The Future of Test-Based Educational Accountability

Edited by Katherine E. Ryan and Lorrie A. Shepard

First published 2008
by Routledge
270 Madison Ave, New York, NY 10016

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Typeset in Minion by Wearset Ltd, Boldon, Tyne and Wear
Printed and bound in the United States of America on acid-free paper by
Edwards Brothers Digital Book Center

Library of Congress Cataloging in Publication Data
Ryan, Katherine E.
The future of test-based educational accountability/Katherine E. Ryan, Lorrie A. Shepard.
p. cm.
Includes bibliographical references.
1. Educational tests and measurements–United States. 2. Educational accountability–United States. I. Shepard, Lorrie A. II. Title.
LB3051.R98 2008
379.1'58–dc22
2007051799

ISBN10: 0-8058-6470-9 (hbk)
ISBN10: 0-203-89509-6 (ebk)

ISBN13: 978-0-8058-6470-0 (hbk)
ISBN13: 978-0-203-89509-2 (ebk)

Shepard, L. S. (2008). A Brief History of Accountability Testing, 1965-2007. In K. E. Ryan, L. S. Shepard (Eds.), *The Future of Test-Based Educational Accountability* (pp. 25-46). New York, NY: Routledge.

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

2
A Brief History of Accountability Testing, 1965–2007

Lorrie A. Shepard

Standardized testing has a long history in the United States, and testing is more salient in the U.S. education system than it is in any other country (Resnick, 1982). Predominantly, tests have been used to make decisions about individual students, especially to place students in special programs and to select students for college (Goslin, 1963). Accountability testing—focused on judging the quality of schools—is a more recent phenomenon, but it has its roots in the technology of IQ testing and the ardent belief among Americans that tests can scientifically determine merit and worth.

A hundred years ago, Goddard and Terman brought IQ tests to America in a climate of Social Darwinism and survival of the fittest. They were strict hereditarians who believed that mental tests could be used to measure innate ability and thereby assign students to education levels and even to their jobs later in life (Terman, 1916). Although beliefs about fixed, innate intelligence lost favor with scientists many decades ago, these ideas continued to have great sway with the public and with educators. Indeed, educational reformers at the end of the twentieth century specifically sought to challenge these endemic attitudes and practices by announcing that "*all* students can learn" and calling for "high standards for *all* students."

A less visible strand of educational testing, with an even longer history, focuses on the use of tests to evaluate the quality of schooling—though without voicing the notion of accountability. In 1845, Massachusetts State Superintendent of Instruction, Horace Mann, pressured Boston school trustees to adopt written examinations because large increases in enrollments made oral exams unfeasible. Long before IQ tests, these examinations were used to classify pupils (Tyack, 1974) and to put comparative information about how schools were doing in the hands of a state-level authority (Resnick, 1982). In the 1890s, in hopes of spending more time on richer subject matter (Cronbach et al., 1980), Joseph Rice administered spelling tests to 30,000 students and found no difference between students taught spelling for 15 minutes per day versus those taught for 30 minutes. Beginning in 1908, Thorndike and his students developed hundreds of achievement tests that then were implemented on a wide scale through

university-based bureaus of cooperative research established to conduct school surveys (Cook, 1941).

Three general points are worth noting about these precursors to today's school accountability. First, achievement testing programs grew up alongside IQ testing, relied on the same statistical techniques for test construction and for evaluating test quality, and suffered from the same limitations. Second, both Mann and Thorndike instituted testing programs because they had already concluded that schools were failing (U.S. Congress Office of Technology Assessment, 1992); gathering data would help them promote school reform. Third, focusing attention on standardized tests often produces perverse results, as Rice discovered when educators spent *more time* on spelling after his study, despite his finding that more time made no difference (Cronbach et al., 1980).

Before 1970, testing programs were mostly local but relied on standardized test batteries available from commercial test publishers. Results from individual aptitude and achievement tests were used to make high-stakes decisions about individual children that could have crushing self-fulfilling consequences (Heller, Holtzman, & Messick, 1982), but test scores were rarely used to make judgments about individual schools. All of that changed relatively abruptly 40 years ago with the emergence of large-scale assessment systems and their use for school accountability. In this chapter, I trace the history of state and national assessments and the origins of educational accountability with its cycles of revision from minimum competency testing, to basic skills testing, to standards-based reform.

It All Started With Title I

Title I of the Elementary and Secondary Education Act (ESEA) of 1965 launched the development of the field of educational evaluation and the school accountability movement. The 1960s are remembered as a time of social unrest, when issues of equality were paramount. It was also a time when the federal government shifted its management practices to focus on cost–benefit analysis and production outcomes (Resnick, 1980), and when in many sectors of government and social services, evaluation research became the handmaiden to public policy (Cronbach et al., 1980). In education, evaluation of post-Sputnik curriculum projects predated Title I, but it was the ESEA mandate for evaluation of every Title I and Title III project that literally created the field of educational evaluation (Worthen & Sanders, 1973). The American Educational Research Association began a monograph series in 1967 to disseminate the latest thinking in evaluation theory, and several educational evaluation organizations and journals date from this period. The most important aspect of Title I evaluation, however, was the new implied contract with local districts whereby federal dollars would be spent on education in exchange for evidence of program effectiveness. It was this bargain—which tied funding to measured outcomes—that created the accountability movement.

The evaluation provisions in Title I came about because Senator Robert Kennedy doubted whether school administrators understood the problems of or knew how to provide effective programs for disadvantaged children. He expected that evaluation data could be used by parents as a "whip" or a "spur" to leverage changes in ineffective schools (Halperin, 1975; McLaughlin, 1975). Kennedy's intention was almost identical to present-day accountability rhetoric. For example, in Colorado, Governor Bill Owens pushed for the development of school report cards because he believed that giving low grades to low performing schools would cause the school community to rally. Parents and business leaders would become involved and make sure that school performance improved (Owens, 1999).

