AYP targets approach 100% most schools will fail to make AYP. One response to the impending failure of the vast majority of schools has been to introduce some form of growth model to replace the rising targets for a fixed percentage of students scoring proficient or higher. There are several reasons for preferring growth models to achievement status models such as the one specified by NCLB. Learning is demonstrated by growth in student achievement rather by current status. Hence, growth is seen as fairer than status alone for holding schools or teachers accountable (see, for example, Betebenner & Linn, 2010).

The appeal of growth models led the U.S. Department of Education to introduce the growth model pilot program. The states that were approved for the pilot program by the U.S. Department of Education to use some form of a growth model adopted a variety of different models (see, for example, Yen, 2009). Another reason for the appeal of growth models in the context of NCLB is that it was believed that they would provide a means for many schools to make AYP that would not do so using the NCLB status model. However, few schools made AYP using the pilot program growth models that would not already have made it using the status model because the Department of Education required that growth to the 100% AYP targets still be retained.

End-of-Course Exams

End-of-course exams have been introduced in a substantial number of states in the past few years. There are several reasons for the appeal of end-of-course exams for state policy makers. First, there is a widespread desire to increase the uniformity and rigor of high school courses to make high school graduates better prepared to take credit-bearing college courses or enter job training programs without remediation. End-of-course exams are, perhaps, the most effective way of accomplishing these goals.

Second, the NCLB testing pattern used for grades 3 through 8 has not been very satisfactory for meeting the NCLB requirement to test students in at least one grade in high school. Unlike the earlier grades, high school students differ greatly in their course taking patterns. Thus, one 10th grade student may have already taken Algebra I and be enrolled in a

geometry course while another 10th grade student is yet to take Algebra I. Third, there is a greater concern that high school students will be less motivated than younger students to put forth their best effort taking a generic accountability test that has no direct consequences for them than students in earlier grades.

While a few states have used end-of-course exams for some time, a substantial number of states have introduced them in the past few years. The most commonly tested subjects are Algebra I, English, and Biology. Some states have a much broader array of end-of-course xams. Texas, for example, offers end-of-course exams for 12 courses. States may or may not require students to pass end-of-course exams to graduate.

Teacher Evaluation

Belief that some teachers are ineffective has led to a desire for better teacher evaluation systems. Principals' evaluations of teachers typically provide little discrimination among teachers. Dissatisfaction with principal ratings of teachers and a focus on student test scores as the primary accountability mechanism has led a number of states and districts, with encouragement from the U.S. Department of Education, to devise systems of teacher evaluation using student test scores. At the federal level, the American Recovery and Reinvestment Act of 2009, which funded some states through the U. S. Department of Education's Race to the Top initiative, called for states to develop teacher evaluation systems based on student test scores. The primary means of using student test scores to evaluate teachers is to use some form of "value-added" approach.

There are several different value-added approaches that states and districts use. They all link student test scores, obtained at two or more points in time, to teachers. It is thought that teachers can be evaluated in terms of the scores obtained by students in their classes in a given year after taking into account student test results for the previous year or for two or more previous years. Teachers are considered effective if their students score higher than expected based on their performance in previous years. Similarly, teachers are considered ineffective if their students do worse than expected given their test results in earlier years.

Value-added results are intuitively appealing. They purport to place teachers on a level playing field by looking at student growth in achievement during the year in a teacher's class. The reasoning is that it would be unfair to use student achievement status to evaluate teachers

because the students in one teacher's class differ greatly from their counter parts in another teacher's class in terms of their background and preparedness to learn the material in a given grade. Value-added results, on the other hand, are purported to level the playing field by taking prior achievement into account.

Value-added results in the aggregate provide substantial evidence that teacher quality makes a difference in student test scores. A one standard deviation difference in teacher value-added scores s has been found to translate into approximately .1 standard deviations difference in student test scores (e.g., Kane & Staiger, 2008, Chetty, Friedman, & Rockoff, 2011). Although a tenth of a standard deviation may seem small, according to Kane and Staiger's (2012) analyses it is roughly equivalent to about a third of school year. That is, a student in a class of a teacher who has a value-added score one standard deviation above the mean would be expected to have an increase in his or her test score in a school year of one and a third school years, while a student with an average teacher would gain only one year for a year in school.

