education review
a journal of book reviews

Volume 10 Number 1 **February 4, 2007**

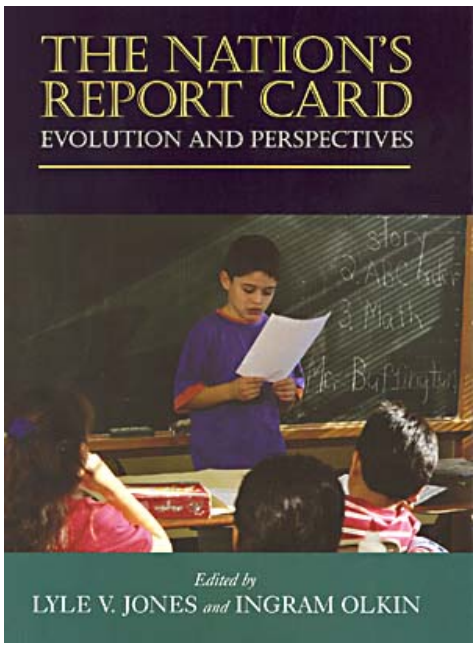
NAEP, Report Cards and Education: A Review Essay

Robert E. Stake
University of Illinois at Urbana-Champaign

Jones, Lyle V. & Olkin, Ingram.(Eds.) (2004). *The Nation's Report Card: Evolution and Perspectives.* Bloomington, IN: Phi Delta Kappa Educational Foundation.
Pp. 598 ISBN 0873678486

Citation: Stake, Robert E. (2007, February 4). NAEP, report cards and education: A review essay. *Education Review, 10*(1). Retrieved [date] from http://edrev.asu.edu/essays/v10n1index.html

The National Assessment of Educational Progress (NAEP) was started in the early1960s by education policy people for political and technological reasons. It came to be called the Nation's Report Card. It was expected to inform national policy. It was to authenticate Cold War and War on Poverty education reform efforts. And it was to further educational research. The political instigator was Francis Keppel, U. S. Secretary of Education. The Carnegie Foundation put up the first money. New and different purposes emerged as opposition to a federal role in U.S. Education died away. And as once feared, it helped move the nation toward standards-based teaching, uniform curricula and coast-to-coast standardized testing (Jennings & Rentner, 2006; Sarason, 1990).



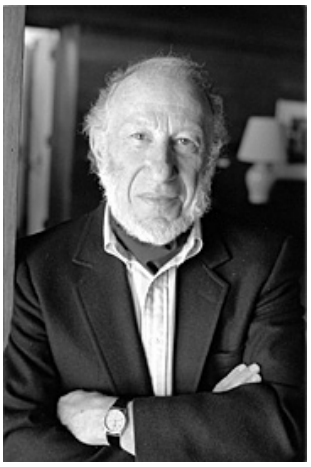
The conceptual force behind NAEP was Ralph Tyler, a pragmatic educational psychologist, measurements specialist, and curriculum consultant. He had been a member

of a strong curriculum studies faculty at the University of Chicago. Some of his students became outstanding psychometricians. Tyler (1966) drew upon psychometrics, but not curriculum expertise, for help in developing this assessment of educational quality.

Early NAEP leaders held the view that sampling examinees was the critical problem for a large-scale indicator of educational progress. The problem of sampling the vast terrain of educational content was finessed. There was an abiding presumption that student performance on standardized tests—when aggregated over a small number of items by a large well-selected group of examinees—*could* indicate group scholastic achievement in a subject (social studies, science). And a second abiding presumption that such an indication of scholastic achievement *could* indicate the quality of schooling and the extent to which a nation is being educated. Both presumptions could have been subjected to empirical validation but no prior or parallel effort was made to assure that validity.

Lyle V. Jones

A professional specialization in educational program evaluation had emerged at the time. Tyler was part of it but he did not hold it in high regard, perhaps partly because limited enthusiasm was expressed therein for his own conceptualization of evaluation (comparing performance to goals). Tyler's idea of critical evidence of learning was measurements of increase in knowledge and skill (performance) by the individual student. Evidence of national progress could simply be aggregates of those increases. Tyler's view of behavioral objectives and goal-based evaluation shaped NAEP theory and practice at first but political pressures later moved NAEP toward more conventional aptitude testing.



Ingram Olkin

People in the assessment vocation (extending widely into business, health care, and government) saw opposition to NAEP as political (which it was) and irrational (but it was well reasoned). Opposition was treated by NAEP advocates with "damage control" more than "co-construction." There was little compromise with those who saw that national assessment might ultimately hurt education.

In the book's appendix (p 557) is a summary by David Goslin (1964) of two of the earliest NAEP planning meetings, noting that the participants recognized some potential hazards, including: "Tests may have the effect of *defining* the legitimate boundaries of educational concern in the eyes of Congress, the public, and even educators." And "It is clear that the group constructing the test would in many respects be setting educational standards." And "Inevitably, there will be a tendency on the part of teachers to teach for the test …." Perhaps because mandated state testing programs have long worked in the same scholastic territory, these concerns were little apparent in the remainder of the book. Yet

the consequences are real and a part of the full history of NAEP. They will be discussed in the next-to-last section of this review.

Book review. I am offering here a review essay of *The Nation's Report Card*, a collection of papers solicited by Lyle Jones and Ingram Olkin and published in 2004 by the Phi Delta Kappa Foundation and the American Educational Research Association. It is an insiders' version of the history of NAEP. I was Associate Director of the Illinois Statewide Testing Program when this history began, and became acquainted with many of the chapter authors and interviewees. But I was an outsider, a skeptical friend (Stake, 1970). The editors and authors of this book were engaged in the creation and nurturing, the advocacy and protection, of the National Assessment of Educational Progress.

The collection starts with two historical accounts, one by Irving Lehmann, a leading measurements spokesperson of the day and the first NAEP Assistant Director for Research; the other by Frederic Mosher, the Carnegie Foundation Program Officer dealing with NAEP at its inception. These pages were written sensitively and personally, telling of an intimacy and intellectual integrity that set the tone for the papers to follow. (The order of subsequent paragraphs follows the order of chapters in the book.)

In my view, John Gardner was the senior educational statesman of the 1960s. As President of the Carnegie Corporation, he found $100,000 for early exploration of national assessment possibilities. In an interview with Ingram Olkin and Marshall Smith for the book, he recalled early conversations with Tyler and Keppel; didn't stay involved, but saw NAEP opponents (mostly school administrators of the day) as "absolutely cold, tough, self-interested." Gardner asked his interviewers, "But don't you think the whole thing of standards is somewhat crass (p 121)?"