Evaluations of the early 1970s, however, were quite benign with low stakes compared to today's context. The Colorado Accountability Act of 1971, for example, required only that districts conduct evaluations of their programs and report to their constituencies, causing one evaluation expert to grouse that requiring educators to conduct their own evaluations was like "asking banks to conduct their own audits" (Worthen, 1974, p. 26). Similarly, because of the need to mitigate the threat of federal intrusion, early Title I evaluations were "chaotically diverse" and could not be aggregated so as to inform policy decisions (Cronbach et al., 1980, p. 33). A few years later, when it was recognized that little could be learned from a multitude of different tests, score metrics, and research designs, a more uniform system of reporting was imposed, which led to a huge burgeoning in the amount of standardized testing (Tallmadge & Wood, 1978).

The National Assessment of Educational Progress: From Achievement Census to Policy Instrument

The National Assessment of Educational Progress (NAEP), begun in 1969, was part of the same general trend toward large-scale data gathering, but NAEP was intended to be an information source and neutral monitor, not an accountability device. Over time, however, as accountability pressures and political interest in test scores intensified, the independence and neutrality of NAEP would be increasingly challenged.

NAEP Beginnings

Ralph Tyler, NAEP's primary architect, called it a census-like data system and likened its purpose to the collection of health statistics on the incidence of heart disease and cancer for different age and occupational groups. Tyler (1966) specifically distinguished this large-scale use of evaluation data—"to help in the understanding of educational problems and needs and to guide in efforts to develop sound public policy regarding education"—from the kinds

of information needed for individual pupil appraisal, teaching decisions, and even curriculum evaluation.

The independence of the National Assessment from specific educational programs or political jurisdictions was further assured by both its data collection methods and administrative structure. Matrix sampling of test items within a content domain would help to ensure that the assessment provided a much broader representation of subject matter fields than was possible on traditional standardized tests (but see Stake's (2007) perspective on the limitations of an assessment conceived by measurement specialists rather than curriculum scholars). At the same time, students were sampled to represent regions of the country and urban, suburban, or rural districts rather than specific states or districts. The contract for overseeing the National Assessment was given to the Education Commission of the States (ECS), a nonprofit organization of governors, chief state school officers, and legislators, again to buffer NAEP from the specter of federal control of education. Interestingly in the beginning, the one political purpose intended for NAEP—again using the disease analogy—was to obtain more generous appropriations for education (Cronbach et al., 1980) because it was expected that the identification of problems would naturally bring more resources to bear in solving them.

Over time, the purpose (and correspondingly the characteristics) of NAEP have become increasingly more politicized, although still relatively immune from politics compared to state assessments. The very features of the assessment that had been designed to shelter it from politics were later blamed for the lack of public interest in the assessment's results and systematically targeted for correction. In 1983, the Educational Testing Service won the contract for NAEP away from ECS by proposing a significant redesign that would be more responsive to policymakers' needs (Messick, Beaton, & Lord, 1983). The frequency of assessments was increased, reporting by grade level rather than age was begun, background and program variables were added to help in interpreting results, and sophisticated scaling methods were introduced to produce a single summary score that could be more readily understood by the public.

NAEP and Comparative State Data

Efforts to increase the visibility and usefulness of NAEP occurred in the context of concerns about education that led at that same time to *A Nation at Risk: The Imperative of Educational Reform* (National Commission on Excellence in Education, 1983). In addition to appointing the Commission to study the quality of American education, Secretary of Education Terrel Bell and his successor, William Bennett, stimulated interest in comparative state education data by publishing their famous "Wall Charts." Annual Wall Charts provided data on student characteristics and education resources, such as per pupil

expenditure, but most heatedly they compared states on average ACT and SAT scores. Obviously, tests administered to non-representative samples of students could not be used to say anything about the quality of education in any state. But the flurry over Bennett's press conferences certainly generated enthusiasm for gathering state-by-state data using more legitimate means.

In 1987, a study group chaired by Lamar Alexander, former Governor of Tennessee, and directed by the Spencer Foundation's President, H. Thomas James, recommended that the NAEP assessment design be expanded to include state-by-state comparisons and possibly even district and school-level data (Alexander & James, 1987). When called upon to review the Alexander–James report, a National Academy of Education committee chaired by Robert Glaser expressed a few concerns but basically endorsed the idea that NAEP could be expanded and used as a "catalyst for school improvement." Specifically, the Glaser (1987) commentary cautioned (a) that future assessments, limited in the competencies they measure, might come "to exercise an influence on our schools that exceeds their scope and true merit" (p. 51) and (b) that "simple comparisons are ripe for abuse and are unlikely to inform meaningful school improvement efforts" (p. 59). Glaser's committee was optimistic, however, that by using more extended-response assessment formats, NAEP could serve "as a model of what students should know and how it should be assessed" (p. 47). Following from these recommendations, voluntary participation of states in the national assessment, called the NAEP Trial State Assessment project, was formally authorized by Congress in 1988. As anticipated, the availability of state comparisons greatly heightened the interest of policymakers and the media in assessment results.

NAEP as a Policy Instrument

In its 1988 reauthorization of NAEP, Congress implemented another recommendation of the Alexander–James and Glaser reports, creating a National Assessment Governing Board (NAGB) for the purpose of making NAEP more responsive to the concerns of various constituencies. One of the most visible and controversial acts of NAGB was to change the way that assessment results were reported. Instead of average scores and descriptive anchors showing what American students "could do," achievement levels were developed on the NAEP scales to show what students "should be able to do." The achievement levels, set through a judgmental process involving educators and lay citizens, were criticized in several evaluation reports (Shepard, Glaser, Linn, & Bohrnstedt, 1993; Stufflebeam, Jaeger, & Scriven, 1991; U.S. General Accounting Office, 1993). Beyond technical and validity problems, one of the main concerns was that judgmentally set standards—that varied dramatically from grade-to-grade and across subject areas and that departed dramatically from normative expectations of grade-level proficiency—would cause confusion and seriously mislead the public as to the meaning of assessment results.

