Economic Policy Institute

Presentation | February 26, 2014

WHAT NAEP ONCE WAS, AND WHAT NAEP COULD ONCE AGAIN BE

BY REBECCA JACOBSEN AND RICHARD ROTHSTEIN

A presentation to the National Assessment Governing Board's 25 Anniversary Celebration, February 26, 2014, Washington, D.C.

Contemporary education policy, whatever else it may or may not have accomplished, has narrowed the school curriculum by holding schools accountable primarily for their student scores in math and reading. This narrowing is incontrovertible, revealed in surveys and confirmed by logic – unless you believe that teachers and administrators, unlike all other actors, behave irrationally, accountability for only some of the many goals of education must create incentives for schools to pay more attention to goals for which they are accountable, and less attention to those for which they are not.

Social scientists have documented this problem in virtually every field of human endeavor – health care, labor market policy, criminal justice, transportation, and so on – and it was most famously summarized in 1979 by Donald Campbell's "Law of Performance Measurement" – "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

Examples of Campbell's Law are widespread – when Medicare held cardiac surgery practices accountable for the survival rates of their patients, physicians turned away the sickest patients who most needed surgery but whose survival rates were lower; when the Department of Labor held employment offices accountable for the proportion of jobseekers placed

in employment, the offices turned attention away from training programs for higher-skilled, longer-lasting positions and towards easier-to-fill, short-term, poorly-paid, unskilled jobs; when cities hold police officers accountable for writing more traffic tickets, you get speed traps; and when schools are held accountable for math and reading scores, less attention is paid to history, science, citizenship, physical education, oral presentations, cooperative learning, the arts and music.

In surveys, the American public, school administrators, school board members, and state legislators all insist that they want schools to pursue multiple goals; basic academic skills are important, but so are critical thinking, the arts, preparation for skilled work, social skills and a work ethic, good citizenship, and physical and emotional health. Yet schools suffer no sanctions for ignoring all other goals, provided math and reading scores improve.

NAEP has contributed to this distortion, by its highly publicized biennial release of national and state-level results in math and reading alone. Because of NAEP's focus and influence, we have little sense of how American schools perform in any other areas. For a public education system on which we spend more than half a trillion of public dollars annually, this is inexcusable.

But it was not always thus, and it need not be thus in the future.

NAEP was first designed in the 1960s by a team led by Francis Keppel, John Gardner, and Ralph Tyler. In a 1949 book, Tyler proposed that evaluation of education should not rely exclusively on standardized test scores but "must appraise the behavior of students since it is changes in these behaviors which is sought in education." He recommended that an assessment be administered after students leave school because material not well taught may be "rapidly dissipated or forgotten." While agreeing that some skills can be assessed with paper and pencil tests, he insisted that other objectives of education, such as social skills, are "more easily and validly appraised through observations of children under conditions in which social relations are involved," both when they are playing and working together. Evaluation should include collection of actual products made by students, such as paintings or samples of writing. He suggested that if a school's reading program aimed to develop students who had increasingly broad and mature interests, evaluators could assess this by seeing what books students checked out of the library. In a 1963 memo to Commissioner of Education Keppel, Tyler explained how such approaches might be developed for a national assessment.

Keppel appointed a design committee, headed by the Carnegie Foundation's Gardner, that decided the NAEP should assess any goal area for which schools devote "15-20% of their time…," the "less tangible areas, as well as the customary areas, in a fashion the public can grasp and understand." These should include all areas which "thoughtful laymen consider important for American youth to learn;" in art, citizenship, career and occupational development (including vocational education), foreign language, health and physical fitness (including such aspects of emotional health as self-image and self-confidence), literature, mathematics, music, reading, science, social studies, and writing. The citizenship area covered more than political knowledge; it also covered ethical principles and interpersonal behaviors that committee members believed comprised democratic citizenship. Consumer protection was also a subject of assessment; the committee felt that the ability to resist false advertising claims, and budget appropriately, was an important life skill with which youth should emerge from schools.

Following Ralph Tyler's insistence that educators should assess the behavioral outcomes of education, not only the abstract skills that *might* lead to such outcomes, the committee designed survey questions about behavior as well as tests

of skill. Schools, committee members agreed, don't teach reading skills as an end in itself. Schools want students to use the skill by reading a newspaper, for example, and effective teaching must lead to such use. An assessment of outcomes should determine not only whether students have basic reading skills, but whether, as they grow older, students actually read newspapers. This behavioral outcome does not stem from direct instruction in language arts classes, but also from classes in other curricular areas, social studies, for example.