Olkin also interviewed Lloyd Morrissett, Gardner's aide on the project. Morrissett remembered the diversity of participants in the early conversations, such as from Wall Street and the Atomic Energy Commission, yet mostly were trusted acquaintances, some Yale connections. " … it was clear from the beginning," he said, "that to undertake something like a national assessment meant overcoming very substantial political resistance in the educational community (p 123)."

In another interview: Russell Sage staffer David Goslin, conference recorder, said: "I pointed out in my report that NAEP eventually would have an impact on curriculum. … Today (2001) NAEP is the closest thing we have to national standards for education. … if the conference participants had predicted these outcomes at that time, NAEP might not have come into existence (p 137)." And in another interview, measurements specialist Lee Cronbach lamented the mid-80s changes when NAEP passed from operation by the Educational Commission of the States to operation by the Educational Testing Service, saying:

> … the efforts of the (1965 Technical Advisory) committee were diametrically opposed to what has gone on since NAGB (1988) took over. … We were interested in framing questions so that a student who had not studied the desert at all could still think about

> an ecological question and put the pieces together. So there was an attempt to free the questions from dependence on the lesson system, free it from the lessons the student had studied. But the understanding was that it was going to influence the teacher to think about whatever this teacher had been doing to promote ecological thinking … That is quite different from anything that ETS was thinking about. (p 144)

The chapter by Clay Allison (*nom de plume*) identified National Assessment's Technical Advisory Committee—in TAC member Bob Abelson's words—as " the heart and guts of NAEP (p 158)." In 1969, TAC had become ANAC and merely advisory but, early on, it had analyzed the data and written the first national reports. Member Lyle Jones said, "We not only gave advice on agency contractors, but we even designed exercises[1] (p 157)." Princeton statistician John Tukey chaired the committee. Jack Merwin, educational measurements specialist at the University of Minnesota, was officially NAEP Director then, but his job was to deliver the data for the committee's interpretations.

Allison described the work of the committee as intensive, meeting about ten times a year, usually two long evenings and until next day's last plane home. Each of the four members was responsible for a mass of data interpretation, writing and critiquing drafts, interactively as a team. Abelson repeatedly said, "The first question is, Will the exercises do their job? (p 158)." He took on the race comparisons; others took the comparisons by region, education of parents, type of community, and sex. Increasingly the emphasis was on these comparisons, less on the content of the exercises.

Earlier the committee had labored on examinee sampling procedures but by 1972, their heavy work had turned to summary comparisons of performance on individual items for reporting to the public. As reported by member Lee Cronbach, their early commitment was to avoid reporting items by subject matter (math, citizenship, etc.) and to stick to criterion referencing—that is to create items of performance meaningful to a lay reader (understanding a newspaper item, explaining a chemical experiment) and issuing a report on each item. Item reporting as opposed to subject reporting turned out to be politically unacceptable, easy to trivialize, and it got shot down in early negotiations. The myth of single-dimension subject-matter achievement took over.

Technical choices. Much of the Technical Advisory Committee work was judgmental, and however wise the members, potentially biased. Tukey said, "… we hope that whatever biases the members of [the committee] have are sufficiently diverse so that we get good answers and good judgment (p 159)." That they challenged each other was—given the intellect and integrity of the members—assured. But the effort to answer Abelson's question, "Will the exercises do their job?" remained secondary partly because educational progress remained vague and multiply perceived. Was NAEP's purpose to indicate the quality of American education or the educational maturity of its youth. Or was it to direct public policy? Or to cause teachers to reconsider their instructional goals? Yes,

[1] Not just multiple-choice items, the early exercises included laboratory tasks, discussion topics, and esthetic issues.

to all, but these purposes competed. Intent on avoiding individual score reporting and state-by-state comparisons, the members were focused on the performance of a national population of individuals. They *did not want* to give attention to quality of teaching, leadership, and taxpayer investment in education (Pellegrino, Jones & Mitchell, 1999). NAEP statements vigorously pointed out that there were limits to what the scores meant, but those cautions were often lost on people who wanted answers.

Much of the education community at that time was reluctant to support NAEP, partly because it anticipated invidious comparisons, but also because it was reluctant to yield public definition of educational progress to a committee made up of a statistics professor, two quantitative psychology professors and an educational measurements professor, with the director of the Center for Advanced Study in the Behavioral Sciences sitting in. Mrs. Tukey was quoted as saying, "They all are John's friends (p 163)." I found no indication in the book that the Technical Advisory Committee (nor its successors: ANAC, OPAC, NAPC, and NAGB) called for validity studies to examine the match between score reports on educational progress and the consequent[2] conceptualizations of policy setters and other readers of the reports.

Although in the early 1970s Analysis Advisory Committee (ANAC) members continued to think of NAEP as providing indices of student achievement on specific tasks, NAEP managers and funders (increasingly the U. S. Office of Education) called for more direct and holistic policy relevance. In part, that meant comparisons with respect to the subject matter learned (e.g., for all of science) rather than to topics or tasks (within science). People were familiar with course grades not broken into topical learnings, and little need was seen outside ANAC for the "publicly understandable" topical performance results conceptualized in Tyler's original design of NAEP.

The question "Will the exercises do their job?" is a quite different question for task assessment versus subject assessment. At frequent meetings, TAC committee members labored at what Tukey called, "icking the questions," identifying exercises (items) to be dropped. When the emphasis is on single task interpretation, the selection of items could be based on publicly-meaningful content of a single item. But with subject matter interpretation, item selection shifted to asking if each item contributed to a homogeneous collection ostensibly representing broad achievement across the subject.

From then on, the item collectivity needed major scrutiny. With attention shifting away from public interpretation of single items, the item selectors could concentrate on the internal coherence of items and on policy-maker expectation of curricular achievement. They were still responsible for matching the knowledge of the students and public interpretation of subject matter performance with the content domain of the items. This book gives the reader little reason to believe that the Technical Advisory Committee and its successors mobilized the expertise or called for broader validity studies to authenticate their item selection procedures.

---

[2] Later, Samuel Messick (1989) advocated studying the "consequential validity" of educational assessment.

For curricular policy, a skeptic might well ask: Is it not possible that topics within a subject-matter such as mathematics or social studies are too internally unrelated from topic to topic to be validly represented by a single score of *achievement*? Is it not possible that for general educational policy making, subject matter performance is too broad a construct? Indicator advocates urge caution but worry little about misrepresentative homogeneity of ingredients (Shavelson et al., 1987), nor about consequences. What makes sense for political policy makers should not determine what teachers teach and what testers test.[3] It is easy to see that *aptitude* for learning can be a much more homogeneous construct than actual learning. And it is scholastic aptitude far more than achievement that has been measured with standardized tests for almost a hundred years. Whatever its face validity, the National Assessment of Educational Progress is probably a scholastic aptitude test battery more than an achievement test battery. We don't have the research to say for sure. More on this later.