Each time efforts have been made to increase the uses of NAEP results, debates have ensued about whether expansion would harm the integrity of assessment data. At issue are two chief concerns: (1) testing more often or in more jurisdictions increases costs, which if not adequately funded will likely reduce the substantive quality of the assessments; and (2) political attention to results could lead to the same kind of teach-the-test distortions that have affected state testing programs. In 1992, as standards-based assessments were being developed, the National Council on Education Standards and Testing called for a system of assessments that reserved for NAEP the role of program/system monitor while encouraging states, national professional associations, or consortiums of states to develop assessments that could be used for individual students. A key idea was to maintain the independence of NAEP so that it could be used to evaluate whether reported gains on assessments used locally for accountability purposes were accurate and thereby determine whether standards-based reforms were effective or ineffective in improving education. Checking on the validity of reported test score gains may have been what President George W. Bush had in mind when he proposed as part of the "No Child Left Behind" (NCLB) legislation that NAEP be used to confirm progress on state assessments. However, many feared that tying funding to outcomes on NAEP would undermine its independence, and as a result of this controversy, the language of the No Child Left Behind legislation was softened, requiring that states participate in NAEP but leaving unspecified how NAEP results would be used to check on the authenticity of achievement gains reported by state assessments.

The history of NAEP over several decades reflects a gradual shift from mere data collection, like the U.S. Census, to an increasingly powerful policy instrument used to garner attention and mobilize educational reform efforts. In this chapter, I pursue this theme of politicization of large-scale assessments, especially of state assessments, which have been much more dramatically affected. Before doing so, however, it is important to consider a larger change in the policy context, a change that shifted the reporting of assessment results from good news to bad news about public education.

The SAT Test Score Decline: Bad News About Public Education

During the 1960s and the nation's war on poverty, public education was viewed with approbation. The only criticism of education was that its benefits had not been extended to poor and minority children. The willingness of policymakers to invest in the Elementary and Secondary Education Act of 1965 was, in fact, a sign of their faith in the power of education to redress many of society's ills. Within a few years, however, the minimum competency testing movement was born in a political climate that had become hypercritical of education. Messick et al. (1983) offered several explanations for this change, including the Vietnam War and the disillusionment of the late 1960s. A very central cause of the decline in public opinion about education, however, was the famous SAT score decline.

In 1963, after two decades of steady or rising scores, SAT averages took a downward turn and continued downhill for the next 14 years. The loss over the entire period was dramatic: 49 points in verbal scores (one-half standard deviation) and 32 points in mathematics. A Blue Ribbon Panel commissioned by the College Board (1977) later found that two-thirds to three-quarters of the score decline was attributable to changes in the composition of the test-taking population during this period, that is, more women and minority group members were going to selective colleges and thus needed to take the test. Nevertheless, what the public remembered was the precipitous decline and the gist of the Panel's speculations about the causes of the smaller but real decline—too many electives instead of required courses, too much TV, and a decline in family participation in the learning process. In his analysis of the factors leading to the Minimum Competency Testing movement, Resnick (1980) cited as well public fears about rising unemployment and the tendency to blame the schools for lack of preparation.

Minimum Competency Testing

When the National Assessment first began, several states created their own state assessment programs modeled after NAEP with its emphasis on system evaluation rather than the performance of individual students. For example, in 1974 California stopped administering an off-the-shelf standardized test and developed the California Assessment Program using matrix sampling for the new purpose of "broad program evaluation rather than diagnostic assessment of individual students" (California State Department of Education, 1973, p. 1). Rhetoric surrounding the SAT test score decline, however, and concerns about an economic downturn quickly overtook the system-level data collection purpose of large-scale assessment and redirected efforts toward enforcement of minimum academic standards. By 1978, 33 states had taken action to mandate minimum competency standards for grade-to-grade promotion or high school graduation (Pipho, 1978). By 1980, all states had a minimum competency testing program or a state testing program of some kind (Baratz, 1980). By mandating state-administered tests and standards, legislators intended to improve the quality of schooling and "put meaning back into the high school diploma."

The minimum competency movement of the 1970s, like the accountability movement today, was driven by a business model. Wise (1978) identified the following management concepts adopted from business into the education sphere: accountability planning, programming, budgeting systems (PPBS); management by objectives (MBO); operations analysis; systems analysis; program evaluation and review technique (PERT); management information systems; and several additional planning and budgeting terms. A simplistic,

bottom-line mentality made it easy to rely on single test scores, like the Gross National Product, as sufficient indicators of system health. Policymakers in both periods gave relatively little attention to the intervening variables needed to achieve mandated ends. In 1978, Wise argued that minimum competency testing programs would fail to improve education because they lacked a "theory of education"; what today would be called a "theory of action." That is, legislators were mandating desired outcomes of schooling without having an understanding of how the mandate might or might not cause changes in curriculum and instruction that would in turn produce the desired outcomes.

The problem of setting performance standards—that is, determining the passing score for the test—also began with minimum competency testing and has continued unabated (Brickell, 1978; Glass, 1978). Because the testing program was intended to be the reform, not just measure its outcomes, minimum competency testing also marked the beginning of serious consequences attached to test results. The only differences between accountability testing then and now—and these differences are striking—were the levels of the standards (minimum standards then, world class standards now) and the content of the test. Figure 2.1 provides an illustration of the extremely low level of content included in minimum competency tests. For example, the mathematics items in this example are roughly at the third grade level according to present day curriculum standards.

Minimum competency graduation tests are still in place in some states, but the movement lasted less than a decade. By the time some slow moving states had developed and implemented a competency program, the movement was already judged by many to have failed in its efforts to improve the quality of education. The authors of *A Nation at Risk* (National Commission on Excellence in Education, 1983), which began the next wave of educational reform, specifically faulted minimum competency examinations as part of the problem, not part of the solution: "'competency' examinations (now required in 37 States) fall short of what is needed, as the 'minimum' tends to become the 'maximum,' thus lowering educational standards for all" (p. 2).

A Nation at Risk, Basic Skills, and the Excellence Movement

Among countless reports on education, *A Nation at Risk* (National Commission On Excellence in Education 1983) is perhaps the single most visible education policy report of the century. It blamed the mediocre performance of U.S. students and U.S. schools on neglect, low standards, and a dilute curriculum. Within two years of its publication, 30 national reports and 250 state reports had been issued on educational reform (Pipho, 1985), and nearly every state had introduced reform legislation. The excellence movement, launched by *A Nation at Risk*, sought to ratchet up expectations by reinstating course-based graduation requirements, extending time in the school day and school year, requiring more homework, and—most importantly—requiring more testing.