In civics education, the NAEP design committee was also interested in assessing behavior as well as factual knowledge. It included whether students showed concern for the welfare and dignity of others, supported rights and freedoms of all, helped maintain law and order, knew the main structure and functions of government, sought community improvement through active democratic participation, understood problems of international relations, took responsibility for their own personal development, and helped and respected their own families.

Attempting to assess each of these goals meant that paper and pencil tests could not be exclusively relied upon. Some outcomes of schools could only be assessed by observation of student behavior, or by survey techniques that verify the activities in which students engage.

Survey-type questions, which are critical for gathering certain types of information, may also produce inaccurate data because when students or adults are asked about things they have done, there is a tendency for respondents to exaggerate. Therefore, early NAEP assessors were trained to ask follow-up questions to make certain that accurate data were obtained. For example, in a citizenship assessment, if a respondent answered that he or she had written a letter to a Congressman, the NAEP assessor then asked for details, such as what the letter was about, to ensure that a real event was being described.

Other goals of education, such as social skills and work ethic, must be observed in order to assess their development. To see whether students were learning to cooperate effectively in small groups, NAEP sent trained observers to present a game to 9-year-olds in sampled schools. In teams of four, the 9-year-olds were offered a prize (such as crayons or yo-yos) to guess what object was hidden in a box. Students could ask yes-or-no questions; two teams competed to see which could identify the toy first. Cooperation was necessary–all team members had to agree on which question to ask, and the role of posing questions was rotated. Trained NAEP observers rated the students on whether they suggested a new question, gave reasons for their viewpoints, sought additional information that supported the team's work, helped organize the procedure, or otherwise demonstrated cooperative problem-solving skills. Students were also rated on whether they impeded teamwork–for example, by making discouraging or irrelevant comments. NAEP then reported on the percentage of 9-year-olds who were capable of cooperative problem solving.

NAEP assessors also gave cooperative exercises to 13- and 17-year-olds. Assessors presented groups of eight students with a list of issues about which teenagers typically had strong opinions. Students were asked to reach consensus on the five most important issues and then write recommendations that the group supported on how to resolve at least two of them. The list included, for 13-year-olds, such issues as whether they should have a curfew for going to bed, whether they should be allowed to watch movies with adult content, and whether parents should have the right to approve their choices of friends. For 17-year-olds, the list included compulsory school attendance and military service requirements as well as the age eligibility minimums for voting, drinking, and smoking. As they did with 9-year-olds, NAEP observers rated whether students took clear positions, gave reasons for their points of view, sought additional relevant informa-

tion, helped organize internal procedures, or defended a group member's right to hold a contrary viewpoint. They also noted whether students demeaned the group's work or did something totally unrelated to the task.

NAEP also assessed the development of attitudes considered essential to our democratic way of life. For example, NAEP attempted to determine whether students understood that individuals should be judged on their own merits and not be held responsible for others' misdeeds. Interviewers asked 9- and 13-year-olds whether, if the father of a friend was jailed for theft, they would still invite the friend to their homes to play. To assess students' commitment to free speech principles, 13- and 17-year-olds were asked if they thought that someone should be permitted to say on television that "Russia is better than the United States," that "Some races of people are better than others," or that "It is not necessary to believe in God." NAEP reported that only 3 percent of 13-year-olds and 17 percent of 17-year-olds thought all three statements should be permitted.

The results of these assessments weren't satisfactory. NAEP's national report showed, for example, that only 4 percent of 13-year-olds defended the right of another group member to voice a different opinion, and only 6 percent were willing to defend their own viewpoints in the face of opposition. Have we improved since then? We have no way to know.

In 1969, the United States was in the throes of a civil rights revolution, so NAEP assessed whether schools were preparing young people for responsible citizenship in this context. NAEP interviewers asked 13- and 17-year-olds what they believed they should do if they saw a public park attendant barring children from entering because of their race. NAEP reported that 82 percent of 13-year-olds and 90 percent of 17-year-olds knew they should do something, such as tell their parents; report it to a public authority or to a civil rights or civil liberties organization; write letters to the newspaper; or take social action, such as picketing or leafleting.

Early NAEP assessed whether 17-year-olds were able to consider alternative viewpoints by asking them to state arguments both for and against one of the most heated public issues of the time–whether students enrolled in college should be drafted. One question asked 9- and 13-year-olds if something might be false, even if it was reported as being true in the newspaper.