The federal role. Back to the book. According to the chapter by distinguished psychometrician and U S. Assistant Commissioner of Education Dorothy Guilford, government statisticians recognized that the NAEP student performance data could serve their own statutory mission to measure "the character and progress of American education." Mostly what traditionally they collected had been background and input information; what they needed, she said, if they were to be of greater use to policy makers, was outcome measures. Guilford identified herself as a strong proponent of NAEP, in contrast to, for example, John Evans, Assistant Commissioner for Planning, Budgeting and Evaluation, who at one time tried to eliminate all federal funding for NAEP because "the data lacked policy relevance." Many of us in the measurements field thought Evans overly political, but the cry for years to come was for greater NAEP usefulness. Guilford tried to increase the dissemination and interpretability of NAEP data and she spoke of the need for studies to increase their validity. "Will the exercises do their job?"

In 1983, the National Institute of Education awarded a grant to the Educational Testing Service, taking over from the Educational Commission of the States. In his chapter, Archie Lapointe, the new Executive Director, documented the change. He and Willard Wirtz had headed an independent evaluation of NAEP that recommended expanded activity, including state-by-state comparisons. Their primary observation was that "educational standards are here to stay" but could be done more effectively. Although the evaluation team and contributors included educational specialists, little attention was given to the validity of student performance testing as an indicator of educational progress. Pressures had increased over a fifteen-year span to separate technical and policy issues, and by 1983, the close combination held by the original Technical Advisory Committee had disappeared. Separating the two lessened further the possibility that the external validity of NAEP would be researched.

It was the politically inspired concept of "accountability," according to Ramsay Selden, then with the Council of Chief State School Officers, that moved NAEP toward

[3] And here is the critical difference between indicator-making and school policy-making. Indicators are good if they are internally coherent, that is, if items are highly correlated across students. Curricular policy is good if it protects the territory of heterogeneous subject matter.

state-by-state comparisons. Pressed by U. S. Commissioner Terrel Bell's charge (National Commission on Excellence in Education, 1983) of "educational mediocrity" in America, the state superintendents, once states-rights protectionists, now advocated state-by-state comparisons "if a common set of educational standards could be reached." Somehow the ideas of the founding fathers that education should be left to the states, presumably because the states could better serve the interests of families and communities, had wafted away. Bell's federalist vision of mediocrity in teaching and learning was seen to be best remedied by having all states arrive at a common curriculum and by measuring performances uniformly. The fears of the original opponents to NAEP had thus been realized in about twenty years. The resistance was broken partly by federally contracting in 1969 to have the Education Commission of the States operate NAEP. It was a coup spurred by collaboration of the states more than by federal legislation. As a developer of assessment policy for state offices, Selden claimed that it had become "widely recognized that state comparisons were healthy and appropriate (p 195)."

As political assemblages the Chief State School Officers and the Education Commission of the States promoted an ethic of consensus. In terms of content domains and assessment topics to be tested, that meant that only non-controversial matters would be included within the definition of educational progress.[4] According to Mary Lou Bourque, NAGB's chief psychometrician, this ethic of avoiding controversy (in item content if not in governance) remained well past 1994. In 1988, feeling that the values of NAEP had been set too much by technicians and civil servants, and following recommendations of the Alexander-James Report (1987), Congress created a new policy group, the National Assessment Governing Board to replace the ECS Policy Board which had replaced ANAC and TAC

NAGB was a diverse 25-member committee (with little Yale connection) having broadly identified responsibilities, with vested interests (as have most committees), and inclinations to participate (as had the original TAC) in the management of the program. Seeing that the Technical Advisory Committee and successive policy committees had left much undone, Bourque made it clear that NAEP responsibilities were formidable. NAGB essentially was a lay group, neither technical nor professional, quite unaware of the difficulty and danger in its new commitment to set cut-score standards of student competence, namely for *Basic, Proficient*, and *Advanced* performance. The commitment was political, not educational (Glass, 2003).

When an external evaluation of NAGB's standards-setting procedure was completed in 1991,[5] Bourque reported that the findings were so negative that the evaluation contract

[4] John Dewey and Joseph Schwab, Tyler's mentor and colleague. two leading educators of the century, saw the study of competing ideas as central to education.

[5] The evaluation was headed by the highly respected head of Western Michigan University's Evaluation Center, Daniel Stufflebeam (1991). In a 2006 email to me he said, "As I look back on the experience, I think that NAGB was under severe pressures from the Bush Administration to show through test results that previous Democrat-led education policies and programs had failed and from Congress to show that NAGB was providing competent, fair-minded leadership and oversight of NAEP.

was terminated, without indicating what the findings were.[6] The standards-setting work was subcontracted to the American College Testing Program, which set standards for most NAEP testing in the 1990s. The standards-setting mechanism remained controversial, often said to be "fatally flawed," drawing in the National Academy of Education and many measurements experts—all apparently without confronting the Congressional presumption that the societal life and institutions of America are sufficiently homogeneous to warrant setting more or less arbitrary national standards of student competence.

Congress was clearly part of the problem, seeking to define in legislation matters better left to professional practice and local decision-making. Bourque pointed out how NAGB, low on technical resources, secured legislative support to bring in more psychometric expertise. But it seemed to me that the discipline of psychometrics was not able to generate what Congress prescribed. NAGB policy required reporting to be free of political considerations, but the entire NAEP effort had technocratic priorities, some linked to political partisanship. Bourque concluded with a description of how the No Child Left Behind legislation[7] changed NAEP from low-stakes testing to high-stakes testing. Intentionally, NAEP helped narrow the curriculum and, unintentionally, diminish instruction oriented to student diversity and assure more teaching to the test. Although neither children nor schools were directly hurt by their NAEP scores, the nation's educational progress—I will claim at the close of this review—was hurt.

Emerson Elliott and Gary Phillips, Commissioner and Deputy Commissioner of Educational Statistics, wrote a chapter for the book on NAEP as it appeared from the U. S. Office of Education, particularly in 1988 when Congress overhauled it, creating the National Assessment Governing Board and requiring interpretation by student-achievement cutting scores. In 1976 the General Accounting Office had rebuked NAEP for offering too little help to educational decision makers, calling even then for specific performance standards. The Alexander-James report agreed (1987), calling for NAEP to take full advantage of state-of-the-art advances in test technology and then setting requirements beyond the state-of-the-art.

