Mapping the Landscape of High-Stakes Testing and Accountability Programs

Margaret E. Goertz, Mark Duffy

Theory Into Practice, Volume 42, Number 1, Winter 2003, pp. 4-11 (Article)

Published by Ohio State University College of Education

Margaret Goertz
Mark Duffy

Mapping the Landscape of High-Stakes Testing and Accountability Programs

The No Child Left Behind (NCLB) Act of 2001 requires states to test more, set more ambitious improvement goals for their schools, and increase sanctions for schools that fail to meet these goals. To gain an understanding of the potential impact of the new law, this article describes the types of state assessment and accountability policies that were in place at the time the U.S. Congress enacted the NCLB Act, and how selected school districts in eight states responded to these policies. It concludes by identifying four challenges facing states and school districts as they implement the NCLB Act.

Introduction

In the 1990s, all 50 U.S. states embarked on education initiatives related to high standards and challenging content. A central focus of these efforts was to establish a common set of academic standards for all students. Other components of these standards-based reforms included assessments that measure student performance and accountability systems that are at least partially focused on student outcomes. Although assessment has always been a critical component of the education system (Glaser & Silver, 1994), the growing focus on standards and accountability has dramatically changed the role of tests in the lives of students, their teachers, and their schools. While teachers continue to use the results of classroom and other types of tests to plan instruction, guide student learning, calculate grades, and place students in special programs, policy makers are turning to data from large-scale statewide assessments to make certification decisions about individual students, and to hold schools and school districts accountable for the performance and progress of their students.[1]

Provisions in the federal government's Title I program have reinforced the role of assessment in standards-based reform. Title I of the Improving America's Schools Act (IASA) of 1994 required states to develop high quality assessments aligned with state standards in reading and mathematics in one grade per grade span (elementary, middle, and high school), and to use these data to track student performance and identify low-performing schools. The most recent amendments to Title I, contained in the No Child Left Behind (NCLB) Act of 2001, give even greater prominence to state assessment. The law expanded state testing requirements to include every child in grades 3 through 8 in reading and mathematics by the 2005-2006 school year, and in science by 2007-2008. These assessments must be aligned with each state's standards and allow student achievement to be comparable from year to year. The results of the tests will be the

Margaret Goertz is a professor of education at the University of Pennsylvania; Mark Duffy is an educational consultant in New Jersey.

THEORY INTO PRACTICE, Volume 42, Number 1, Winter 2003

primary measure of student progress toward the achievement of state standards. States will hold schools and districts accountable for "adequate yearly progress" toward the goal of having all students meet their state-defined "proficient" levels by the end of school year 2013-2014. Students attending Title I schools that fail to make adequate progress are given the option of transferring to other public schools or receiving supplemental educational services outside the school. Title I schools that fail to improve over time can be restructured, converted into charter schools, or taken over by their district or state.

This article describes the kinds of assessment systems states had in place when the NCLB Act was enacted and how states used these data.[2] We then briefly discuss how a sample of districts in eight states responded to their state assessment and accountability policies. The article ends with a set of issues that states, districts, and schools face as they address the multiple policy demands placed on their assessment systems by the new federal legislation.