Not only are positive value-added estimates of teacher quality derived from greater gains in student test scores, but they have also been shown to predict long-term outcomes such as the likelihood that a student will attend college, the quality of college that a student will attend, the likelihood of having a baby when still a teenager, and the their life-time earnings (Chetty, Friedman, & Tockoff, 2011).

Although these associations of aggregate value-added results are impressive, it does not follow that that value-added scores are adequate for the evaluation of individual teachers. The adequacy of value-added results for evaluating teacher effectiveness is debatable. Critics (e.g., Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, & Shepard, 2010) have identified a number of reasons why value-added results alone are not an adequate basis for judging teacher effectiveness. There is substantial volatility in a teacher's value-added scores from one year to the next. Value-added results depend on the particular analytical model that is used and on the variables included in the model. Gains in test scores are not the same as gains in achievement because of limitations in the tests and there are able research findings documenting that a narrow focus on test scores often leads to inflated test scores. Furthermore, the goals of education that teachers are expected to pursue for their students are broader than the knowledge and skills that are measured by standardized achievement tests.

Another limitation of value-added analyses for teacher evaluation is that they are generally available only for teachers who teach English language arts or mathematics in grades 4 through 8. The restriction to grades 4 through 8 is due to the NCLB requirement that tests be administered in reading or English-language arts and mathematics in grades 3 through 8 and one grade in high school. Since test data are usually not available prior to grade 3 and value-added analyses require test data at two points in time, the typically available test data do not provide a basis for computing value-added scores for teachers who teach below or at grade 3. This limitation also applies at the high school level since testing is required at only one grade in high school. Furthermore, many high school teachers teach subjects other than English-language arts or mathematics.

Because of these limitations of value-added results for the evaluation of individual teachers and results of their own analyses using teacher observations, Kane & Staiger (2012) have advised against the use of value-added results alone for teacher evaluations. They recommend that high-quality observations and student survey results be combined with value-added results for purposes of teacher evaluation.

Effects of Accountability Systems

Trends in Test Scores

The majority of states with sufficient data on state tests that were comparable over the years 2005-09 showed increases in their state test results (Chadowsky & Chadowsky, 2010). The increases were generally larger in mathematics than in reading and in the lower grades (most commonly grade 4) than the higher grades (most commonly grades 8 or 10), but, in both subjects and in both the higher and lower grade, most states had upward trends using either the percent of students who scored proficient or above or the test score means. Chadowsky and Chadowsky (2010) used state trends on the National Assessment of Educational Progress (NAEP) to check on the degree to which the trend on state test scores were confirmed by trends on NAEP. They found that most states had gains on NAEP, but the gains on NAEP were generally smaller than the gains for the state-specific tests.

NCLB is intended to provide incentives that will result in increases in student achievement. The observed trends are consistent with the conclusion that NCLB accountability requirements have had a positive impact. There are many factors other than NCLB that may

have produced the observed test score gains. The observed gains on NAEP suggest that the state gains are not entirely spurious, but the fact that NAEP gains tend to be smaller than the state test scores suggests that those scores may be somewhat inflated.

In addition to the goal of increased achievement, NCLB is also intended to reduce the gaps in student achievement. The gaps in state scores declined slightly between 2005 and 2009 in a number of states. The decreases in the gaps between black and white students, between Hispanic and white students, and between low income students and their more affluent counterparts were generally quite small, however, and in a few states there was either no change or a slight increase in the gaps (Center on Education Policy, 2010). As in the case of the trends in state test scores for all students, there are many factors other than NCLB that may have led to the changes in the size of the gaps in test scores.