In 2001-2, co-author Ingram Olkin interviewed Chester Finn, a Ronald Reagan functionary and first NAGB chair; Marshall Smith, Stanford professor and Bill Clinton's Deputy Secretary of Education; and Jack Jennings, long-time, widely-respected Congressional staffer on education. As before, the questions were genial, not confrontational. Finn said he helped Secretary Bill Bennett identify NAGB members and try to resolve the question of testing for what the American curriculum was or should be. Smith said that when NAEP moved from the Education Commission of the States to the Educational Testing Service, the test reporting moved from "a passive thermometer to a much more aggressive posture (p 269)," and that the Clinton White House paid increasing

[6] Strong criticism of NAEP technical operations seldom appears in the book. Acclaim appears more often. Bourque herself said "… short term trends [reported by NAEP] … have been so influential in guiding state policies and moving the educational reform movement forward during the last decade.

attention to NAEP, using it in promoting its education policies. Jennings described long-running battles for federal funding of education, with NAEP of minor importance, but with House Democrats opposing extending testing and Conservatives favoring exposure (through testing) of the weak payoffs of federal funding.

The many complexities. Returning to the technical, Bob Linn's chapter offered his view of the influence of external evaluations of NAEP. Funding for NAEP regularly called for external evaluation, internal evaluation was even more frequent and penetrating, and the results were remarkably similar: NAEP was needed, was applying state-of-the-art technology, was unable—partly because of funding—to do needed background research, and was under stress because political and technical choices sometimes conflicted. The administrators of NAEP tried to operate by consensus, to do only what a variety of advisors could fully agree upon, but that may have restrained the circle of advisors and taken certain issues off the agenda.

The evaluations relied largely on blue-ribbon panels, i.e., small groups of experts, almost all distinguished in large-scale testing. One panel had two specialists in educational program evaluation and a third in educational measurements. Linn concentrated on studies occurring between 1980 and 2000. He found that the conclusions drawn usually had already been expressed by NAEP's policy board (NAGB). But it was clear too that the Board had a strong political leaning regarding educational needs and expected more precision of measurement and meaningfulness of test performance than ETS and its many collaborators were able to provide. From early times on, the evaluators pointed out that NAEP was expected to do more than it could do, whether indexing educational progress, invigorating policy, or assisting state assessments. An evaluation by the National Research Council concluded:

> The nation's educational progress should be portrayed by a broad array of educational indicators that includes but goes beyond NAEP's achievement results. The U.S. Department of Education should integrate and supplement the current collections of data about education inputs, practices and outcomes to provide a more comprehensive picture of education in America. (Pellegrino, Jones, and Mitchell, 1999, p 22)

Too much had been asked. Too much had been promised. Unrealistic expectations are not just impossible dreams, but can be a political ploy to discredit past initiatives and discourage public investment in education.

Wayne Martin, Director of the CCSSO State Assessment Center, wrote a chapter on how NAEP was viewed from the Education Commission of the States, NAEP's home, from 1969 to 1983. As an organization of State School Superintendents, something of a small lobby group, ECS increasingly found the delivery of large-scale testing a burden. Others claimed ECS had not done enough to make NAEP useful to researchers. Politically it made sense to have NAEP run by the states, who originally feared the intrusion of federal measurements, but once established, it made sense to transfer it to a large-scale testing corporation, ETS, the Educational Testing Service (Messick, Beaton & Lord, 1983). Martin also described the burden on the Colorado State Department of Education in the early 1990s

to participate in state NAEP testing. And he described the State Superintendents' initial opposition to NAGB's decision to set "achievement levels," thus nationally standardizing definitions of academic competence; but soon Colorado Governor Roy Romer and the State Superintendents joined the New Standards Project to set national educational standards. Across the country, speakers for the constituencies of school reform, however disagreeing on tactics, joined in increasing clamor for standards, hushing the traditions of local control and advocates of curricular diversity.

Tyler's advocacy of item scores was acknowledged as impractical by Frederic Mosher of Carnegie in his chapter, "What NAEP Really Could Do." Collections of student scores on individual exercises turned out to be of small use (to the public, to policy makers, and to the profession) as general indicators of educational progress (Linn, Baker & Burstein, 1991). Nor did clusters of items. Nor did IRT scaled scores. It was commendable that the creators of NAEP wanted the scores to have public meaning, but it was too much to expect. It was not because the public is dumb. Education is complex. The constructs of good education and educational progress defy reduction to scaling. Only by reducing the concept of school system performance to what the tests measure can we work with scales such as NAEP's. Evaluating education requires much more.

Reading comprehension and mathematics achievement (and such) are constructs of mythical character. Of course something gets scaled—but, as Mosher meticulously pointed out, the educational meaning of these scales is hard to pin down. Teachers and curriculum specialists use them as slang, giving them little challenge. The constructs have little pedagogic and curricular definition. (Of course they have personal and political value.) Student achievements are mosaics of performance across a great variety of situations. Task performances are different from traits such as reasoning ability and spatial relationships aptitude. Grade point averages collect dissimilar knowledges, many of them pertinent to a context. Teachers, testers, and parents describe a student's academic performance using a grade or a point on a scale, and it has meaning, a gross meaning, a crude meaning. It can be reliable. But that letter or numeral is not a refined representation of how well educated in a content domain the student has become. And its error does not disappear by adding scores for a large number of students to provide a district or national score. Mosher said, "[NAEP] failed to invest in developing the kind of understanding of subject matter that would provide a basis for developing sound standards (p 334)." He suggested that ETS's introduction of Item Response Theory was the giant step moving NAEP away from proper respect for the complexity of education.

Mosher noted that the goals of American education have thus been restated to "helping every student reach proficiency in the core subjects." (p 332) He implied that switching from what professionals define as education to what technicians purport to measure has lowered the educational progress of a nation.

A small part of the book attended to the development of assessment materials. Vincent Campbell and Daryl Nichols wrote on "Assessing Citizenship," noting that the original emphasis on measuring what ought to be taught as well as what was taught was a distinction not prominent in NAEP reports. They asked, "What have been the effects of NAEP assessment? What are the side effects?" They saw that school curricula have

narrowed, in spite of an ever-increasing domain of knowledge to teach. Narrowing the curriculum is partly due to setting common goals and state standards, but also the increased politicization of setting frameworks and state review of textbooks. It is also due to the testing, and the narrow expectation of what will be on the tests.

Writing on "Assessing Writing and Mathematics," Ina Mullis observed that NAEP had lots of educators involved in its test development, working hard at goal setting and item review; but they came into a workplace where education had already been conceptualized in terms of student behaviors and psychometric traits. In this shop, the test items were to be authentic in the eyes of measurements scholars rather than teachers,[8] and desirable in the eyes of citizens. Cost was a severe restraint on item type. Very little advanced content, especially that taught only in special schools, was represented in item pools; such content did not meet consensus criteria, so that students taking those classes were denied the opportunity to show part of what they had been taught.