Consumerism

1. Group health insurance, offered by an employer, will cost you less than a health policy you purchase by yourself.

2. One must use credit to have a good credit rating. A good way to keep a satisfactory rating is:

   a. borrow money from friends
   b. make a budget to avoid using credit
   c. pay your bills promptly
   d. pay cash for everything you buy
   e. none of the above

3. Steve borrowed $200 from his bank. He repaid it in six monthly installments of $37.50 each. What was the "cost" in dollars?

   a. $15 b. $25 c. $37.50 d. $237.50

4. Match the letter of the consumer protection agency with the function it performs:

   a. Federal Trade Commission
   b. Better Business Bureau
   c. City Health Department
   d. Food and Drug Administration

   _____ investigates false claims in advertisements of nationally sold products

   _____ provides information regarding the reputation of local business firms

   _____ inspects public eating places and hospitals

   _____ analyzes foods, drugs and cosmetics suspected of being harmful for human use

5. A brand of cola is available in four bottle sizes. Which of the following bottles has the lowest price per ounce?

   a. 6 oz. at 36 cents
   b. 8 oz. at 42 cents
   c. 12 oz. at 56 cents
   d. 24 oz at $1.20

Mathematics

1. Which digit represents hundredths in: 1234.567?

2.
$$\begin{array}{r} 21\ 5/7 \\ +2\ 3/7 \\ \hline \end{array}$$

   a. 24 1/7 b. 23 1/2 c. 19 2/7 d. 16 1/7 e. 168/7

3. Chain-link fencing costs 59 cents a foot. Approximately how much will it cost for 50 feet of fencing?

   a. $10 b. $25 c. $30 d. $ 40 e. $2,950

4. How long should a roast cook if it weighs 5 pounds and must cook 20 minutes for each pound?

   a. 2 hours
   b. 1 hour and 20 minutes
   c. 2 hours and 40 minutes
   d. 1 hour and 40 minutes
   e none of the above

5. Express 15% as a decimal

   a. 15 b. 15 c. 1.5 d. 150 e. 0.15

Democratic process

Which of the following would you expect to find in a democratic society?

	Would	Would not
1. Joe Smith gives $5 each to vote for him	____	____
2. Citizens legally picket and protest a court decision	____	____
3. A group of people go to the city council to ask for an investigation of the mayor	____	____
4. Congress overrides a Presidential veto	____	____
5. A citizen is arrested for breaking a law that is not written down	____	____

Answers:

Consumerism: 1. True 2. c 3. b 4. a, b, c, d 5. c
Mathematics: 1. 6 2. a 3. c 4. d 5. b
Democratic process: 1. Would 2. Would 3. Would 4. Would 5. Would not

Figure 2.1 Examples of Low-Level Questions Typical of Minimum Competency Graduation Tests in the 1970s.

Although the rhetoric of the excellence movement called for "new basics" and a rigorous academic curriculum for all students, critics even at the time warned that reliance on quantitative rather than qualitative factors was more likely to ensure educational adequacy rather than excellence (Duke, 1985).

In retrospect, it may seem odd that the excellence movement, with its aversion for low standards, did not provoke a more thorough reexamination of the kinds of tests used to lead as well as measure the reform. Some states did forego their minimum competency tests, but even the new tests adopted in the mid 1980s were predominantly multiple-choice basic skills tests. It was not until the effects of high-stakes tests began to be evaluated that any doubt arose about whether rising test scores on limited tests could be trusted as evidence that achievement was improving. Initially, gains on these tests, mostly in reading, math, and writing (measured by multiple-choice questions) were applauded as evidence of the success of reforms. Popham (1987), for example, used the gains in percent passing the tests in five different states to show the effectiveness of "measurement-driven instruction."

Ultimately, however, there were several validity challenges to the rosy picture painted by steadily rising test scores. John Cannell (1987), a West Virginia physician, frustrated at discovering above average test scores reported for a patient with grave school difficulties, conducted a survey and found that all 50 states claimed to be performing above average on nationally normed tests. More systematic evidence from the National Assessment for the 1980s showed gains in basic skills, but the gains were not so great as those reported on state assessments. Moreover, trends on higher-order skills were either flat or declining (U.S. Congress, Office of Technology Assessment, 1992).

Prompted by complaints that "high-stakes" accountability tests were narrowing the curriculum and producing inflated test score gains, numerous studies were undertaken to examine the effects of testing on teaching and learning. Several large-scale surveys of teachers showed essentially the same patterns. Because of pressure to improve test scores, teachers reduced or eliminated time for non-tested subjects, spent considerable amounts of time practicing test-taking skills, and changed their instructional materials and activities to imitate test formats as closely as possible (Darling-Hammond & Wise, 1985; Rottenberg & Smith, 1990; Shepard & Dougherty, 1991). These practices, which reduced the curriculum to drill and practice for the test, were the most pronounced in schools and districts serving large numbers of poor and minority children (Madaus, West, Harmon, Lomax, & Viator, 1992). Other studies, designed to investigate the effect of such practices on learning, used independent measures to evaluate whether apparent learning gains were real (Koretz, Linn, Dunbar, & Shephard, 1991). Unfortunately, high levels of student performance on accountability tests could not be replicated on independent measures of the same content, suggesting that students drilled

constantly in preparation for the test lacked understanding of underlying concepts.

By the end of the 1980s, concerns about the huge increase in the amount of testing, as well as concerns about potential negative effects, prompted Congress to commission a comprehensive report on educational testing (U.S. Congress Office of Technology Assessment, 1992). Evidence of negative effects from high-stakes testing was sufficient to cause framers of the 1994 reauthorization of Title I to redirect substantially evaluation requirements that had theretofore driven the mandate for norm-referenced assessments. It would be a mistake to conclude, however, that policymakers, educators, and researchers all shared a common understanding of what had gone wrong with previous reforms. Researchers and teachers in subject matter fields were the most likely to be knowledgeable about research on the distorting effects of test-driven instructional decisions. Cognitive researchers, new to the assessment game, were aware of severe distortions caused by teaching to the test, but were inclined to believe that this problem could be solved by making better tests (Frederiksen & Collins, 1989; Resnick & Resnick, 1992). Policymakers, with little time for academic quibbles, were willing in many states to invest in the development of new forms of assessment, but at the same time continued to interpret the results from all different sorts of tests as if they were equally trustworthy.