Increased Test Scores vs. Increased Achievement

As was noted above, increases in test scores from year to year may or may not reflect increases in achievement. State tests tap only a fraction of the content specified in state content standards. Some content standards may be too difficult or too expensive to measure on a state test. In addition, scores on tests that are administered by states each year may be inflated due to narrow teaching to the test content rather than the broader domain of the content standards (Haut & Elliott, 2011). Several studies (e.g., Chadowsky & Chadowsky, 2010; Jacob, 2005, 2007; Koretz, 2002; Koretz & Barron, 1998) have found that state increases on low-stakes tests are smaller on than the increases reported on high-stakes tests used for holding schools accountable. These results raise serious questions about the validity of the interpretations of educational test results from tests used for purposes of school accountability (see, for example, Baker & Linn, 2004).

State tests used for purposes of NCLB are revised each year so the simple explanation that applied in the 1980s when The Lake Wobegon Effect identified by Cannell (1987) that familiarity with and teaching to a specific test form no longer applies. In order to equate the tests from one year to the next, however, a substantial subset of items from the previous year are used as an anchor test. Thus, only part of the test is unique each year. Furthermore, test specifications remain the same from one year to the next and, on some state tests, a common "item shell" is used to assess a particular concept. For example, the specification for a test to assess student

knowledge and ability to use the Pythagorean theorem may always be tested by items that ask students to find the length of a ladder given the height of a building, and the distance of the base of the ladder from the building (Koretz, 2008b). Teachers may drill students on items that ask questions in this form resulting in students being able to answer the test item without a more general understanding of and ability to apply the Pythagorean theorem in a broader array of situations.

Based on its review of the research evidence, the National Research Council's (NRC) Committee on Incentives and Test-Based Accountability in Public Education (Haut & Elliott, 2011) concluded that test-based accountability had led to only small improvements in student test scores. Since the test score gains are inflated to some unknown degree, this would seem to imply that incentives provided by test-based accountability have not been effective in improving student achievement. The conclusions of the NRC Committee, however, have been sharply criticized by (Hanushek (2011) who argued that the Committee's conclusions were based on a selective review of the evidence and are far too negative. Whatever the magnitude of the effects of test-based accountability on student achievement is, those positive intended effects need to be weighed against unintended negative side effects.

Negative Side Effects

Teachers' responses to surveys indicate that they do use test preparation materials and narrow their teaching to material covered on tests when high stakes are attached to the results (Hamilton, Stecher, Marsh, McComb, Robyn, Russell, Naftel & Barney, 2007; Koretz, Mitchell, Barron & Heath, 1996). This narrowing of the focus of instruction can lead to inflated test scores that give a misleading impression of the changes in student achievement. Unfortunately a small, but notorious, fraction of teachers and administrators go beyond test preparation and narrowing of instruction and find ways to inflate test scores by giving clues during test administration or even changing student answers. Such unethical behavior has led to a rash of cheating scandals in the last few years.

Narrowing the focus of instruction, teaching to the test, and cheating are all unintended negative consequences of high-stakes, test-based accountability in the content areas tested (usually reading and mathematics). There is also evidence that content areas such as history, art, music and science that are usually not part of the accountability system get less attention than

reading and mathematics. Thus, not only are scores in the tested subjects inflated, but achievement in other subject areas may suffer.

Lessons Learned

Experience with test-based accountability systems has shown that the same test form cannot be used year after year. Even when the test form is changed from one year to the next the validity of the results may be compromised by the use of test items that present problems in the same way each year. Accountability systems need to include mechanisms to evaluate score inflation and guard against it. One approach is to use multiple measures (Haut & Elliott, 2011). It is also useful to use progress on low-stakes tests such as NAEP to monitor progress shown on high-stakes accountability tests. Another promising approach is to design tests to have a self-monitoring mechanism (Koretz, 2008a; Koretz & Bequin, 2010).

Some Options for Future Accountability Systems

Although the reauthorization of NCLB is long overdue, it is not clear when it will be reauthorized. It is evident, however, that there is considerable dissatisfaction with accountability requirements of the current law. It remains to be seen what implications that dissatisfaction will have for the accountability requirements in the reauthorization, but there are several activities that are likely to help shape the requirements.

Common Core State Standards

The Common Core Standards (CCSS) were developed under the auspices of the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA) (http:///www.corestandards.org/). They were developed in response to the concerns that state content standards varied substantially from state to state and in many cases lacked rigor.