NAEP assessed social responsibility in more private situations as well. In 1977, 17-year-olds were asked, in a multiple choice exercise, what they should do if they noticed that their friend's 6-month-old baby had bruises. The correct answer was "Suggest that your friend call her baby's doctor about the bruises." Incorrect choices included "Ignore the bruises because they are none of your business" and "Accuse your friend of beating her child." Follow-up prompts, however, explained that on a subsequent visit, when the baby still appeared to be bruised, the friend said the baby had fallen out of her crib. The prompt asked what you should do next, with the correct answer being, "Call the local child health agency and report your suspicions."

Just as NAEP then was more comprehensive in *what* it covered, so too was it comprehensive with regard to *whom* it covered. The committee recognized that if the assessment was administered only in high school settings, some students would be missed because they had dropped out. Indeed, the greater the number of students who were no longer in school, the more likely the assessment results would inaccurately reflect the average abilities of the full age cohort. Therefore, an assessment of youth in their late teens must include students who were no longer in school as well as those who were still enrolled.

One member of the design team was Peter Rossi, a prestigious sociologist who headed the National Opinion Research Center. Rossi observed that "[a]lthough the immediate goal of school systems is to produce graduates with appropriate skills, knowledge and values, the long range goal is to produce an adult population with levels of these attributes sufficient to meet the needs of our society." Supporting Ralph Tyler's convictions, Rossi observed that educators often face the problem of decay – students may learn something that enables them to pass a test in the short run, but does not stay with them after the school year is over. Therefore, Rossi argued, to truly measure whether an educational system is effective, young adults must also be assessed, to see what has stuck with them a few years after they left school. A school may produce lower tested achievement for 13 year-olds, but if its graduates have better achievement as young adults, it may be a superior school. "In short, whether or not a school system is an effective institution depends as much on whether its products as adults retain enough of the knowledge, skills and values that were characteristics of graduates."

The design team was explicit that it did not want to model NAEP on the standardized tests then available and used by American schools. Rather, it wanted to model NAEP on the kinds of surveys conducted by the Bureau of the Census, the Bureau of Labor Statistics (BLS), or the Centers for Disease Control (CDC). These surveys provide data that citizens and policymakers must interpret themselves. No overall scores are provided; nobody passes or fails the Census or the National Health Survey. Rather, when official surveys report, for example, that unemployment was higher this year than last, or that the diabetes rate among women was rising from one survey to the next, policymakers can take this information to design improvements in job training programs or public health campaigns.

The design committee wanted NAEP to serve a similar function for education. If NAEP were to report that the percentage of 13 year-olds who could calculate the area of a triangle had declined, this should spur policymakers to examine mathematics instruction to see where it might be falling short. Or if NAEP were to report that the percentage of 17 year-olds who knew the correct technique for applying artificial respiration had increased, this should encourage policymakers that this aspect of health education was improving. Gardner and his fellow committee members believed that such percentages would be easily understood by the public, which could (with its representatives) then draw its own conclusions about educational policy from these numbers.

Thus, the only reporting in which early NAEP planned to engage was the kind of reporting done by the Census, BLS, or CDC. NAEP would release no overall scores, but only specific test questions or activities, along with the percentages of students who could successfully answer the question or conduct the activity, broken down by region, gender, poverty status, and race.

The NAEP design was an entirely mainstream effort. It was broadly supported, not only by educators but by politicians and policymakers. In 1970, as the design was first being implemented, President Richard Nixon called for an accountability system, not simply a NAEP-like survey, that would embrace the same principles. An educational accountability system, he said, should "pay as much heed to what are called the 'immeasurables' of schooling… such as responsibility, wit, and humanity as it does to verbal and mathematical achievement."

Assessment of behaviors and attitudes across a wide range of educational outcomes, a focus on retention of knowledge, skills and behaviors into adulthood, and reporting that enables public deliberation – these principles all guided early NAEP. The assessment was first administered nationwide in 1969 and continued pretty much according to the original design through 1973. NAEP assessed 9, 13, and 17 year-olds, as well as young adults (ages 26 to 35), regardless of what

grade students were in, or even of whether they were still enrolled in school. For students not in school, trained NAEP staff assessed a random sample of youth and young adults of the appropriate ages in their homes and communities. NAEP covered all important outcomes that schools aimed to produce, some of which could be assessed by paper and pencil tests, and others of which could only be assessed by trained observers who recorded student behavior. There were no scores; the national results reported were percentages of students or young adults who successfully completed specific exercises, broken down by region, gender, poverty status, and race. Because results were meant to spur conversation about our nation's schools, educational experts were invited to write commentaries based on their review of the results.