Using Standards to Correct Previous Reforms

Just as *A Nation at Risk* was both a rejection and extension of minimum competency testing, so too were standards-based reforms of the 1990s both a rejection and extension of the recent basic skills reforms. Unlike the prior reform, which reaffirmed traditional curricula, the standards movement called for the development of much more challenging curricula: focused on reasoning and processes of inquiry, as well as content knowledge, and directed toward engaging students in using their knowledge in real-world contexts. Leading the way, the National Council of Teachers of Mathematics report on *Curriculum and Evaluation Standards for School Mathematics* (1989) expanded the purview of elementary school mathematics to include geometry and spatial sense, measurement, statistics and probability, and patterns and relationships, and at the same time emphasized problem solving, communication, mathematical reasoning, and mathematical connections rather than computation and rote activities.

As an extension of previous reforms, the standards movement continued to rely heavily on large-scale accountability assessments to leverage changes in instruction. In contrast to previous reforms, however, standards-based reformers explicitly called for a radical transformation of the substance of those assessments as a corrective for the distorting effects of existing high-stakes testing programs. Various terms such as *authentic*, *direct*, and

performance-based assessments were used in standards parlance to convey the idea that assessments themselves had to be reformed to reflect more faithfully how learning would be used in non-test situations.

A great many standards documents provided sample assessment tasks both to exemplify and to enact curricular reforms. For example, the Mathematics Sciences Education Board of the National Research Council developed a set of prototypes for mathematics assessment. Intended for fourth graders, the tasks illustrated how different education would have to be to build students' confidence as well as provide them with the proficiencies needed to do well. Consistent with the reform's intentions, the tasks called for connections with other academic areas, and promoted higher-order thinking by asking students to justify their answers, draw a picture to explain their solution, make predictions, and draw generalizations from their problem solutions. Similarly in science, assessment tasks devised to mirror the new standards required students to formulate a question, design and conduct scientific investigations, use tools for data collection, formulate and defend a scientific argument, evaluate alternative explanations on the basis of evidence, and communicate the results of a scientific study.

The standards movement also differs from earlier reforms in that it has been informed and guided by an underlying theory of teaching and learning drawn from the cognitive sciences. Learning is no longer thought to be a mechanical process of memorization and accumulation of information but is rather an active process that requires reasoning and sense making on the part of the learner. Correspondingly then, effective teaching involves creating the necessary social supports (activities and patterns of interaction) so that students become accustomed to working on interesting problems, reasoning aloud or explaining their thinking, and monitoring and reflecting on their own learning. These substantially more challenging curricular goals place heavy demands on both the content knowledge and pedagogical skills of teachers.

Given the ambitious and unprecedented aims of the reform, nearly every report involved in the creation of the standards movement said something about the need for capacity building. For example, Smith and O'Day (1990), who were among the early architects of standards-based reform, envisioned a reform that was systemic, affecting all aspects of the educational system. They emphasized the need for professional development for both pre-service and in-service teachers and for conditions that would enhance teacher professionalism. Similarly, the National Council on Education Standards and Testing (NCEST) called for the development of school and system "delivery standards," acknowledging that ambitious goals would not be met without shared responsibility for improvement at both the state and local levels (NCEST, 1992). The National Academy of Education Panel on Standards-Based Educational Reform verified that compelling research evidence existed to support much higher expectations for students under fundamentally different conditions of teaching and learning; but the Panel cautioned that the knowledge base was fragmentary. Considerably more development would be needed before these ambitious ideas could be implemented on a wide scale (McLaughlin & Shepard, 1995).

No Child Left Behind and the Standards Movement: Contradictions and Controversies

Standards-based reform, begun in the early 1990s, is the most enduring of test-based accountability reforms, yet the version of reform instantiated in No Child Left Behind contradicts core principles of the standards movement. Understanding the current accountability scene requires greater awareness of the competing versions of reform, wildly different performance standards, and conflicting findings about accountability's beneficial and harmful effects.

Competing Models

Although the standards movement, in principle, has a theory of action—what it would take to get from here to there—in fact, the reform cannot be said to rest on a sound theory if most of the participants do not have access either to the theory or to its enabling conditions. An honest look at the current scene suggests that there are at least two fundamentally different models, and perhaps many, underlying standards-based reforms, though all are dressed up in the same rhetoric.

We might label the original vision of systemic change put forth by Smith and O'Day (1990), Resnick and Resnick (1992), and Frederiksen and Collins (1989) as examples of the *teaching and learning* or *cognitive science* version of standards-based reform. In contrast, in a 1999 NRC report aimed at helping states develop new Title I assessment and accountability systems, Elmore and Rothman (1999) retrospectively describe a simplified, *basic* standards-based reform model:

> The centerpiece of the system is a set of challenging standards for student performance. By setting these standards for all students, states would hold high expectations for performance; these expectations would be the same regardless of students' backgrounds or where they attended school. Aligning assessments to the standards would allow students, parents, and teachers to monitor student performance against the standards. Providing flexibility to schools would permit them to make the instructional and structural changes needed for their students to reach the standards. And holding schools accountable for meeting the standards would create incentives to redesign instruction toward the standards and provide appropriate assistance to schools that need extra help.
>
> (pp. 2–3)

Elmore and Rothman concluded that such a model has failed to improve the system significantly because it omits direct efforts to "build the capacity of teachers and administrators to improve instruction" (p. 3).

Most politicians are unaware of the original learning theory and research-based arguments as to why it should be possible to hold all students to high "world class" standards. To the extent that policymakers subscribe to a theory of action, they are more likely to hold with Elmore and Rothman's basic model or to adopt a *high-stakes incentives* version of standards-based reform. For example, Hess (2002) argues for "minimum standards" and for what he calls "the coercive force of self-interest":

> High-stakes accountability systems link rewards and punishments to demonstrated student performance in an effort to transform the quality of schooling. Such systems press students to master specified content and force educators to effectively teach that content. In such a regime, school improvement no longer rests upon individual volition or intrinsic motivation. Instead, students and teachers are compelled to cooperate by threatening a student's ability to graduate or a teacher's job security. Such transformative systems seek to harness the self-interest of students and educators to refocus schools and redefine the expectations of teachers and learners.
>
> (p. 70)

These competing views of both ends and means surely hinder the ability of states and school districts to implement the kind of coherent and mutually supportive system envisioned by the teaching and learning model advocates. Although NCLB includes teacher quality provisions and mandates for scientifically-based reading instruction, its testing and accountability requirements were modeled after proclaimed successes in Texas and Florida and rely primarily on the threat of sanctions to induce greater effort and improved achievement.