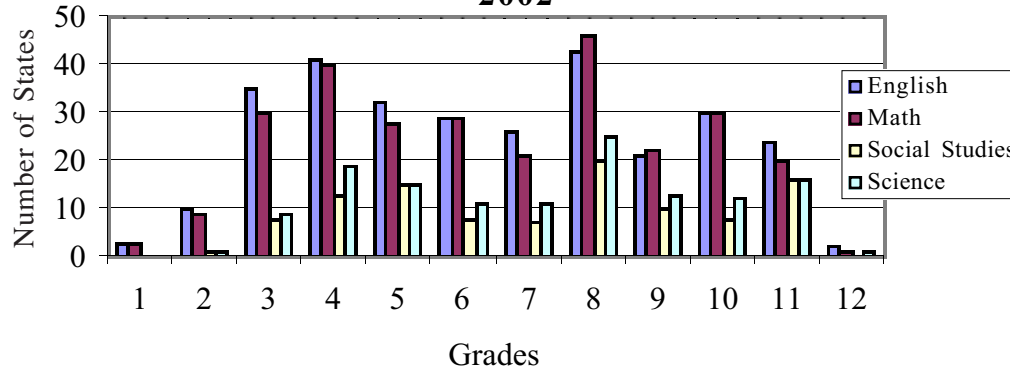
State Assessments: Form and Function

Form

By 2001, 48 states had implemented statewide assessments in reading and mathematics. We define a statewide assessment as one that is required by the state and does not allow for district discretion in the selection of a test. The other two states, Iowa and Nebraska, required their districts to test students in specified grades or grade spans, but left the choice of assessment instrument to the locality. These statewide assessment systems varied widely, however, in the subject areas and grade levels assessed, type of test used, and use of test data.

The IASA required that states test students at least once during each of three grade spans: third-to-fifth; sixth-to-ninth; and tenth-to-twelfth. But states assessed students considerably more often, with some states testing students in almost every grade. Thirteen states and the District of Columbia tested consecutive grades between grades 2 or 3 and at least grade 8 in the same subject areas using the same assessment, as required by the NCLB Act. Another three states tested consecutive grades between grades 2 or 3 and 8 in different subjects and/or using multiple assessments. The other 32 states tested students in only one or two grades per subject in elementary school, middle school, and high school. States most often tested students in grades 3, 4, 5, 8, and 10 (Figure 1). All 48 states with state assessment systems tested students in mathematics and English/language arts or reading in some mix of grades. Fewer states tested writing (31), science (34), and social studies (29).

Figure 1
Subjects and Grades Assessed Across the States: 2001-2002



Source: Data from *Education Week* ("Quality Counts," 2002)

State assessment systems included a mix of tests and testing formats. In 1999-2000, 29 states administered a combination of criterion-referenced and norm-referenced tests.[3] Seventeen states used only criterion-referenced tests, and two states used only norm-referenced tests. *Education Week* ("Quality Counts," 2002) reports that many state assessments incorporated open-ended as well as multiple-choice items. While only two states, Kentucky and Vermont, included portfolio assessments (which were judgments of students' best classroom work) in their state systems, nearly all states reported using extended response items in English (defined as written responses of at least one paragraph). Eighteen states included multiple-step problems that required students to explain or show their work in their mathematics assessments. Many states included other components to measure student performance, such as local assessments and non-cognitive measures (e.g., attendance, dropout, and graduation rates). However, only one state in five had a local testing requirement, just one result of the expansion of state testing programs. A growing number of states are requiring districts to assess early literacy skills as a means of identifying students who need help in reading in the primary grades.

Function

States use test results for varied and multiple purposes, including student diagnosis or placement, student promotion, high school graduation, school and district performance accountability, and program assessment. The use and consequences of the tests determine the level of "stakes" associated with the assessment system. Tests carry high stakes for students, for example, when they are used to assign students to schools, programs, or classes based on their achievement level (tracking), to make promotion decisions, and/or to determine whether a student will receive a high school diploma (Heubert & Hauser, 1999). Tests have high stakes for schools when low test scores or failure to show student progress are a major criterion in a district's or state's decision to intervene in or take over the administration of a school.

The accountability systems that emerged in the 1990s as part of the standards-based reform movement shifted the focus of accountability from education inputs to educational outcomes, and from school districts and students to schools. State accountability systems were designed to create incentives for schools to focus on student achievement and continuous progress. While state assessments are at the center of these accountability systems, the type and strength of incentives is determined by how states measure student performance and adequate progress, who sets what goals for the system, and the consequences of meeting (or not meeting) these goals. State accountability systems vary along all these dimensions.

