Reflections on the Future of NAEP:
Instrument for Monitoring or for Accountability?

CSE Technical Report 499

Lauren Resnick
CRESST/University of Pittsburgh, LRDC

February 1999

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Construct Validity: Understanding Cognitive Processes and Consequences Lauren Resnick and Michael Young, Project Directors, University of Pittsburgh, Learning Research and Development Center

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student achievement, Curriculum, and Assessment, the Officer of Educational Research and Improvement, or the U.S. Department of Education.

REFLECTIONS ON THE FUTURE OF NAEP: INSTRUMENT FOR MONITORING OR FOR ACCOUNTABILITY?

Lauren B. Resnick

CRESST/University of Pittsburgh, LRDC

Abstract

The National Assessment of Educational Progress (NAEP) is now accepted as the nation's report card, but getting there took over 30 years and a convoluted process beset by technical and political issues. After summarizing the history of NAEP and its successful use as a monitoring instrument, the author discusses the ramifications of its use as an accountability instrument, including how it could negatively affect educational outcomes and the validity of the NAEP tests.

The history of the National Assessment of Educational Progress (NAEP) encapsulates and highlights a thirty-plus-year history of American education in which political and technical issues have become intertwined. NAEP has figured in public debates over whether and what kind of national presence there should be in American education—a function traditionally and constitutionally reserved to the states and, in some cases, to localities. This most fundamental of political and public concerns has influenced the ways in which technical issues—including measurement technology and psychometrics, statistical inference and sampling, and comparability of populations and of test instruments—have been addressed. Closely related to these political and technical concerns has been a set of questions about what American schools are teaching, what they ought to teach, and what role tests (especially, but not only, NAEP) should play in curriculum debates. In recent years, furthermore, the question of whether and how NAEP should set *standards* of performance for its tests has brought together all three lines of NAEP's history—national presence, technical quality, and role in curriculum—into interaction and contention.

NAEP and the Question of National Presence

Today NAEP is accepted as the nation's report card, our system for monitoring the nation's educational progress. These easily accepted terms belie the controversy surrounding the original creation of NAEP in the 1960s. Until then, the country had had no systematic data collection on the academic performance of the nation's schools, a situation that would have been unthinkable in countries with national responsibility for the public education system. In the post-World War II era, with a growing national investment in education (e.g., the G.I. Bill, the National Defense Education Act, compensatory education legislation), many people began to feel the need for systematic, nationwide evidence on educational performance. NAEP's birth coincided with the creation of a system of national centers for research on education, federal funding of regional laboratories for disseminating research knowledge, and a host of other efforts to create a research- and data-based federal presence in an otherwise state and locally guided education system.

Not everyone shared this thirst for national attention to education. Indeed, so much resistance met the idea of nationally sponsored tracking of school performance that it was politically possible to establish NAEP only by prohibiting reporting of local, or even state, results on the NAEP instruments. For more than two decades, NAEP reports were given only at the national and regional levels. The reporting regions grouped states together so that no single state's government could be held to public judgment based on NAEP data. NAEP was, thus, politically safe but also largely ignored by the public.

In the late 1980s, as education became a priority of many governors and state legislatures, states trying to bring educational issues to public attention began looking for ways to track education performance in their schools. By then, the NAEP tests had won general acceptance as non-intrusive measures of achievement: The tests were given to relatively few students, took relatively little time, and did not pretend to specify appropriate curriculum. It was a natural step for governors, legislators, and some educators responsible for state education policy to request NAEP scores for their own states. State-level NAEP results, they argued, would provide an already established, high-quality monitor of how the state was doing and would automatically allow comparisons with other states and with the country as a whole. Thus was born the "State NAEP" experiment, leading to a widespread view that, as long as NAEP stays out of the curriculum specification business, this *national*

test can be a useful tool for states in monitoring how they are fulfilling their constitutional obligations to provide effective education.

According to many educators and policymakers, the next logical step would be to provide NAEP scores for individual school districts and even individual schools. Most who call for such use think of it as little different from state-level NAEP reporting. That is, a high-quality instrument would be available for local use. It would allow comparison with national achievement levels but would not prescribe a particular curriculum and so would not interfere with local control privileges and responsibilities. Localities could benefit from a national investment in testing, it is argued, without accepting a national presence in their governance and programming.