From early on, as Bourque pointed out, NAEP used "the consensus process" for item selection; later it was required by law:

> …each learning area assessment shall have goal statements devised through a national consensus approach, providing for active participation of teachers, curriculum specialists, subject matter specialists, local school administrators, parents, and member of the general public. Public law 98-511, Section 405 (E)(19 October 1984).

While politically advantageous and sometimes educationally desirable, consensus has a narrowing effect on education. Whatever the statistical characteristics, consensus items inadequately represent non-consensus achievements.

Technical highpoints. Item design was important but secondary. Survey design was seen to be more important, particularly as to the groups to be compared, such as ages and regions. NAEP's greatly-deliberated early technical planning and student sampling were chronicled by James Chromy, Alva Finkner, and Daniel Horvitz. The Research Triangle Institute team was hard-wired for getting random student participation. They shrugged off random sampling of scholastic knowledge and skill: "Even though no formal sampling process was followed, the set of exercises finally appearing in an assessment was viewed as a sample from a larger universe of potentially available exercises (p 389)." The nonchalance was mine at the time too. Reviewing the progress of national assessment in 1970 (Stake, 1970), I claimed that the meaning of NAEP would come with its use. I did not anticipate the narrowing of the curriculum (See section below on Educational Consequences).

[8] Teacher review of items is often pitched to esoteric fault-finding and topical advocacy, especially for items submitted by another group. Agendas seldom allow for well-rounded discussion of the domain.

Sampling and field operations in the middle years, 1983-2001, were described by Keith Rust, an officer of Westat, subcontractor at the time for those matters. "Designs and operations that would have been considered impossible in the early years are now implemented routinely (p 447)." He pointed out what others saw at the turn of the century, that NAEP would be changed by the interest of the federal administration and Congress in having it play a significant role in determining the accountability of education in the states. Movement to standardized state assessment would reconfigure the original ideas for this assessment. Rust predicted future changes in NAEP given the "federal interest in determining the accountability of states for improvements in education in the elementary and middle school years" and "the increasing burden that testing programs … are placing on the states, districts, and especially schools (p 447)."

Albert Beaton and Eugene Johnson, multiply-honored psychometricians, provided a chapter on emerging technical innovations in NAEP. Stephen Lazer, recently Executive Director of NAEP, contributed a chapter on innovations in instrumentation and dissemination in NAEP. Both cover essential topics for a history of one of the most profound technical accomplishments in measurement and assessment of education.

Although not debated in the book, it is widely presumed that even crude measurement is a step in the right direction. Social scientists often justify indicators on the basis of inter-ingredient correlation. And, in education, there often is high correlation between crude and sophisticated indicators of achievement. But interpretation of crude indicators is a poor guide to policy and practice. For example, there are no scientific grounds for penalizing a school for failing to improve test scores. Embarrassment sometimes may be effective in changing behavior, but it is a deceit to say that student test performance is a valid basis for evaluating school effort. NAEP scores do not adequately represent quality of schooling. The trend lines do not adequately represent the rise and decline of American education.

Jones and Olkin's *The Nation's Report Card: Evolution and Perspectives* tells an important part[9] of NAEP history, presenting the views of many of those most influential in its rise to prominence. There is a handsome redundancy of NAEP technology and politics across the chapters, revealing, as expected, several different views of what NAEP was supposed to do. NAEP was a noble and elegant effort to produce periodic stop-action photos of youth competence. Alas, there is little evidence here or elsewhere that NAEP

[9] My review, of course, tells much less of the story than the book does, and very little of the technical accomplishment. (For that see Shepard, 1995, and Jones, 1996.) Making a suggestion on the first draft of my review, co-editor Lyle Jones remarked, " … With respect to the earliest years, you fail to note the breadth of coverage by subject area (see p 562 of J&O) or the range of exercise types included in each area—laboratory tasks, discussion topics, esthetic issues, etc., with minimal use of multiple choice formats. This was in keeping with Tyler's (and TAC's) desires, and a hallmark of the initial efforts. Early NAEP also entailed out-of-school samples, so that findings, reported by age and not grade, pertained to the entire age population. Some of these features were dropped ostensibly because they were costly. Others were changed by NAGB. The changes all had consequences, allowing findings to focus more on selected topics (e.g., reading, math, science), to pertain to grade instead of age, and to be adapted to purposes of accountability—all to Tyler's deep regret."

added to the refinement of education policy either for the Congress or the administration of an individual school. In well-crafted story-telling, Jones and Olkin portray how NAEP has been a creature of its times, educationally and politically. But there is more to think about.

Should scholastic ability tests be the Nation's Report Card?

What is a report card anyway? My dictionary first says it is "a report on a student periodically submitted by a school to parents or guardian." And then generalizes, "an evaluation of performance." We might suppose a Nation's Report Card to be an assessment of how its students are performing, but we seize the implication that it tells how well the nation's school system is performing.

It is reasonable to expect that a national report card will tell about the quality of both teaching and learning. Common pedagogic chatter has it that "you haven't taught if they haven't learned"—implying that a good assessment of learning would be a good assessment of teaching. But it is obvious that children learn more outside school than in, so the degree to which the children are becoming educated is not a good assessment of school teaching and learning. Even in school, children learn much from each other, and learn much the teacher did not intend to teach, and many teachers arrange experiences so that children learn independently and beyond intention. Scholastic learning and teaching need to be assessed separately, and both are worthy entries in a report card, and perhaps non-scholastic learning as well (Linn, in press).

We would like to have valid assessments of the growing (and sometimes deteriorating) competence of our children. And we need assessment of the quality of our schools, and of different components of schooling. Our needs for assessment are several. One assessment cannot substitute for all the others. The several assessments of progress need to be thought of separately and thought of together.

Not everything that children are learning in school needs to be examined, but some does. (Practically, the assessment of student learning is carried out partly because it assists the management of schools and aids in the pacification of parents. NAEP assessment of learning is seldom used as input for classroom teaching.) Most needing of assessment are three major arenas of student academic sophistication. One is indication of the child's scholastic knowledge. Another is the child's readiness for further learning (scholastic aptitude). And a third is the experience that a child has for which the school is responsible. The three can be packaged in many ways, but any summarization on a Report Card, for child or nation, will be simplistic.

Knowledge. For political reasons more than educational, some knowledge gets classified as essential. States and school districts (of the U.S. and many places) identify essential knowledge in statements of "learning standards." These standards are believed to be uniform, alluded to as applicable to each and every student. And test items are written to provide rankings and sometimes a cut-score—above or below a level of knowledge seemed minimally desirable—to separate for possibly legitimate purpose those who have learned at that level from those who have not, about topical areas such as economic geography, geometric proofs, and English poets. The topics are clustered into subject matter domains.