Cacophonous Standards

Hess's comments also point to another source of confusion underlying the standards movement despite its seemingly monolithic form. Hess calls the Virginia reforms ambitious, but then says that the standards represent—at a minimum—the knowledge and skills that should be taught. This rhetorical slight of hand—labeling rigorous standards minimal—has become commonplace. When the standards movement began, the phrase "world class standards for all students" was used to indicate that new expectations would be created that required all children to attain a level of proficiency theretofore achieved by only an elite group of students. World-class language was used with teachers involved in setting standards, and they were encouraged to

eschew normative expectations and to dream about what might be. The result has been very high standards, in many cases set at the seventieth or even ninetieth percentile, as well as great variety in the level of standards from state to state. In 1990, the baseline year for the new NAEP Mathematics Assessment, for example, the standard for proficiency was set at a score level corresponding to the eightieth percentile for eighth graders and at the eighty-seventh percentile for fourth and twelfth graders.

No Child Left Behind increased the amount of testing and the potential negative consequences attached to test results. As a result, some states adjusted their proficiency standards, thus increasing the variability in state standards. Linn (in this volume) documents the tremendous differences between the percent proficient reported on NAEP for each state versus the percent proficient determined by the states' own tests. For example, in 2005 only 18% of fourth graders in Mississippi met the proficiency standard in reading on NAEP, but 87% of fourth graders were reported to be proficient on Mississippi's own test. Differences among states and between states and NAEP could be due to differences in content standards, differences in tests, differences in the stringency of the passing score, or to real differences in student achievement. A recent report by the National Center for Education Statistics (2007), however, reveals that the greatest source of the variability in state results is the differences in the stringency of proficiency standards.

Figure 2.2 is a simplified graphic intended to illustrate that proficiency standards might be set anywhere from the top to the bottom of the test score distribution. These different levels correspond roughly to different eras in the history of test-based accountability. However, these trends are not pervasive, so although there has been a general ratcheting up of standards over time, current practice includes a hodgepodge of leftover minimum competency standards, world-class standards, and "adjusted" proficiency standards adopted by some states for purposes of NCLB.

Unfortunately, most policymakers are not aware of how high some standards have been set and are inclined to treat all standards as if they refer to the same level of accomplishment. Policymakers and journalists also use pass–fail language without realizing that standards are no longer set at a minimum level. In Colorado, for example, there are four reporting categories: unsatisfactory, partially proficient, proficient, and advanced. The unsatisfactory level more closely corresponds to what traditionally would be thought of as inadequate or failing performance. The partially proficient category includes students who are both below average and above average in comparison to national norms; but partially proficient students are now identified in newspaper headlines as "failing" grade-level standards. A striking consequence of reporting assessment results in relation to world-class proficiency levels is that failure rates are alarmingly high and media stories constantly report bad news about public education.

Good and Ill Effects

If the purpose of large-scale assessments is now not to monitor change but to lead it, then how effective have standards-based assessments been in directing positive changes in curriculum and teaching? Studies after a decade of standards-based reform still show the strong influence of high-stakes tests on what gets taught (McNeil & Valenzuela, 2000; Stecher & Chun, 2001; Taylor, Shepard, Kinner, & Rosenthal, 2001). To the extent that the content of assessments has improved, there have been corresponding improvements in instruction and curriculum. In Washington state, for example, teachers reported spending more time during writing instruction on the genres to be

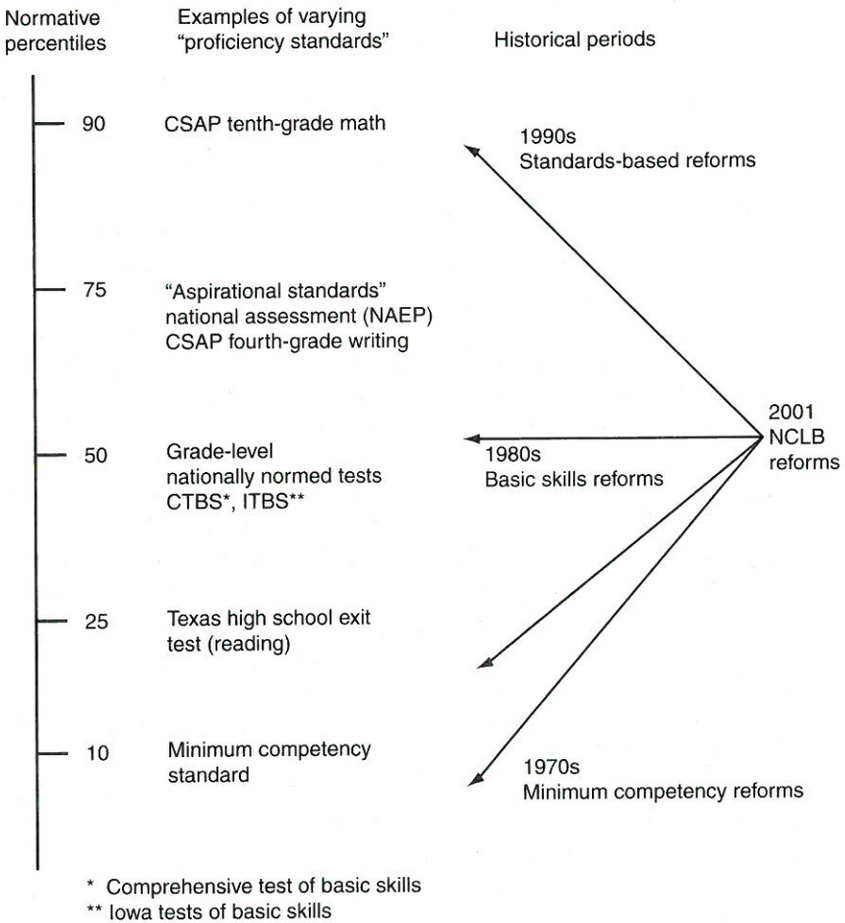

Figure 2.2 "Proficiency Standards" from Different Historical Periods Shown in Comparison to National Norms.

tested and attending to rubric-based writing strategies such as topic, audience, and purpose. In mathematics, the state learning goals and assessments prompted increased instructional time devoted to topics such as probability and statistics and to sense-making activities such as representing and sharing information, relating concepts, and formulating questions (Stecher & Chun, 2001). In Colorado, districts invested in professional development and new writing curricula, which teachers said had genuinely improved instruction (Taylor, et al., 2001).