Public reporting is the most basic form of accountability, and the most common. Reporting of assessment results allows the public to become aware of how a school and its students are achieving based on test scores and other data. The public can use this information to demand improvement in their schools, or possibly to choose alternate schools for their children. All 50 states currently produce or require local school districts to produce and disseminate district or school report cards (CCSSO, 2000). The report cards contain, at a minimum, student performance on state and/or local assessments. In 2001, only 17 states disaggregated performance by student characteristics, such as race/ethnicity and poverty, a requirement of the NCLB Act ("Quality Counts," 2002).

In 2000, 13 states used public reporting as their primary accountability mechanism. A few other states allowed districts to establish criteria for school performance, but used strategic plans or district and school improvement plans to hold districts accountable for student performance. The majority of states, however, set performance goals for schools or school districts and held these units directly accountable for meeting these outcome goals. These states also established rewards for meeting or exceeding state goals, sanctions for not meeting their targets, or both. The states' performance goals varied, however, along several dimensions, including how performance is measured and whether the performance goal is fixed or relative.

The IASA called for states to establish at least three levels of student performance (advanced, proficient, and below proficient) on state assessments to show how well students were mastering the material outlined in state content standards. Although

the terminology has been changed from *below proficient* to *basic* in the NCLB Act, the three levels of achievement are still required. Nearly all of the states with statewide assessments had student performance levels in place for the 2000-2001 school year. There was a wide variation, however, in school performance goals across the states. State targets appeared to vary along three dimensions: (a) the expected level of student performance (e.g., which test cut scores states use to define basic or proficient performance); (b) the percentage of students that schools had to get to these standards; and (c) the length of time schools were given to meet these goals.

Once states have established performance goals, they must determine how they will measure annual progress toward these goals. The IASA required states to define what they considered "substantial and continuous progress" toward performance goals; using these definitions of adequate yearly progress, states identified schools and districts in need of improvement. The 33 states with performance-based accountability systems used at least one of the following approaches in 2000 to measure school progress:

1. achieve a performance threshold or thresholds to make satisfactory progress (absolute target);
2. meet an annual growth target that is based on each school's past performance and often reflects its distance from state goals (relative growth); or
3. reduce the number or percentage of students scoring in the lowest performance levels (narrowing the achievement gap).

Under the NCLB Act, schools must show incremental and linear progress toward the attainment of academic proficiency in 12 years. In addition, states must develop separate progress goals for subgroups of students, including economically disadvantaged students, students from major ethnic and racial groups, students with disabilities, and limited English proficiency students, as well as all public school students.

Most states direct rewards and sanctions to the school level. Twenty-eight states and the District of Columbia provided assistance to low performing schools in 2001, while 18 offered rewards and 20 levied sanctions. Such sanctions included closure (9 states), reconstitution (15 states), and student transfers (11 states) ("Quality Counts," 2002). The majority of states, however, required schools to develop improvement plans as an initial step in addressing areas of weakness and creating change.

More and more states have developed high-stakes accountability systems for students as well as for schools. Eight states have enacted promotion policies for students in the elementary and middle grades that incorporate state test scores. By 2008, high school students in 28 states will have to pass a state-administered test in order to graduate from high school, an increase of ten since 1996-97. In another seven states, student performance on a state assessment may be noted on a student's transcript or diploma, but passing a state test is not required for graduation. Most state high school tests assess a student's general knowledge of English/language arts and mathematics, and often science and social studies as well. While most of the graduation tests implemented in the 1980s and early 1990s focused on basic skills, many of these states are in the process of revising their high school assessments so they will measure more rigorous content. Eleven states are administering or developing end-of-course high school examinations, which are curriculum-based assessments of specific high school courses, in addition to, or in lieu of, more general competency tests. Only six of these states will require students to pass these end-of-course examinations to graduate from high school, however.