Most school report cards identify at least half a dozen subject matter domains, such as social studies and language arts, covering a vast range of topics and subtopics, some taught, some not.

As the book made clear, NAEP people repeatedly faced the question of whether it is fair to assess children on topics not taught. Certainly the parents and the system deserve to know what has been taught and what of that has been learned. But teachers do not teach all children the same; some get more instruction than others, many do individualized projects. Most teachers try to be fair, but they cannot teach each child the same, partly because no two children are equally ready to learn what is being taught. Fairness is sought by moving each child a little further, but some will spurt ahead and others lag behind, even when the teacher tries hard to distribute opportunity evenly.

State Standards provide little guide as to what should be the targeted development of the content of a standard (Cole, in press). District guidelines, textbooks, tests, and professional practice add some uniformity to the coverage, but the variance is great from classroom to classroom, just as it is from student to student. One teacher satisfies a standard about the Westward Movement by drawing attention to a paragraph whereas another assigns children a project on national expansion and cultural assimilation. This variety is inevitable. It can become an administrative headache when teachers argue the differences, not just because most teachers crave a certain freedom from supervision but because, to a considerable extent, teachers should be allowed to pursue ways of teaching and topics they themselves know they can teach best. There is no national or personal good to have everyone taught the same; yet political advantage and a desire for equity of educational opportunity press us to speak as if good education is standardized.

Thus, on a refined report card, a student's knowledges or all students' knowledge should be portrayed regarding more than a few content domains, with as many subdivisions as are needed to show that different things are and should be learned by different children; yet to show that the divisions of academic discipline that almost everybody recognizes are still respected. Too many content areas on the report card will offend the reader and may imply a precision not attained. Because of equating difficulties, it seldom will be accurate to compare levels from one subject matter to another, such as U.S. history to geography. The levels and differences on the report card will only be an estimate. The use of standardized tests adds to the precision of measurement, but what is tested is only a weak sample of what could be tested. In school, grading is an art, not a technology, and can be informative in the hands of a teacher not using grades as an instrument of control. Nationally, assessment is also an art (Glass, 2003) and is most informative when scores are not being used as an instrument of control.

Goslin's record of early NAEP meetings (p 554) shows that questions were discussed about the number of subject matters to test, but not the curricular issues raised in the paragraphs above. The early plan was to test for important task knowledge, not to represent subject matters, but those tasks were rejected as not relevant to policy. To Cronbach's dismay, the thinking switched to conventional achievement testing, where highly correlated items directed assessment toward a centroid of topics. It is just as easy to think of this centroid as an index of scholastic ability as a sample of subject matter topics,

and the research was not done to assure that NAEP was measuring achievement rather than scholastic ability. Questions of content coverage were raised in the subcontractor's development of exercises but almost never by the 29 strategists talking and writing for this version of NAEP's history.

Scholastic Ability. Scholastic ability is the ability to learn in school, i.e., to profit from academic experience, including learning to decode instructions and solve problems. It is partly a function of native intelligence, the immeasurable capabilities drawn from genetic codes, when given ample opportunity to develop. But the fact is that much of life in the more privileged sites, as well as in the most impoverished, constrains and enhances the development of ability in important ways. The effects of nurturing are great. Nurturing is visibly the domain of mothers and fathers, but siblings, peers, extended family, then teachers, counselors and coaches, plus social and work groups and all the rest of our culture contribute as well to scholastic ability. It is an intriguing but futile speculation to attribute ability to certain experiences and persons, but not a futile responsibility for persons to provide opportunity for each child to become more able.

A hundred years ago we had intelligence tests, purportedly to measure native ability as it had developed in our cultures, and a few efforts to make culture-free intelligence tests. Partly because there is no justification for calling either of them natural or culture-free, and because efforts to declare some groups more intelligent than others are offensive, the name of the tests was changed from intelligence tests to scholastic ability tests. Calculation of intelligence or ability always was recognized as needing to take age into account, acknowledging that mental ability grows, and assuring that what was intelligence would vary in kind as well as amount. There is something called intelligence. We all recognize some aspects of it, but almost no researcher continued to follow the thinking of Charles Spearman (1927), who spoke of a g-factor, a general intelligence. Howard Gardner (1983) sorted out a number of intelligences: linguistic intelligence, musical intelligence, bodily-kinesthetic intelligence, and six others at my last count.

I have heard of report cards that recognize this diversity of intelligences, but most report cards focus on performance in subject areas rather than skills. Much schooling is aimed at skill development, including in language, music, physical development, and deportment. It is apparent in many statements of academic standards that skills are as important as knowledge, but the practice has been to include skills within the content domains rather than to list them separately. There seems good reason why a national report card would identify a few of the skills that are not subject matter specific, or even taught-for directly, such as graphic skills, coaching skills, and social skills. But the NAEP planners were not inclined toward taxonomies of education, but instead sought indicators that would assist policy people in making educational decisions.

It is not clear that NAEP's indicators would be better guides to policy setting and assessment of the school system if they reflected taxonomically the complexity of education and schooling. Whatever their ingredients, indicators come to have meaning as they are used over time. They are likely to be more respected, less opposed, if they have face validity. NAEP purports to assess student achievement, not the other aspects of educational

progress. Still, NAEP seems little hurt by having ignored assessment of teaching and administration. NAEP is a narrow indicator but it was not promised to be broad.

As to utility, NAEP findings are almost never a part of teacher discourse or course preparation. They do not figure in school or district deliberations, although sometimes bolstering an already chosen line of argument. The 1999 Committee on the Evaluation of National and State Assessment of Educational Progress appointed by the National Research Council (Pellegrino, Jones and Mitchell, 1999 p 27) identified eight different ways the 1996 mathematics and science findings were used: "to make descriptive statements, to serve evaluative purposes, and to meet interpretive ends." The evidence of use was not presented. No single case was given of a policy decision resting on NAEP data. It appears that NAEP findings are rhetorical, seldom illuminative. Today we see NAEP data compared to state test averages, probably lowering the credibility of state findings (Kiplinger, in press). It is good to have data for amplifying our problems, extending our discussions, but that was not the policy value promised. Many of the voices in the book spoke of NAEP as successful, usually referring to its technical advances. Indeed it has survived and enjoyed media respect. But the book cited little evidence that NAEP has contributed to the maintenance of our education system, or slowed its demise. And it did not raise the question of NAEP being held accountable for its consequences.