Here, I believe, the argument fails. Using NAEP—or a test closely based on it—to track the performance of individual schools and districts would convert NAEP from a monitoring to an accountability instrument. This would create a national presence in American schools far greater than anything we have seen before. The NAEP tests would be much more influential and constraining, for example, than requirements for Title I, special education, or Goals 2000 expenditures. District- and school-level score reporting would give NAEP influence over the de facto curriculum: that is, over what is taught day-to-day and especially what is taught close to test-taking time. Opinions differ on whether such national influence would be desirable or not. But past research on the ways in which educators teach to high-stakes tests leaves little doubt that there would be such an influence, its extent dependent on what kind of stakes were attached to NAEP test performances. Furthermore, if NAEP scores were reported at district or school levels and educators began to gear their teaching to the NAEP tests, there would likely be a corruption (cf. Koretz) in its value as an independent monitor of educational progress. To see why, we must turn to a consideration of how assessments function in social systems.

Indicators Versus Incentives: The Thermostat Effect in Educational Measurement

I begin with a basic distinction that concerns how we use measurement information in social systems. Assessment data, like other measures, can function as an *indicator* of a system's health or as an integral *part of* the system, specifying actions to take in order to meet certain desired outcomes. Indicator measurements are monitors. They provide broad information on the performance of a social

system, information that can be weighed and interpreted by those who make policy decisions. Monitors inform decision making but do not force or directly encourage particular actions. NAEP has been designed as a monitor and has functioned as one until now.

Measurements can also function as an integral part of a social system, directly influencing how people in the system direct their efforts. Examples of measures designed to function as part of the education system, rather than monitors of it, include Advanced Placement exams, which students prepare for by studying a largely prescribed curriculum, and diagnostic tests that teachers use to decide how to focus the next several lessons for a given student or group of students.

We use indicator measurements to monitor many aspects of our social and economic life, and these indicators play a distinct role in our political and policy debates. Crime statistics are social indicators; they are used in debates over policing, prison, and juvenile offender policies. Economic indicators, from employment and inflation statistics to aggregate stock market reports, track the performance of the economy as a whole and certain segments of it. The measures used are indirect indicators of the health of the economy. They do not tell us directly about the social effects of unemployment, about how individual people and companies respond to inflation, about who is buying and selling stocks or why. All of these effects and causes must be inferred. Most important, economic indicators do not automatically prescribe action. Rather, individuals making decisions for themselves (e.g., investors) or policy decisions for the country (e.g., the Federal Reserve or Congress) weigh the evidence from multiple sources, consider possible interactions, and decide what action to take: buy or sell, raise or lower interest rates, modify unemployment compensation programs.

The economic indicator system provides information to decision makers but does not function as an actual component of the economic system. The aggregate indicators reflect the independent decisions and actions of thousands, perhaps millions, of individual actors in various parts of the economy. No individual or organization is held responsible for the numbers. It is assumed to be right and proper that individuals and organizations act in their self-interest, using the indexes as information but not as prescriptions for action. As a result, there is little possibility or incentive for "manipulating" the indicators—for faking data or for trying to directly influence the numbers by individual action or by persuasion.

Until now, NAEP has functioned in much the same way: as a system of indicator information that can help decision makers but for which no individual or specific agency can be held directly accountable. Few try to manipulate the NAEP data, for example, by teaching to the NAEP test. The test remains outside the educational process, an indicator of how well the system is doing and a check on the effectiveness of education policies. The recent proposals to use NAEP to hold school districts, individual schools, and even—in the case of the proposed national tests based on NAEP—individual students and teachers accountable for achievement levels would make NAEP *part of* the education system rather than a monitor of it. If this happens, NAEP will be fundamentally changed. Its character and usefulness as an indicator system would be lost. The validity of the NAEP tests as monitors of the general health of the education system would come into question. And, unless the nature of the NAEP tests themselves were changed substantially, their effect on the extent and nature of learning is likely to be negative.

This may sound extreme and even, perhaps, illogical. How could using NAEP tests, which we all agree are technically sound, for accountability negatively affect educational outcomes and the very validity of the NAEP tests? I can most easily explain by invoking an analogy with a very familiar measurement system: thermometers that measure temperature and that are sometimes, via thermostats, hooked up directly to a heating system. A thermometer—the height of a column of mercury, for example—provides an indirect measure of temperature. A direct measure of ambient air temperature would require observing the motion of air molecules: Faster movement means higher temperature. A direct measure of the effects of temperature would require examining people's comfort: Are they shivering or sweating? What kinds of clothing are they wearing?

One reason for using thermometers to measure temperature is their very unobtrusive character; taking a thermometer reading does not change the degree of comfort in a room, does not disturb people or make them unnecessarily conscious of possible discomfort due to temperature. On the other hand, if people do become uncomfortable, they can check the thermometer reading and decide whether to don or shed clothing, add another log to the fire or open the windows. The thermometer provides information but does not directly control actions related to temperature.