Unfortunately, recent studies on the effects of standards-based reforms also confirm many of the old negative effects of high-stakes testing. The same surveys that found positive effects in Washington and Colorado also found that time for teaching social studies and science was eliminated or reduced because the state tests focused only on reading, writing, and mathematics. These patterns appear to have intensified under NCLB (Dillon, 2006; Manzo, 2005) with the greatest effects being felt in low performing schools.

Ultimately, an evaluation of the effectiveness of NCLB's high-stakes incentives version of standards-based reform will depend on how well it meets its primary goals of raising student achievement and closing the achievement gap. Nearly three decades of experience with accountability and test-driven reforms has at least provided some wisdom about how these questions should be addressed. In contrast with previous analysts who used score gains on accountability tests themselves as evidence of effectiveness, it is now widely understood by researchers and policymakers that some independent confirmation is needed to establish the validity of achievement gains. For example, two contrasting studies by researchers at the RAND Corporation used NAEP as an independent measure of achievement gains and documented both real and spurious aspects of test score gains in Texas. The study by Grissmer, Flanagan, Kawata, & Williamson (2000) found that Texas students performed better than expected based on family characteristics and socioeconomic factors. However, the study by Klein, Hamilton, McCaffrey, & Stecher (2000) found that gains on NAEP were nothing like the dramatic gains reported on the Texas Assessment of Academic Skills (TAAS). Klein et al. also found that the gap in achievement between majority and minority groups had widened for Texas students on NAEP whereas the gap had appeared to be closing on the TAAS. Both the Grissmer and Klein studies could be true, of course. Texas students could be learning more in recent years, but not as much as claimed by the TAAS.

A 2007 report from the Center on Education Policy (CEP) used state assessment data to evaluate the impact of NCLB on student achievement nationally. They found that most states with three years of data saw increases in reading and math scores, and that there was more evidence of gaps closing than gaps increasing (although gaps remained substantial). CEP attempted to analyze data from all 50 states, but only 13 states had adequate data for analyzing even short-term trends. Lee (2006), using NAEP data through 2005, found quite a

different picture. Lee found that NAEP reading trends were flat before and after NCLB; and that the rate of gain in math was the same before and after the new law. Similarly, when CEP looked at NAEP results they noted low correlations between gains on state tests and gains on NAEP. Many states showing rising scores on their own tests have shown declines or flat results on NAEP. In the period from 1990 to 2005, few states reduced gaps significantly, and Lee found no systematic differences between strong accountability states and weak accountability states in the closing of achievement gaps for blacks, Hispanics, or poor students. Still, the longer-term positive trend on NAEP mathematics might be a sign of general improvements attributable to standards-based reforms more generally rather than NCLB specifically.

Accountability Testing: Lessons Learned

McDonnell (this volume) provides a political science analysis explaining why the core policy ideas of test-based accountability are well entrenched. In addition to the political ideal of democratic accountability, accountability mandates tap into powerful belief systems underpinning Americans' love affair with testing.

Accountability testing and its impacts are not new. Policymakers in successive decades seem to discover, each time for the first time, that U.S. economic competitiveness is threatened by poor achievement, especially in math and science. In response, test-based accountability is seen as an effective top-down means to ensure that schools work harder to improve student learning. Each time, well documented consequences of high-stakes testing have been the narrowing of curriculum and instruction to focus only on tested subjects using test-like formats. In many cases, teaching the test hurt learning rather than helped it. Indeed, the standards and assessment reforms of the 1990s were intended to correct the teaching-the-test consequences of 1980s reforms, which before that had been intended to correct the severe limitations of minimum competency testing in the 1970s. Not remembering any of this, the framers of NCLB took a backward step, imposing more testing, which made it more likely that cost constraints would limit the substantive quality of tests.

Over time, there has been a general ratcheting up of standards but also a proliferation of different standards without any transparency for the public about what has changed and what has stayed the same. Schools look worse and worse if students are said to "fail" when they don't meet high "world class standards" or when "adequate" yearly progress (which seems to imply "normal" progress) is defined in terms of 100% proficiency.

In thinking about how to reform the reforms, the following lessons are the most critical: (1) better quality, substantively challenging assessments are less likely to cause curriculum distortion than limited, multiple-choice-only tests; (2) when tests are used to drive reform, they can't be used as the sole measures of the reform's effects; (3) when an incentives-based coercive model of

standards-based reform is adopted instead of one based on capacity building (including more challenging curricular resources, improved assessments, and teacher professional development), there is little evidence that accountability systems will achieve their desired ends; and, (4) test scores may go up—but in cases without real improvements in teaching and learning—apparent gains have not been confirmed by independent tests.

The claims about the benefits of test-based accountability for improving education should themselves be subjected to audit and evaluation. Given several decades of high-stakes, test-based accountability, it is conceivable that such programs are sometimes the *cause* of poor instruction and limited learning rather than being a guaranteed cure. The most recent study using NAEP fails to find improved achievement or closing of achievement gaps associated with NCLB. Nonetheless, steadily rising gains in mathematics since 1990, especially at fourth grade, suggests that reforms have had beneficial effects. Although it is impossible to isolate the specific causes of large-scale trends, teacher survey data and smaller-scale studies of innovations tell us that content standards and improved curriculum have made more of a difference in effecting these changes than test scores and pressure alone.

Acknowledgments

The author wishes to thank Carl Kaestle, Katherine Ryan, and Nancy Cole for thoughtful comments and critique in response to earlier versions of the manuscript. Portions of this chapter were originally presented at the Spencer Foundation's 30th Anniversary Conference, "Traditions of Scholarship in Education," Chicago, January 24–25, 2002.

References

Alexander, L. & James, H. T. (1987). *The nation's report card: Improving the assessment of student achievement.* Washington, DC: National Academy of Education.

Baratz, J. C. (1980). Policy implications of minimum competency testing. In R. Jaeger & C. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, consequences* (pp. 529–539). Berkeley, CA: McCutchan.