The developers of the CCSS were informed by the strengths and lessons from existing state content standards and by the learning expectations in other high-performing countries. The standards are intended to be rigorous, and they stress higher-order learning and problem-solving skills needed for the 21st century. It is intended that students

who master the high-school level common core standards will be prepared for college and job training programs. Forty five states and the District of Columbia have adopted the CCSS.

State Assessment Consortia

Adoption of the CCSS has direct implications for state testing programs. The requirement that state tests be aligned with the state content standards adopted by a state is likely to require substantial revisions in the tests states use. The U.S. Department of Education has funded two multistate consortia to develop assessments that are aligned with the CCSS. The state consortia are called the Partnership for Assessment of Readiness for College and Careers (PARCC) (www.parcconline.org) and the Smarter Balanced Assessment Consortia (SBAC) (www.k12.wa.us/amarter/default.aspx). All of 45 states that have adopted the CCSS have joined one or both of the consortia. Twenty-four states and the District of Columbia are members of PARCC and 28 states are members of SBAC. Both consortia are expected to have tests available for use by states by 2014. After 2014, when the funding for development ends, there will need to be some mechanism, probably involving the major test publishers, for distribution, administration, maintenance, and updating of the tests.

As the name suggests, PARCC hopes to develop an assessment system that will increase the number of high school graduates who are prepared for success in college and in the workplace. To accomplish this goal, PARCC plans to develop assessments that can be used to determine "whether students are college- and career-ready or on track." It is intended that the assessments will "assess the full range of the CCSS, including standards that are difficult to measure." The assessments are intended to measure achievement of both high- and low-performing students. It is hoped that the assessment system will provide the data needed for accountability purposes including the measurement of growth (www.parcconline.org/parcc-assessment-design).

To accomplish these ambitious goals, PARCC plans to use computer-administered tests that "will include a mix of constructed response items, performance-based tasks and computer-enhanced, computer-scored items" (www.parcconline.org/parcc-assessment-design). There will be two required summative assessments and three optional non-summative components, one of

which will measure speaking and listening skills. Once developed, it is planned that the assessment system will be available to all member states.

The SBAC assessment system plans are equally ambitious. Similar to the PARCC plans, the SBAC plans to develop an assessment system comprised of both required summative and optional interim assessments. SBAC also plans to develop an array of formative tools that will help teachers assess student acquisition of the CCSS and diagnoses student learning needs. Like PARCC, SBAC plans to measure "progress and attainment of the knowledge and skills required to be college and career ready."

SBAC expects to develop computer-adaptive tests and performance tasks for both the summative and interim assessments. The plans call for assessments that include both the computer-adaptive component and "performance tasks that will be administered in the last 12 weeks of the school year in grades 3-8 and high school for English Language Arts (ELA) and mathematics." (www.k12.wa.us/smarter/pubdocs/SBASSummary.2010.PDF).

The emphasis on inclusion of all students that is a prominent feature of NCLB will continue to be important for new systems developed by PARCC and SBAC. Both of these state consortia plan to provide a system of accommodations to make it possible for students with disabilities to participate in the assessments. Because of the desire to include all students, the U.S. Department of Education has funded two state consortia to develop alternate assessments for students with severe cognitive disabilities and two state consortia to develop tests for English language learners. The assessments developed by these consortia should lead to improvements in the measurement of the English proficiency of English language learners and the measurement of the degree to which students with cognitive disabilities have mastered the CCSS.

Non-Test Approaches to Accountability

In the United States there has been a heavy emphasis on quantitative approaches in educational accountability. The test-based accountability requirements of NCLB are consistent with that tradition. In many other countries, however, there is much more reliance on qualitative approaches to educational accountability. Great Britain, for example, has a long tradition of using school inspections for purposes of school accountability. The Office of Standards in Education, Children's Services and Skills (Ofsted) has a well-developed system that relies on a cadre of inspectors to visit schools and provide accountability reports.