NAEP is alluded to be, but is not, a National Report Card. It provides data on curriculum-specific scholastic aptitude, a correlate of gross achievement in a subject-matter domain, but not its measure. It tells us more about how grandly educated and miserably educated the nation's children may become than about how educated they have become.

Educational consequences. NAEP has been evaluated externally and internally, sometimes raising the question of "How has it helped?" But not, to my knowledge, "How has it hurt?" Marshall Smith testified that NAEP had provided data that the Clinton White House used in selling decisions it had already made. The formal meta-assessments across the years found NAEP not as useful as it should be, though the standard of "usefulness" was left vague. Most people, it seems, support the effort to fix it so that it does help. But the answer remains, "It hasn't helped much."

I do not recall anyone else claiming that the Nation's Report Card has hurt the nation. I do not know of poor education policy that has been set because NAEP data pushed the decision one way and not another. It is not apparent that NAEP has provided data giving support or pause to an education reform. NAEP data have been used in research but it is not apparent that NAEP procedures or data caused educational researchers to do better or worse work. So Ralph Tyler's original aims for NAEP have neither robust support nor opposition in evaluative reviews or user experience. Absence of consequences seems to be the meta-assessment conclusion.

But, if NAEP is to be a primary indicator of educational progress, it should be banded together with other determinants of education. NAEP should take some credit and blame. If the schools are bad, this messenger may bear some of the blame. Intentionally, NAEP has influenced the meaning of the news. Intentionally, it has influenced what people expect educational progress to be. It has been the aim of many educational and political

leaders to make the schools more responsive to the wishes of the people, and technology has facilitated the communication, making it easier to know the educational planning and activity. But technology also has hurt communication because the language of public communication is generally much simpler than the language of teaching and learning.[10]

Perhaps the most serious claim against NAEP is its systemic unintended co-conspiracy with other technical structures that come with large-scale management. The structures increase distant covert impersonal control and sometimes bring alienation. Decisions made according to indicators are seen to lack compassion and adaptation to the local culture. NAEP extended the mysterious monitoring of human affairs in contemporary society. With NAEP help, the view of little progress in education is widespread, and low is the willingness to support local efforts at school reform.

But a more direct claim is that, with other standardized testing, NAEP has been a bad influence on pedagogy and curriculum. As NAEP styles its items, so do other test makers, and teachers are increasingly drawn to teach accordingly. As it identifies its topics, teachers adjust and conform, a little more resistant to the voice of conscience saying teach more deeply and more personally into the subject matter. The curriculum narrows and discourages diversity of learning.[11] Teachers are hired and teachers are trained increasingly in response to state standards, state testing, state-approved textbooks, and other standardizing influences such as NAEP. No, I know of no hard evidence that NAEP is a perpetrator or even accessory to the harm that education has suffered in these ways--nor evidence against.[12]

Who should be doing the studies that assure us that high-stakes standardized testing, with NAEP as flag bearer, is not curbing educational progress? These are consequential validity studies that should be a regular part of psychometric work and the evaluations of NAEP. The history of NAEP is saddest in its lack of attention to the question whether teachers and policy makers are influenced to make poor decisions on the basis of what they know about NAEP.

---

[10] In the body politic, many of the hurts are but cuts and bruises. The problems of teaching limited-English-proficient students are probably exacerbated more than relieved by high-stakes achievement testing. It could be argued that knowledge of the high percentage of students scoring below "basic levels" adds to the national malaise more than the low competence itself.

[11] Many advocates of school change have urged a narrower curriculum, especially to curb teacher choices as to what will be taught. And standardized testing has been a principal tool for the advocacy. From a survey of the Center on Educational Policy, Jack Jennings (2006) reported that 71% of the schools were reducing time spent teaching subjects not tested. Diane Ravitch (2006), Seymour Sarason (1990), and many others (Pellegrino, Jones & Mitchell, 1999) have lamented the narrowness.

[12] When these harms are assessed today, great damage is attributed to the federal No Child Left Behind program. NCLB and NAEP are weakly linked, but, as I see it, they affect educational progress in similar ways. NCLB was given little attention in the Jones & Olkin book.

Is a rational National Report Card possible?

For some, the most serious problem with standardized testing is its singularity of voice, its quiet demand that education be evaluated in a single way. Partly because testing's voice and style are declarative and legalistic, we fail to take full advantage of the wisdom and diversity of the professional people in and around education. Relying too much on technical and bureaucratic languages denies public insight into the depth of understanding and uncertainty behind any official report of the quality of education. For teachers and citizens, standardized testing is mystical, not rational.

The practice of high-stakes testing in America and elsewhere has accompanied an effort to treat teaching and learning in a simple but fair manner. But education, as said here repeatedly, is hugely complex, partly because of inequitable distribution of opportunity. High-stakes testing— including NAEP testing—detracts as much as assists our mechanisms of review, meta-evaluation, and the validation implied in our professional standards. It curbs our efforts to be rational about the assessment of teaching and learning.

We need to recognize the hazards in education. When we allow further standardization of education, when we reform education, we need challenges from multiple viewpoints as to the costs and benefits for the children (Jones, 1997). Education requires decisions as to how children, teachers, and schools will be sustained and promoted; but it often appears that measurement and testing increase unnecessarily and hurtfully the formal decisions made (Berliner & Biddle, 1996). Collectively and dialectically, we need to do the research, evaluation, and public and professional deliberation that illuminates the roles of standardized achievement testing. And we might see that illumination as part of the Nation's Report Card.

Within the inventive, aspiring, affluent culture of the last fifty years, the creation of NAEP was deeply reasoned and elegant, the measurements profession at its "shining best." It responded to the metric and meritocratic appetites of the nation. Since its formative years, NAEP has helped legitimate state-by-state orientation to mandated testing. In later years it became a tool of political aggrandizement. As to its original purposes, it has not done very well.

With strong support from the measurements community, the main characters in this book about NAEP created NAEP in their own image. They wanted it to be the best that they could be. To be pure assessment, they disdained curriculum experts and philosophers. But they failed to demand validation of the assessment's core policy. At first, the core policy was tracing performance over time, but gradually the core policy became test-based accountability.

Today, the measurements community is cognizant of and explicit about the shortcomings of NAEP and standardized testing (Linn, in press). At the January, 2007 CRESST conference honoring Bob Linn's retirement, the papers zoomed in on the technical, administrative and social problems of test-based accountability (Ryan & Shepard, in press). Nationwide, misinformative and miseducational practices are prominent, pressed

by the accountability ethic. According to Lorraine McDonnell (in press), the core policy of test-based accountability in the USA is deeply established and remains unchallenged politically by any other ethic of educational progress.