District Response to High-Stakes Testing and Accountability Systems

With its focus on results and consequences for students and schools, performance-based accountability of the type embodied in the NCLB Act has developed much greater public visibility—and with it, controversy—than the more traditional input-based accountability systems. On one hand, researchers and reporters alike have lauded the success of states like North Carolina and Texas in raising the academic performance of their students and narrowing the performance gap between White students and students of color (cf. Grissmer & Flanagan, 1998; Grissmer, Flanagan, Kawata, & Williamson, 2000). On the other hand, civil rights advocates in New York and Texas have charged in court that high-stakes accountability systems discriminate

against poor and minority students; and researchers and educators have argued that high-stakes testing and accountability systems narrow curricula and limit teacher flexibility and creativity (cf. McNeil, 2000).

The Consortium for Policy Research in Education (CPRE) conducted a longitudinal study of standards-based reform in eight states and 23 school districts between 1996 and 1999 to examine how schools and school districts responded to standards-based testing and accountability policies.[4] The researchers found that well-developed state and local standards and performance-based accountability systems provided a clear focus to districts, schools, and teachers regarding the attainment of student outcomes, and created incentives for school and school system improvement (Goertz, 2001; Massell, 2001).

State and school district standards, coupled with aligned assessments, set clear expectations for student achievement in the study districts, and guided curriculum development, school improvement planning, local assessments, and professional development. Most of the districts required schools to develop improvement plans that identified school-level needs and strategies for achieving district goals. These plans were often used to identify teacher professional development needs, justify the expenditure of Title I and other discretionary funds, and/or plan curriculum and instruction. One of the most striking trends across the districts was the remarkably high level of attention paid to using data on student outcomes to inform these decisions.

As the conversation about school improvement became informed by an understanding of the data about students' learning, educators and district staff pressed for more and better measures of student performance. Some districts extended student testing beyond their state's assessment system, adopting local tests in grades not tested by the state. In some cases these were standardized, norm-referenced tests. In other cases, districts used district-designed end-of-unit tests and teacher-generated running records to measure student progress against district goals (which were aligned with state goals). Still other districts sought multiple measures of student performance, supplementing commercial tests with performance assessments. While a major purpose of district assessments was to measure the continuous progress of students toward district and/or state goals and provide instructional feedback to teachers and schools, other reasons for district testing activity emerged. These included: providing information on individual students for parents, teachers, and/or special programs identification (e.g., special education); providing external validation of student performance, especially in the basic skills areas, through the use of national, norm-referenced assessments; evaluating programs (such as Title I, state compensatory education, gifted and talented, and vocational education); and reinforcing the form and language of performance assessments in instruction in an effort to bridge the gap between assessment and instruction.

It appears, however, that some states were prepared to hold students accountable for performance without necessarily holding their schools accountable. While public reporting focused a spotlight on low-performing schools, educators in the CPRE study generally faced few formal consequences for not meeting school, district, and/or state performance goals beyond those imposed by the state. Districts used performance data to provide support rather than to impose sanctions. When formal consequences existed, they fell more heavily on students and principals than on teachers. The stakes were the highest for students, particularly those in states and/or districts with high school graduation and/or promotion requirements. At the school level, the principal was the primary focus of accountability. But principal accountability was generally ill-defined. Few districts had set formal performance goals or consequences for principals, and some respondents spoke of consequences in hypothetical terms. Typically, principals whose schools were posting low scores would be called in to explain their past actions, required to submit new improvement plans, and monitored closely. If their school was identified as "in crisis," media coverage was intensive. These principals might be shifted to another school or placed on a different type of assignment. They were rarely fired or demoted as the result of low student achievement, particularly if a district faced a shortage of principals.

Teachers also faced few consequences for poor student performance. Many of the teachers in the CPRE study reported that their districts and states held schools more accountable for student

performance than teachers. While some districts looked at student performance, teachers were generally held accountable for the delivery of their instruction and for meeting their professional development goals. While several districts in the study were looking more closely at instructional practice, few had clear performance standards or consequences linked to these evaluations. Consequences for poor performance appeared limited to professional development, coaching, and mentoring.