A general overview of the purposes, policies and principles are provided in *The Framework for School Inspection* (http://www.ofsted.gov.uk/resources/framework-for-school-inspection-january-2012). The Framework also specifies the focus of school inspections, which includes a descriptions of the judgments made by inspectors and procedures that are followed in the inspection process. Although quantitative student achievement results are considered in the inspection, the emphasis is clearly on expert judgments of Her Majesty's Inspectors (HMI) and other inspection service providers.

The HMI base their judgments on a variety of information sources and generally includes interviews with teachers and other school personnel and classroom observations. They use the interviews and observations to report on the quality of teaching, the quality of leadership and management, and the behavior and safety of students as well as information regarding student achievement. The inspectors produce reports and assign schools one of four grades: outstanding, good, satisfactory and inadequate. Schools receiving grades of outstanding or good are inspected less frequently than schools than schools that receive one of the two lower grades. Schools that receive a grade of inadequate may be given a notice to improve or special measures may be taken to improve the effectiveness of the school.

Finland, which had an external inspection system of school accountability, has moved away from the inspection system and adopted a system relying on school self-evaluation. A comparative study of schools in England and Finland concluded that both approaches had some advantages and some disadvantages (Webb, Vulliamy, Hakkien & Hamalainen, 1998).

A number of classroom observation systems have been developed in the United States. Although the observation systems have usually been used for research purposes or to provide teachers with feedback intended to improve their teaching practices, the systems have also been used for teacher evaluation and accountability. Some of the observations systems are designed for use in a single content area. For example, as is evident from their names, the Protocol for Language Arts Teaching Observation (PLATO) system (http://cset.stanford.edu/media/PLATO_Overview.pdf) was developed for use in English language arts classrooms while the Mathematical Quality of Instruction (MQI) system (http://sitemaker.umich.edu/lmt/files/lmt-mqi_glossary_1.pdf) was developed for use in mathematics classrooms. Other observation systems such as the Classroom Assessment Scoring

System (CLASS) (http://www.teachstone.org/about-the-class/) system and the Framework for Teaching (FFT) (http://www.danielsongroup.org/Default.aspx) system were designed for use in multiple content areas.

Qualitative approaches to school accountability are clearly possible, and were, in fact, used long before quantitative, test-based accountability approaches. Qualitative and quantitative approaches to school and teacher accountability both have strengths and shortcomings. A mixed-model approach that builds on the strengths of both qualitative and quantitative would seem to be preferable to relying solely on one approach or the other.

Combining Qualitative and Quantitative Approaches to Accountability

A mixed-model accountability system could give priority to either qualitative or quantitative results or might seek a balance between the two approaches. Hall and Ryan (2011), for example, describe what they call a "qualitatively-driven mixed methods approach" to educational accountability that that emphasizes the importance of qualitative information. A modification of traditional test-based accountability could use the student growth on achievement tests as a trigger to identify schools where inspections would be conducted to collect qualitative information that might explain low rates of student growth and suggest possibilities for improvement rather than the current use by NCLB to sanction schools found in need of improvement (Linn, 2005; 2008).

The Measures of Effective Teaching (MET) project (Kane and Staiger, 2012) is a collaborative effort supported by the Bill and Melinda Gates foundation that explicitly includes a combination of both qualitative classroom observation results and test-based quantitative results using value-added analyses. The Kane and Staiger investigation included five classroom observation systems that were found to yield reliable results. In addition to the four previously mentioned systems (PLATO, MQI, CLASS, and EFT) they used the UTeach Observation Protocol (UTOP), (https://wikis.utexas.edu/display/physed/UTeach+Observation+Protocol), for a sample of their schools. They found that all five observation systems yielded results that were correlated with value-added results, and unlike the later, provide a basis for giving feedback to teachers about ways to improve their teaching. Kane and Staiger (2012) concluded that a combination of

classroom observations, student feedback and gains on achievement tests was better than any of the approaches alone.

Maximizing Intended Positive Effects and Minimizing Unintended Negative Effects

Past experience with test-based accountability systems shows that the systems need to be designed with considerable care. The tests need to be designed to measure the knowledge and skills that are important. The validity of the uses of test scores for purposes of accountability needs to be evaluated (AERA, APA, & NCME, 1999, Baker, Linn, Herman & Koretz, 2002). The tests need to be aligned with high-quality content standards that specify the desired knowledge and skills. The recently developed Core Content Standards fit that description and provide an excellent framework to guide test development.