So, it is my speculation that the history of NAEP would not have changed much had curriculum scholars been fully involved in the development of NAEP and had education and measurement professionals waged a more vigorous fight against unvalidated interpretation. Nationally, education is deeply political. Hundreds of thousands of educators, legislators, and citizens wanted to do what they could to help education. They wanted to believe we were creating good indicators of educational progress. Steered away from trusting the nation's teachers, and like Terrel Bell, presuming a faulty wall chart (read "report card") was better than none at all, they cottoned to test-based accountability, with NAEP's view of educational progress. A simple view of education is a valuable political tool. No, in such a nation, I don't believe a rational national report card is possible.

References

Alexander, Lamar; & James, Thomas (Eds.) (1987). *The Nation's Report Card: Improving assessment of student achievement.* Cambridge, MA: National Academy of Education.

Bell, Terrel. (1983). National Commission on Excellence in Education, 1983. *A nation at risk: The imperative for educational reform.* Washington DC: US Government Printing Office.

Berliner, David C. & Biddle, Bruce J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* Reading, MA: Addison Wesley.

Cole, Nancy (in press). Discussion of the context of educational accountability. In Ryan, Katherine & Shepard, Lorrie. *The future of test-based accountability. CRESST Conference honoring the retirement of Robert Linn*. Mehwah, NJ: Erlbaum.

Gardner, Howard. (1983). *Frames of mind: The theory of multiple intelligence*. New York: Basic Books.

Glass, Gene V (2003). Standards and criteria redux. Retrieved January 15, 2007, from http://glass.ed.asu.edu/gene/papers/standards/

Goslin, David. (1964). Summary report: Two conferences on a National Assessment of Educational Progress. Carnegie Corporation. Pp 551-559 in Jones, Lyle V. & Olkin, Ingram. (2004). *The Nation's Report Card: Evolution and perspectives.* Bloomington, IN: Phi Delta Kappa Education Foundation.

Jennings, Jack; & Rentner, Diane. (2006). Ten big effects of the No Child Left Behind Act on public schools. *Phi Delta Kappan*. October, 2006. Retrieved January 15, 2007, from http://www.pdkintl.org/kappan/k_v88/k0610jen.htm.

Jones, Lyle V. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher, 25,* 7, 15-22.

Jones, Lyle V. (1997). National tests and education reform: Are they compatible? The William H. Angoff Memorial Lecture, October 8, 1997, Educational Testing Service. http://www.ets.org/Media/Research/pdf/PICANG4.pdf.

Jones, Lyle V. & Olkin, Ingram (Eds.) (2004). *The Nation's Report Card: Evolution and perspectives.* Bloomington, IN: Phi Delta Kappa Education Foundation.

Kiplinger, Vonda (in press). Reliability and equating of assessments. In Ryan, Katherine & Shepard, Lorrie. *The future of test-based accountability. CRESST Conference honoring the retirement of Robert Linn.* Mehwah, NJ: Erlbaum.

Linn, Robert L. (in press). Educational accountability systems. In Ryan, Katherine & Shepard, Lorrie. *The future of test-based accountability. CRESST Conference honoring the retirement of Robert Linn.* Mehwah, NJ: Erlbaum.

Linn, Robert L.; Baker, Eva; & Burstein, Lee. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics.* Los Angeles: UCLA, CRESST.

McDonnell, Lorraine (in press). The politics of educational accountability. In Ryan, Katherine & Shepard, Lorrie. *The future of test-based accountability. CRESST Conference honoring the retirement of Robert Linn.* Mehwah, NJ: Erlbaum.

Messick, Samuel J. (1989). Validity. Pp. 13-103 in Linn, Robert. (Ed.) *Educational Measurement, 3rd Edition.* Washington, DC: American Council on Education.

Messick, Samuel J.; Beaton, Albert; & Lord, Frederick M. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era.* Princeton, NJ: Educational Testing Service, Report 83-10.

National Commission on Excellence in Education, (1983). *A nation at risk: The imperative for educational reform.* Washington DC: US Government Printing Office.

Pellegrino, James; Jones, Lee; & Mitchell, Karen (Eds.), (1999). *Grading the National Report Card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, DC: National Academy Press.

Ravitch, Diane. (2006). The fall of the standards-bearers. *School and Colleg.* Retrieved 1/15/2007 from http://schoolandcollege.com/articles/2006/03/01a04401/index.html.

Ryan, Katherine & Shepard, Lorrie. (in press). *The future of test-based accountability. CRESST Conference honoring the retirement of Robert Linn.* Mehwah, NJ: Erlbaum.

Sarason, Seymour. (1998). *Political leadership and educational failure*. San Francisco: Jossey Bass.

Shavelson, Richard; McDonnell, Lorraine; Oakes, Jeannie; & Carey, Neil. (1987). Indicator systems for monitoring mathematics and science education. Santa Monica, CA: RAND

Shepard, Lorrie A. (1995). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. In Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board (NAGB). Washington, DC: National Assessment Governing Board and National Center for Educational Statistics.

Spearman, Charles. (1927). *The abilities of man.* New York: Macmillan.

Stake, Robert E. (1970). National assessment. Pp 53-66 in Glass, Gene V (Ed.). *Proceedings of the 1970 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

Stufflebeam, Daniel; Jaeger, Richard M. & Michael Scriven (1992). A retrospective analysis of a summative evaluation of NAGB's pilot project to set achievement levels on the National Assessment of Educational Progress. Paper at the Annual Meeting of the American Educational Research Association, April 20, 1992.

Tyler, Ralph W. (1966). The objectives and plans for a National Assessment of Educational Progress. *Journal of Education Measurement, 3*(1), 1-10.

About the Author

Robert E. Stake

Robert E. Stake is Emeritus Professor of Education and Director of CIRCE at the University of Illinois. Since 1965 he has been a specialist in the evaluation of educational programs, moving from psychometric to qualitative inquiries. Among the evaluative studies he has directed are works in science and mathematics in elementary and secondary schools, model programs and conventional teaching of the arts in schools, development of teaching with sensitivity to gender equity; education of teachers for the deaf and for youth in transition from school to work settings, environmental education, early childhood education, and special education programs for gifted students, and the reform of urban education. Stake has authored *Quieting Reform,* a book on Charles Murray's evaluation of Cities-in -Schools; four books on methodology, *Evaluating the Arts in Education, The Art of Case Study Research, Standards-Based and Responsive Evaluation,* and *Multiple Case Study Analysis*. For his evaluation work, he received, in 1988, the Lazarsfeld Award from the

American Evaluation Association, and, in 1994, an honorary doctorate from the University of Uppsala. Email: stake@uiuc.edu

Editors

http://edrev.asu.edu