Certainly, if we used results such as these–and not just math and reading scores–to evaluate our school systems, incentives would shift. National reporting of low scores on the free speech questions, for example, might spur some members of the public to demand that schools do a better job on citizenship. Such pressure might reduce the incentive to drop cooperative learning in favor of test preparation in math and reading.

However, NAEP never got to develop these approaches beyond some early implementation and reports. Congressional leaders were not persuaded that such a census of educational outcomes could be used to improve policy and that it was worth the expenditure in a time of fiscal restraint. In 1974, Congress cut NAEP's annual budget in half, and important design elements in the original NAEP were dropped. After 1974, NAEP eliminated the young adult sample, although it was briefly reinstated for one year in 1977. After 1976, NAEP ceased assessing out-of-school 17 year-olds. NAEP ceased observing behavioral outcomes and, with very rare exceptions (for example, an arts and music assessment was conducted in 1997), NAEP became exclusively a pencil and paper test.

Statistical sophistication became a higher priority than citizen understandability. ETS began to report overall subject-area scale scores that are less intuitively meaningful to the public than the percentage of students who performed a certain task. It is more difficult for the public to deliberate whether a score of 250 fails to meet educational expectations, and such a score provides little guidance regarding what areas specifically need improvement. In the early 1990s, Congress again began to increase the NAEP budget, but the new money was now mostly used to increase the frequency of math and reading tests and to increase the sample size to support state-level results in basic skills alone, rather than return to assessing a broader range of content.

In 1996, the National Assessment Governing Board issued a policy statement explicitly asserting that NAEP "tests academic subjects and does not collect information on individual students' personal values or attitudes." Although information on individual students was never at issue, the new policy statement made clear that NAEP would steer away from any assessment of the values and attitudes that have always been central goals of American education.

There have certainly been protests against this change in direction. In 1982, NAEP commissioned former Secretary of Labor Willard Wirtz, and educator Archie Lapointe to evaluate the assessment and make recommendations. The Wirtz-Lapointe report recommended reinstatement of the out-of-school assessments of 17 year-olds and young adults, and a reinstatement of the broad goal-area coverage that had been abandoned to focus more narrowly on inexpensive-to-test basic academic skills. Continuing NAEP with such a narrow focus, Wirtz and Lapointe said, "would not make sense" and "will virtually assure the Assessment's not playing a major role in informing the general public about the educational achievement picture…. Covering subject areas not covered by other assessment systems is critical." Not doing so "would contribute to the narrowing of American education."

Four years later, the Secretary of Education appointed another study group to evaluate NAEP, chaired by then-Governor of Tennessee Lamar Alexander. Hillary Rodham Clinton, then a lawyer and wife of the Arkansas governor, was a member. The group's 1987 report protested the narrowing of subject areas NAEP was assessing to basic academic skills, and urged that coverage be re-expanded, especially to include "higher-order cognitive skills," often "involving value judgments of some subtlety," which had been minimized as NAEP moved more towards easily quantifiable and tested basic skills. The study group also urged that the out-of-school 17 year-old and young adult population assessments be restored. In a companion report to Alexander's, the National Academy of Education concluded that NAEP's narrow coverage "may have a distorted impact on our schools."

The Academy's pessimistic prediction has now been fulfilled, and this early history of NAEP has become a quaint curiosity. Few officials in the U.S. Department of Education are even aware of it. But knowledge of the work of John Gardner's design committee, and of NAEP's experiences during its first decade, should be revived. They illustrate how assessment could be used as part of a balanced accountability system for education, a system upon which the public could rely to learn if schools truly perform satisfactorily and where attention to improvement should be directed.

About the authors

Rebecca Jacobsen (rjacobs@msu.edu) is an Associate Professor of Education at Michigan State University. Richard Rothstein (riroth@epi.org) is a Research Associate of the Economic Policy Institute and a Senior Fellow at the Chief Justice Earl Warren Institute on Law and Social Policy, University of California (Berkeley) School of Law. This presentation is drawn from their 2008 book, in collaboration with Tamara Wilder, *Grading Education: Getting Accountability Right*, published by the Economic Policy Institute and Teachers College Press. Bibliographic source citations for all claims in this presentation can be found in the book, http://www.epi.org/publication/books_grading_education/