Now hook the same thermometer to a thermostat. It becomes part of the heating system. When the temperature drops to a predetermined point, a furnace automatically goes on, heat is supplied, and the temperature rises. When it rises to a

predetermined point, the furnace goes off, and temperature gradually drops. This is a self-regulating system. It requires no human decision making, no intentionality.

In fact, the thermostatic system works precisely because no human intentionality is involved once the system is set up. The air molecules don't care what the temperature reading is. They do not act with the goal of changing the reading. If they did have intentionality, and if they thought the temperature reading was too low, they might rush over to the thermometer bulb or sensor and move about quickly in order to increase the air temperature around the thermometer. The result would be an increase in the temperature reading, and the furnace would go off, making the room colder in a little while!

Educational tests, however, cannot share this characteristic of unobtrusiveness unless special efforts are made to isolate them from the intentional efforts of educators. We cannot place a "test thermometer" in a classroom without expecting to change conditions significantly in the classroom. This is because tests are used in a social rather than a physical system, a system in which measurements that matter in the lives of actors in the system can be expected to affect their future actions. In our temperature analogy, the molecules of air are not motivated to produce a temperature reading in a specific range. But teachers and school principals who are held accountable for test scores are motivated to produce scores in an acceptable range. Any educational assessment for which educators are held personally accountable will produce efforts on their part to have their students perform well on that assessment. Primary among these efforts will be teaching the test item—that is, having students practice doing the very things that will appear on the test. Other things educators might teach will receive less attention or be ignored entirely. Unless the test examines directly the full range of what we want students to know and be able to do, these well-intentioned teaching efforts will reduce what students learn about important parts of the intended curriculum. In other words, highly motivated educators can produce an educationally "colder" room.

In this analogy, NAEP is our educational thermometer, an indicator that tells us at periodic intervals how high or low academic achievement is. In education, we would like to see a continual rise in the achievement level, so we look for ways to stimulate the kind of effort by which educators might drive achievement upward. It is natural enough to think of using NAEP as an incentive for such effort. Advocates of such a shift in the use of NAEP complain that educators ignore national assessment data; they want NAEP to be linked more closely to educational effort

and used as a tool for accountability, for holding school districts, schools, or teachers accountable for levels of educational achievement.

Doing that would be the equivalent of hooking NAEP to the educational "furnace" and using it to turn on more educational effort when the level of achievement is not high enough. In a human and social system, however, we would likely get the room-cooling effect that corresponds to limiting the temperature increase to just the area around the sensor bulb. Today, no educator in the system is particularly motivated to do a speed dance around the NAEP sensor bulb. If they were so motivated, educators would cluster disproportionate amounts of effort around the items on the NAEP test and thus draw attention and effort away from everything else that is important to teach. Unless NAEP were to directly measure the kinds of performances that we want from our education, "real" education achievement temperature would go down even as the indicator went up.

Indirect Measurement: Why NAEP Works as an Indicator

What I have just said does not constitute an indictment or even a criticism of NAEP, as long as NAEP functions as an indicator rather than an accountability measure and no one tries to teach to it. A test used to index or monitor levels of achievement need not measure elements of the curriculum directly. It can use any indirect forms of measurement that can be shown to correlate at high levels with the kinds of academic performances that society values.

Take writing as an example. It is expensive to measure writing competence directly, for students must be given time during the test itself to compose essays, and then a reliable grading system, using trained human judges instead of machines, must be mounted. Suppose, however, we found a way to estimate the logical quality of a piece of writing by counting the number of connectors—such as *thus, therefore, however, nevertheless*—that the essay contained. This might be done by obtaining judges' ratings of the logical quality of student essays and using those ratings as the criterion in a predictive validity study of the connector counting measure. That is, the validity of the connector indicator would rest on its ability to predict logical quality as judged by trained humans. With modern scanning technology, student essays could be put into digital form, and a simple algorithm could be used to count the proportion of logical connectors and to assign a logical quality score to the essay. This would provide an inexpensive indicator of logical quality that would be perfectly valid for use by NAEP as long as no one taught students to pepper their

essays with logical connector terms regardless of the logical connections between their sentences. It does not take great imagination to see that if such teaching occurred, we would soon have student essays filled with logical connector terms, but these same essays might make little logical sense.

In a similar vein, multiple-choice items in which students select among possible problem solutions may be good indirect indicators of how well students could solve problems independently. That is, they may predict independent problem-solving performance adequately. They can, therefore, serve NAEP's monitoring function well. But if NAEP were to become an accountability test, so that there was an incentive to directly teach the NAEP items, it is likely that students would, over time, get better NAEP problem-solving scores but, in fact, not be improving on the real criterion of independent problem solving.