High-quality content standards, while critical, are insufficient. They need to be accompanied by tests and instructional materials that are well aligned with the standards. The tests need to provide broad and representative coverage of the content standards at a depth of knowledge called for by the standards. They need to assess those content standards that are hard to measure as well as those that are more readily measured by a standardized achievement test.

Both the SBAC and PARCC consortia promise to develop tests that are aligned with the CCSS and meet the other characteristics noted above. The consortium plans are ambitious and promise to use technology that will make it possible to assess hard-to-measure standards. If the goals of the consortia are realized, states will have tests that are substantially better than those that are currently being used. Although promising, it remains to be seen the degree to which the ambitious goals of the consortia will be realized.

Past experience has clearly shown that it is unacceptable to use the same test form from year to year when high stakes are attached to results. A new form of the test that is equated to forms used in previous years must be developed each year. It is also important that the way in which key concepts and understandings are assessed needs to change from year to avoid the problems caused by teaching to a specific representation of a concept rather than to a broader understanding of the concept (Koretz, 2008a, b).

Accountability systems need to stop relying on current status scores, which are fundamental for the current NCLB requirements, and put the emphasis on growth. Learning implies a change in achievement rather than a particular level of achievement at a fixed point in

time. Growth also provides a fairer basis of comparison than current status for comparing schools whose students start at different levels of achievement.

Systems also need to include indicators of the degree to which tests used for accountability provide an inflated impression of achievement gains. This may be done by comparing trends in achievement on the high-stakes accountability tests to trends for low-stakes tests such a NAEP. Alternatively, accountability systems could include some form of self-monitoring mechanism such as that suggested by Koretz and Bequin (2010).

It is important that testing systems used for accountability purposes include as many students as possible. Thus, there is a need to provide accommodations for students with disabilities to participate. It important that the systems include a means of determining when English language learners have sufficient proficiency in English to take subject area tests in English. It is also important to have a system of alternative assessments for students with severe cognitive disabilities so that those students can be included in the overall accountability system.

Test-based accountability systems should be supplemented by qualitative information about the quality of teaching. Such information might be obtained from interviews or classroom observations either as a routine part of the system or at least for schools that are rated low based on the growth in achievement test scores. The qualitative information is likely to be more useful than the test results in suggesting ways that teachers can improve.

Providing Instructionally Useful Data

Instructional utility has long been a goal of standardized tests that are used primarily for accountability purposes. There are several factors, however, that make this goal illusive, if not unattainable. Teachers need the results immediately if they are to be used to modify day-to-day instruction. There is a lag of at least several weeks, however, between the time state tests are administered and the time results are reported. Even, if the use of computer-administered tests allowed the immediate reporting of results, the tests are generally administered near the end of the school year when there is little time left for teachers to use the results.

End-of-year tests do provide teachers with targets for the following year. This has both positive and negative sides. On the positive side, the tests often clarify aspects of the content standards that are particularly important. On the other hand, as was noted above, they can also

narrow the focus to only those aspects of the content standards that are regularly assessed and to particular ways of presenting problems that may lead to inflated test scores.

The difficulty of getting information that is truly useful for the most important instructional uses from end-of-year accountability tests has led to the introduction of a variety of other tests for use in earlier times of the school year. Perie, Marion, Gong, and Wurtzel (2007) distinguish three types of assessments (summative, interim, and formative) that are often included in accountability systems. The summative assessments are the end-of-year tests that have been the focus of this paper.

Interim assessments, which are sometimes called benchmark or diagnostic assessments, are tests that have are administered at various points during the school year. These tests are intended to provide teachers and students with information that is predictive of subsequent performance on the end-of-year summative assessment. It is expected that the early warning of potential difficulty will help teachers better prepare their students for the end-of-year tests that count for accountability purposes.