States, in turn, have been slow to apply sanctions to schools that fail to improve student performance, in large part because they lack the fiscal and human capacity to provide the necessary support. California, for example, reduced the number of low-performing schools subject to state takeover in 2002 because the potential numbers were so large (Sanders, 2002). In that same year, Maryland had reconstituted only four of the 107 schools that have been identified as reconstitution eligible, and North Carolina targeted its limited resources on a small number of failing schools.

High-Stakes Assessment and Accountability in the 21st Century

The "soft sanction" approach taken by many states and districts may be a thing of the past under the NCLB Act. The federal government now requires states and school districts to test more and report more, to set more ambitious improvement goals, and to apply sanctions more quickly to schools that do not meet these goals. These provisions will have major implications for state and district assessment and accountability policies.

First, most states must expand the size and scope of their assessment programs, in some cases testing six additional grade levels. This expansion has major cost and capacity consequences for states. Although the federal government has promised aid to cover the expense of developing these assessments, states must absorb the additional cost of administering and scoring the tests. Some states are finding needed funds by eliminating tests in science (at least temporarily), social studies, and other subjects not covered by the NCLB Act. This could have the unintended consequence of narrowing educators' focus on the tested subjects of reading and mathematics. The impact of expanded state testing on districts will be mixed. As discussed earlier, many districts already test additional grades, so the increased test burden on students may be limited. This is particularly true in districts that have back-filled with tests that are similar to state assessments in type and coverage. Districts that use more performance-based or instructionally based assessments face a dilemma, however. They can continue their assessment programs and test students more, or, facing substantial test burden, they may eliminate their local tests. Some California districts chose the latter option when the state implemented a grade-by-grade assessment in the late 1990s. District educators expressed concerns, however, that the format of the state assessment signaled a movement away from prior state, and current district, reform efforts.

This growing reliance on a single test raises a second set of issues: whether one test can serve multiple purposes. Policy makers expect one assessment system to provide indicators of the performance of the education system, hold schools and educators accountable for their performance, certify student performance as students move from grade to grade or out of the K-12 education system, motivate students to perform better and teachers to change their instructional content and strategies, and aid in instructional decisions about individual students (McDonnell, 1994). Assessment experts, however, question whether one test, no matter what the format, can address these multiple needs. For example, scores on assessments that are best suited for classroom instruction, such as portfolios, are difficult to aggregate on a district basis for accountability purposes. Performance-based and open-response items are better suited than multiple-choice tests in measuring complex skills and understanding. But, for a test of equal testing time, multiple-choice tests produce more reliable scores for individual students.

Third, while states will face many technical and political problems in responding to the stronger and more prescriptive accountability provisions of the NCLB Act, the law does take steps to bring student and adult accountability into greater balance. With a 12-year timeline, and a focus on the performance of subgroups of students, many more schools will be identified as not meeting state improvement goals.

The threat of sanctions, which range from school choice to school takeover, will get the attention of teachers and principals. The question then becomes, will educators pay attention to the right kinds of student performance data, which is a function of the quality and appropriateness of the test, and will they know how to act on that data?

This question leads us to our fourth and final issue: the capacity of the system to support change in practice. Research on school-based performance awards programs shows that clear goals and incentives are necessary, but not sufficient, to motivate teachers to reach their school's student achievement goals. Teacher motivation is also influenced by the presence of various capacity-building conditions, such as meaningful professional development. In addition, teacher knowledge and skills related to improved instruction are important (Kelley, Odden, Milanowski, & Heneman, 2000). Yet, the assessment and accountability provisions in the new federal legislation, like earlier state and federal policies, emphasize accountability over capacity-building. States and districts need knowledge, human resources, and financial resources to turn around poorly performing schools.