NAEP is filled with indicator measurement. It now includes a short writing sample (and it scores these short essays directly, not by counting logical connectives), but restrictions of testing time and scoring cost limit how much real writing students can do on the test, and so short-answer items are used as surrogates. Most of NAEP's reading items are indirect, multiple-choice measures of reading comprehension that do not assess the breadth of students' reading or the depth of their ability to interpret texts. Measures of mathematical problem solving and scientific concepts are similarly indirect. The vast majority of its items are multiple choice or short answer that do not directly measure the kinds of complex educational performances called for in national, state, and district standards. There is nothing wrong with this form of indirect measurement if NAEP's indirect indicators can be predictively validated against more direct measures of valued educational performances. But, as I argued earlier, it can do enormous harm if people begin to teach directly the kinds of items that are on NAEP and to make that the major focus of teaching and curriculum.

NAEP, the Curriculum, and Standards: A Political and Technical Dilemma

This discussion of indicators versus direct measures brings us to another fundamental question about the future of NAEP: What should its relation be to the school curriculum? Should it lead, pointing educators and the public toward an image of what the schools *ought to be teaching*, or should it reflect actual practice? This question, a dilemma for NAEP since its inception, is intimately tied to the indicator versus direct measure choice. Designed primarily to track how well

students are learning what is actually taught (that is, to follow the curriculum) and to register changes over time, it contains an important core of items that reflect the curriculum of the past. Over the years, additional items have been added to reflect more current curriculum in use. But even this does not satisfy those who would like to see NAEP lead educators toward a more demanding curriculum in which students actually solve problems, write regularly, and come to grips with complex concepts rather than performing routine exercises. If NAEP were to lead the curriculum, it would have to be designed to be taught to. Published standards, along with the test items themselves, would need to signal to educators, parents, and students exactly what they should be teaching and studying. In other words, the NAEP tests would have to avoid indicator items and focus on direct observations of performances. This would require a substantial technical redesign.

It is possible to create assessments designed to lead teaching. The New Standards Project, which I have co-directed with Marc Tucker for the past eight years, has been leading the way in showing how. Working with a partnership of states and school districts, we have developed a system of assessments that are *standards-referenced*. Standards referencing is an extension of the concept of criterion-referenced testing. Like criterion-referenced tests, standards-referenced examinations set an absolute level of performance and report whether a student has met that level. Standards-referenced examinations begin with a very clear definition of the learning standards to be met, including multiple examples of student work that meets the standards. A test is then systematically constructed to assess the extent to which individual students meet those standards. The standards are made available to test users, including teachers, students, and parents. Standards-referenced assessment involves a set of technical processes, including creating assessment tasks to explicitly assess particular standards, assembling a collection of tasks to represent the standard as a whole, and using panels of judges to decide which patterns of responses meet the standard and which do not.

Because standards-referenced examinations are systematically validated against known and publicly available standards, it is possible to prepare students directly for the examinations. This is done by "teaching to" the standards, not to specific test items. Standards-referenced examinations provide a way of ensuring that curriculum, instruction, teacher professional development, and assessment are all aimed at the same clear and public academic expectations.

NAEP's standard-setting exercises do not constitute a standards-referencing process of the kind that I have been describing. The process does not begin with standards and then build the test to the standard, but rather works the other way around. It uses a framework to loosely specify the content that test items should sample but does not decide in advance what kinds of performances should be expected of competent students at given stages of learning. NAEP's *standards*, established only *after* the test is built, are essentially a judgment about cut points on a test scale, about how many correct items justify a *basic*, *proficient*, or *advanced* score on the test. NAEP's approach to standard setting, like its use of indirect measures, is acceptable for an indicator system. It would be a weak and possibly destructive foundation on which to build an accountability system. Using the NAEP standards to guide teaching would require teaching a collection of test items, something that is not a desirable use of the present NAEP tests.

Thus, substantial technical redesign would be called for if an attempt were to be made to use NAEP standards to lead the curriculum. But that is not the greatest problem NAEP would encounter. The real problem is one of fundamental policy. Using NAEP to set standards for the American curriculum would mean that NAEP was setting curriculum for the nation's schools. It would, in no uncertain terms, be specifying what needed to be taught in order to earn the test scores that "counted." This would be true, as well, for tests for individual students that are closely based on NAEP—as is proposed today for the voluntary national tests in reading and mathematics. If NAEP and the National Assessment Governing Board (NAGB) embark on this course, they need to behave as what they will have become: the agency for guiding America's national curriculum. And, as educators begin to prepare students directly for the NAEP tests, they should expect that the value of NAEP as an independent *monitor* of education achievement will decline substantially.