Although "formative" is frequently used to describe instruments that fit the Perrie, et al. (2007) definition of interim assessments it is misleading to call them formative assessments. Formative assessment is a process that teachers engage in to monitor student achievement on a day-to-day basis rather than a test with a fixed set of items (Perie, et al. 2007, Popham, 2008). Popham defines formative as follows.

> Formative assessment is a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning-tactics. (Popham, 2008, p.6)

This notion of formative assessment is clearly different from a test provided by an external source. Nonetheless, external supports for teachers to use in doing formative assessment can be useful. Both PARCC and SBAC plan to offer optional interim assessment and tools for formative assessment in addition to their mandatory summative assessments. It is expected that the tools for formative assessment will help teachers obtain information about student progress in acquiring the knowledge and skills called for by the CCSS and to suggested steps that may be taken to help students reach those standards. Although it remains to be seen

how effective the consortium efforts will be in aiding teachers' use of formative assessment, it is a step in the direction that any future test-based accountability system should have.

As was discussed above, test-based accountability systems should be used in tandem with qualitative methods to help teacher effectiveness. Systematic observations of teaching practice have to been shown to be especially important in providing teachers with actionable information about their teaching practices.

Summary and Conclusion

Test-based accountability systems have taken a variety of forms over the last fifty years or so. Such systems are seen as tools that can help improve education by clarifying goals and increasing motivation of students, teachers, and other educators by holding them responsible for student achievement outcomes. There is evidence that when accountability systems are put in place that test scores increase. The magnitude of the increases is generally modest. The gains on tests used for accountability tend to be larger than gains on low-stakes tests in the same content domains such as NAEP, which suggests that some portion of the gains is due to test score inflation caused by teaching to the test. Nonetheless, demands for test-based accountability are relentless.

There are reasons to hope that the tests used in the future will be better than current tests and will be aimed at the ambitious goals articulated in the CC SS. The two federally funded state consortia, PARCC and SMAC, promise not only to develop more adequate summative tests for accountability purposes, but also to develop a system of interim assessments and tools for teachers to use for formative assessments. It is hoped that the total packages developed by PARCC and SBAC will lead to improved school and teacher accountability which will, in turn, lead to higher student achievement and better preparation of students for college and the work place. The realization of these goals will take a concerted and sustained effort.

The quantitative approach to accountability through test-based accountability is not the only approach to holding schools and teachers accountable. Qualitative approaches using school visits and classroom observations have enjoyed wider use in some other countries than they have in the United States. The qualitative and quantitative approaches both have strengths and limitations. A hybrid system that capitalizes on the strengths of each approach is preferable to either of the two approaches alone. An accountability system that used high-quality tests tied to

ambitious content standards to measure student growth coupled with the tools to help teachers with formative assessments and the use of observational and other qualitative information to suggest steps teachers might take to improve their teaching would be a vast improvement over the current practice which bases sanctions on measures of current achievement status.

References

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

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. EPI Briefing Paper #278. Washington, DC: Economic Policy Institute, August 29.

Baker, E. L. & Linn, R. L. (2004). Validity issues for accountability systems. In S. Fuhrman & R. Elmore (Eds.). *Redesigning accountability* (pp. 47-72), New York: Teachers College Press.

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* (CRESST Policy Brief 5, pp. 1-6). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Betebenner, D. W., & Linn, R. L. (2010). *Growth in student achievement: Issues of measurement, longitudinal data analysis and accountability*. Princeton, NJ, K-12 assessment and Performance Management Center, ETS.

Cannell J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV. Friends of Education.

Center on Education Policy. (2010). *State test score trends through 2005-09, Part 2: Slow and uneven progress in narrowing gaps*. Washington, DC: Author. Available at www.cep.org.

Chadowsky, N. & Chadowsky, V. (2009). *State test score trends through 2005-09, Part 1: Rising scores on state tests and NAEP*. Washington, DC: Center on Education Policy. Available at www.cep.org.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. NBER Working Paper Series. Cambridge, MA, National Bureau of Economic Research.

Elmore, R. F. (2004). The problem of stakes in performance-based accountability systems. In S. H, Fuhrman & R. F. Elmore (eds.), *Redesigning accountability systems for education*. New York, Teachers College Press.