Research on capacity-building activities in the CPRE study states and districts identified some promising strategies. At the state level, these included creating decentralized support systems involving individuals and organizations that work directly with schools, nurturing professional networks of teachers and other education experts, providing curriculum frameworks and other curricular materials that included examples of standards-based instruction, and producing professional development and training standards (Massell, 1998). District strategies included enhancing teacher professionalism, curriculum reform aligned to state standards, data-driven decision making, and assistance targeted on low performing schools (Massell, 2000; Massell & Goertz, 2002). But states and districts reported having insufficient resources to help the number of schools that have, or that will be, identified as in need of improvement.

Conclusion

High-stakes testing and accountability policies are here to stay, at least in the near future. The challenge for policy makers and practitioners is to make the system work in ways that benefit students and their teachers. Well-designed assessments and accountability systems can focus attention on schools and students who need the most help, motivate students and educators, and foster the development of better curriculum and instruction. But policy makers must recognize the limits as well as the promise of such policies.

Notes

1. We use the terms *assess/assessment* and *testing/test* interchangeably in this article.
2. These data are drawn from a 50-state survey of state assessment and accountability policies conducted in 2000 by the Consortium for Policy Research in Education of the Graduate School of Education at the University of Pennsylvania (Goertz & Duffy, 2001), and updated with data reported by *Education Week* (2002).
3. Criterion-referenced tests measure knowledge and skills that are specific to a state or district, while norm-referenced tests measure the knowledge and skills of students across the country.
4. The eight states—California, Colorado, Florida, Kentucky, Maryland, Michigan, Minnesota, and Texas—were selected to represent a range in the age, stability, and type of state accountability policies. Together these states educate approximately 40% of U.S. public school students. The 23 school districts, three per state except for California, were selected for their activism in school improvement and standards-based reform but were demographically diverse.

References

Council of Chief State School Officers (CCSSO). (2000). *State accountability reports and indicators reports*. Retrieved November 2000 from www.ccsso.org/pdfs/AccountabilityReport2000.pdf.

Glaser, R., & Silver, E. (1994). Assessment, testing and instruction: Retrospect and prospect. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 393-419). Washington, DC: American Educational Research Association.

Goertz, M.E. (2001). Standards-based accountability: Horse trade or horse whip? In S. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (pp. 39-59). Chicago: National Society for the Study of Education.

Goertz, M.E., & Duffy, M. (2001). *Assessment and accountability systems in the 50 states: 1999–2000* (Research Rep. No. RR-046). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

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

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND.

Heubert, J.P., & Hauser, R.M. (1999). *High stakes testing for tracking, promotion and graduation*. Washington, DC: National Academy Press.

Kelley, C., Odden, A., Milanowski, A., & Heneman III, H. (2000). The motivational effects of school-based performance awards. *CPRE Policy Briefs* (No. RB-29). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Massell, D. (1998). State strategies for building local capacity: Addressing the needs of standards-based reform. *CPRE Policy Briefs* (No. RB-25). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Massell, D. (2000). The district role in building capacity: Four strategies. *CPRE Policy Briefs* (No. RB-32). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Massell, D. (2001). The theory and practice of using data to build capacity: State and local strategies and their effects. In S. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (pp. 148-169). Chicago: National Society for the Study of Education.

Massell, D., & Goertz, M.E. (2002). District strategies for building capacity. In A.M. Hightower, M. Knapp, J.A. Marsh, & M.W. McLaughlin (Eds.), *School districts and instructional renewal* (pp. 43-60). New York: Teachers College Press.

McDonnell, L.M. (1994). *Policymakers' views of student assessment*. Santa Monica, CA: RAND.

McNeil, L.M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. London: Routledge.

Quality counts 2002: Building blocks for success [Special Report]. (2002, January 7). Bethesda, MD: *Education Week*.

Sanders, J. (2002, May 1). School takeover threat relaxed. *Sacramento Bee*. Retrieved May 3, 2002, from www.sacbee.com.