Brickell, H. M. (1978). Seven key notes on minimum competency testing. *Phi Delta Kappan,* 59 (9), 589–592.

California State Department of Education. (1973, January). *Feedback, 1.* Sacramento, CA: California State Department of Education.

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

Center on Education Policy. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Washington, DC: Center on Education Policy.

College Board. (1977). *On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline.* New York: College Board.

Cook, W. W. (1941). Achievement tests. In W. S. Monroe (Ed.), *Encyclopedia of educational research* (pp. 1283–1301). New York: Macmillan.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements.* San Francisco: Jossey Bass.

Darling-Hammond, L. & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *Elementary School Journal*, 85 (3), 315–336.

Dillon, S. (2006, March 26). Schools cut back subjects to push reading and math. *New York Times*. Online, available at: http://www.nytimes.com/2006/03/26/education/26child.html (accessed February 21, 2008).

Duke, D. L. (1985). What is the nature of educational excellence and should we try to measure it? *Phi Delta Kappan*, 66 (10), 671–674.

Elmore, R. F. & Rothman, R. (Eds.) (1999). *Testing, teaching, and learning: A guide for states and school districts*. Washington, DC: National Academy Press.

Findley, J. (1978). Westside's minimum competency graduation requirements: A program that works. *Phi Delta Kappan*, 59 (9), 614–618.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18 (9), 27–32.

Glaser, R. (1987). Commentary by the National Academy of Education. In L. Alexander & H. T. James (Eds.), *The nation's report card: Improving the assessment of student achievement* (pp. 43–76). Washington, DC: National Academy of Education.

Glass, G. V (1978). Standards and criteria. *Journal of Educational Measurement*, 15 (4), 237–261.

Goslin, D. A. (1963). *The search for ability: Standardized testing in social perspective*. New York: Russell Sage Foundation.

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

Halperin, S. (1975). ESEA ten years later. *Educational Researcher*, 4 (8), 5–9.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity*. Washington, DC: National Academy Press.

Hess, F. M. (2002). *Reform, resistance,... retreat? The predictable politics of accountability in Virginia*. In D. Ravitch (Ed.), *Brookings Papers on Educational Policy: 2002* (pp. 69–122). Washington, DC: Brookings Institution Press.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends*. Cambridge, MA: Civil Rights Project at Harvard University.

McLaughlin, M. W. (1975). *Evaluation and reform: The Elementary and Secondary Education Act of 1965*. Cambridge, MA: Ballinger.

McLaughlin, M. W. & Shepard, L. A. (1995). *Improving education through standards-based reform. A report of the National Academy of Education Panel on standards-based reform*. Stanford, CA: National Academy of Education.

McNeil, L. & Valenzuela, A. (2000). *The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric*. Cambridge, MA: Harvard University Civil Rights Project.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4–12*. Boston, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Manzo, K. K. (2005, March 15). Social studies losing out to reading, math, *Education Week*, 24 (1), 16–17.

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era*. Princeton, NJ: National Assessment of Educational Progress.

National Center for Education Statistics. (2007). *Mapping 2005 state proficiency standards onto the NAEP scales* (NCES 2007–482). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Council on Education Standards and Testing. (1992, January 24). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. Washington, DC: U.S. Government Printing Office.

Owens, B. (1999, December 8). Announcement of "Putting Children First: A plan for safe and excellent public schools" [Remarks as prepared for delivery]. Online, available at: http://www.state.co.us/childrenfirst/ChildrenFirstRemarks.htm (accessed February 21, 2008).

Pipho, C. (1978). Minimum competency testing in 1978: A look at state standards. *Phi Delta Kappan*, 59 (9), 585–588.

Pipho, C. (1985). The excellence movement: On ice for the summer? *Phi Delta Kappan*, 66 (10), 669–670.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, 68 (9), 679–682.

Resnick, D. P. (1980). Minimum competency testing historically considered. *Review of Research in Education*, 8, 3–29.

Resnick, D. P. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies, Part II* (pp. 173–194). Washington, DC: National Academy Press.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Boston, MA: Kluwer Academic Publishers.

Rottenberg, C. & Smith, M. L. (1990, April). Unintended effects of external testing in elementary schools. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Shepard, L. A. & Dougherty, K. (1991, April). Effects of high stakes testing on instruction. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, CA: National Academy of Education.

Smith, M. S. & O'Day, J. (1990). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). London: Taylor & Francis.

Stake, R. E. (2007). NAEP, report cards and education: A review essay. *Education Review*, 10 (1). Online, available at: http://edrev.asu.edu/essays/v10n1index.html (accessed September 28, 2007).

Stecher, B. & Chun, T. (2001, November). *School and classroom practices during two years of educational reform in Washington state*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991, August). *Summative evaluation of the National Assessment Governing Board's inaugural 1990–91 effort to set achievement levels on the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

Tallmadge, G. K. & Wood, C. T. (1978). *The user's guide: ESEA Title I evaluation and reporting system*. Mountain View, CA: RMC Research Corporation.

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2001). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost*. Boulder, CO: Center for Research on Evaluation, Standards, and Student Testing, University of Colorado.

Terman, L. M. (1916). *The measurement of intelligence*. Boston, MA: Houghton Mifflin.

Tyack, D. (1974). *The one best system: A history of American urban education*. Cambridge, MA: Harvard University Press.

Tyler, R. W. (1966). The objectives and plans for a national assessment of educational progress. *Journal of Educational Measurement*, 3 (1), 1–4.

U.S. Congress Office of Technology Assessment. (1992, February). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

U.S. General Accounting Office. (1993, June). *Educational achievement standards: NAGB's approach yields misleading interpretations* (Report No. GAO/PEMD-93-12). Washington, DC: U.S. General Accounting Office.

Wise, A. E. (1978). Minimum competency testing: Another case of hyper-rationalization. *Phi Delta Kappan*, 59 (9), 596–598.

Worthen, B. R. (1974). *A look at the mosaic of educational evaluation and accountability.* Portland, OR: Northwest Regional Educational Laboratory.

Worthen, B. R. & Sanders, J. R. (1973). *Educational evaluation: Theory and practice.* Belmont, CA: Wadsworth Publishing.