Haladyna, T. M., Nolan, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test pollution. *Educational Researcher*, *20*, 2-7.

Hall, J. N. & Ryan, K. E. (2011). Educational accountability: Qualitatively driven mixed-methods approach. *Qualitative Inquiry*, *17*(1), 105-115.

Hamilton, L. S., Stecher, B. M., Marsh, J. R., McCombs, J. S., Robyn, A., Russell, J. L., Naftel, S., & Barney, M. E. (2005). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states. Santa Monica, CA: Rand.*

Hanushek, E. A. (2012). Grinding the anti-testing ax: More bias than evidence in NRC panel's conclusions. *Education Next*, Spring, 2-8.
Hanushek, E. A. & Raymond, M. E. (2005). Does school accountability lead to improved student performance. J*ournal of Policy Analysis and Management*, *24*(2), 297-327.

Haut, M. & Elliott, S. W (Eds.). (2011). *Incentives and test-based accountability in education.* Committee on Incentives and Test-based Accountability in Public Education, National Research Council. Washington, DC: National Academies Press.

Jacob, B.A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89, 761-96.

Jacob, B.A. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments* (Working Paper No. 12817). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J. & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. NBER Working Paper No, 14607.

Kane, T. J. & Staiger, D, O. (2012). Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. *MET Project Research Paper*. Seattle WA. Bill and Melinda Gates Foundation.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources*, *37*(4), 752-777.

Koretz, D. (2008a). Further steps toward the development of an accountability-oriented science of measurement. In K. E. Ryan & L. A. Shepard (eds.), *The Future of test-based educational accountability*, pp. 72-91. Mahwah, NJ: Lawrence Erlbaum Associates.

Koretz, D. M. (2008b). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Koretz, D. & Barron, S. L. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS).* MR-1014-EDU. Santa Monica: Rand.

Koretz, D. & Bequin, A. (2010). Self monitoring assessments for educational accountability systems. *Measurement*, *8*, 92-109.

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *The perceived effects of the Maryland School Performance Assessment Program*. CSE Technical Report No. 409. Los Angeles: UCLA Center for the Study of Evaluation.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 31*(7), 3–13.

Linn, R. L. (2005). Issues in the design of accountability systems. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education* (pp. 78-98), Vol. 104, Part I.

Linn, R.L. (2008). Educational accountability systems. In I. E. Ryan and L. A. Shepard (eds.), *The future of test-based accountability*. New York: Routledge.

Linn, R. L., Graue, M. E, & Sanders, N. M (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*, No. 3, 5-14.
National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

Nolan, S. B., Haladyna, 9–15, T. M., & Hass, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice*, *11*(2).

O'Day, J. & Smith, M. (1993). Systemic reform and educational opportunity. In S. H. Fuhrman (Ed.), *Designing coherent educational policy: Improving the system*. NY: Jossey-Bass.

Perie, M., Marion, S., Gong, B., Wurtzel, J. (2007). The role of interim assessments in a comprehensive assessment system: A policy brief. Available at http://www.nciea.org/.

Popham, W. J. (2008). *Transformative Assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Porter, A. C., Chester, M. D. & Schlesinger, M. D. (2004, June). Framework for an effective assessment and accountability program. Teachers College Record, 106 (6), 1358-1400.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Smith, M. & O'Day, J. (1991). *Putting the pieces together: Systemic school reform. CPRE Policy Brief.* New Brunswick, NJ: Eagleton Institute Politics.

Stecher, B. M., & Hamilton, L. S. (2002). Putting theory to the test: Systems of educational accountability should be held accountable. *Rand Review, 26*(1), 16–23.

Webb, R., Vulliamy, G., Hakkinen, K., Hamalainen, S. (1998). External inspection or school self-evaluation? A comparative analysis of policy and practice in primary schools in England and Finland, *British Educational Research Journal*, *24*(5), 539-556.

Yen, W. M. (2009). *Growth models for the NCLB growth model pilot*. Princeton, NJ: Educational Testing Service.