DOCUMENT RESUME

ED 344 939 TM 018 334

AUTHOR Koretz, Daniel
TITLE NAEP and the Movement toward National Testing. Draft.
PUB DATE Apr 92
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Accountability; *Educational Assessment; Educational History; Elementary Secondary Education; *Evaluation Methods; National Competency Tests; National Norms; *National Programs; Political Influences; Student Evaluation; *Testing Programs; Test Use
IDENTIFIERS *Monitoring; *National Assessment of Educational Progress; Trial State Assessment (NAEP)

ABSTRACT

Roles suggested for the National Assessment of Educational Progress (NAEP) in proposed national testing are discussed. Recent proposals center around monitoring, evaluation and accountability, and serving as a benchmark for other tests. Monitoring has been the traditional function of the NAEP, and in this role, its influence has been substantial. The movement to use this assessment for evaluation and accountability is well under way, as exemplified in the Trial State Assessment of 1990 and its extensions. Proposals to use the NAEP would, in effect, use it as a substitute for publisher's norming samples or as a national anchor test. These proposals raise the following technical issues: (1) linking a broad matrix-sampled test to narrower student-level tests; (2) linking taught-to tests to uncorrupted tests; (3) keeping the NAEP itself uncorrupted; and (4) erroneous evaluations based on insufficient data. It is concluded that the NAEP remains best suited for its original monitoring function. There is a 13-item list of references. (SLD)

NAEP and the Movement Toward National Testing

Daniel Koretz
Senior Social Scientist
RAND

In Sharon Johnson-Lewis, Chair, *Educational Assessment: Are the Politicians Winning?* Symposium presented at the annual meeting of the American Educational Research Association, San Francisco, April 22, 1992.

The views expressed here are solely those of the author and do not represent a position of RAND or its clients.

Current proposals for national testing, while in some respects path-breaking, are in other ways a continuation of the reforms of the 1980s. Those reforms left control at the state level, but they were nonetheless national in that a few principal elements--in particular, increases in externally mandated testing, heightened consequences for test scores, and stiffened requirements for graduation from high school--were common to the initiatives in many states. The current debate echoes the 1980s themes of stiffened standards and greater reliance on externally mandated testing, but the focus of the debate has become more clearly national. Even though formal decision-making power still rests with the states, policies are now being formulated at the national level--for example, by the Administration, the National Education Goals Panel, the National Council on Education Standards and Testing, and the Congress.

Central to many of the proposed new reforms is national testing. Some proposals call simply for one or more national tests, while others call for a national system of independent, "voluntary" exams (voluntary for whom, one should ask) that will be linked in some manner via national standards.

When Michael Kean first suggested this symposium, the Administration's proposal called for using the National Assessment of Educational Progress (NAEP) as an interim national examination until a new set of "American Achievement Tests" could be readied. This aspect of the administrations proposal seems to have faded into the background, and none of the other leading proposals for a national examination system would use NAEP in that manner. Nonetheless, NAEP remains a central part of

major proposals, and the roles that are proposed for it highlight some of the major technical and political themes of the current debate.

"WE" VERSUS "THEY?"

Before turning to the NAEP, I would like to digress for a moment about the title of this symposium, *Educational Assessment: Are the Politicians Winning?*

This title might suggest "we" versus "they:" politicians trying to fiddle inappropriately with assessment, while we technically knowledgeable good guys hold the line, or try to, against unrealistic expectations or even outright misuse of tests. Unfortunately, in my experience, the teams have not lined up so tidily. Certainly, some politicians have called for questionable uses of tests and have paid too little heed to expert advice and historical evidence. But so but so have some social scientists. The opposing side--those who have voiced objections to the proposed national testing schemes or urged greater attention to technical concerns or possible undesirable side-effects--also has mixed membership. In my view, many people in the research and testing communities were slow to address publicly the difficult issues raised by the proposed national testing systems and hesitant to criticize them. Indeed, to the extent that matters have slowed down enough for more reasoned consideration of the proposals--and I am not confident that we are yet assured of much of a breather--some of the people who deserve the greatest credit are politicians, most notably Representatives Bill Goodling and Dale Kildee and their staffs. One of the most striking moments in the debates of the National Council on Education Standards and Testing occurred when the Congressional members of the Council distributed a letter insisting that the

Council's call for national testing be held in abeyance until issues of validity, reliability, and fairness were adequately addressed.

THREE ROLES FOR NAEP

Tuning back to the NAEP itself: the functions of NAEP have already changed substantially from its early days, and current proposals would transform them further. To sort out the recent proposals, it is helpful to distinguish three basic roles: (1) monitoring; (2) evaluation and accountability; and (3) serving as a benchmark for other tests.

Monitoring aggregate trends in performance

This is of course the traditional role of NAEP and was until recently its only function: describing what American youth know and can do, and how those proficiencies change over time. Traditionally, NAEP was designed to monitor achievement in the nation as a whole, and its reporting was limited to the national populations of students or youth at specified ages and a very small number of subgroups, such as regions and racial/ethnic groups. Reporting at other political levels, such as the state level, was avoided by design.

How valuable is NAEP as a monitoring tool? Some participants in the current debate disparage the usefulness of merely monitoring achievement, but NAEP's influence as a monitoring tool has been substantial indeed. The current reform movement and that of the past decade arose because of widespread dissatisfaction with the performance of American students. That dissatisfaction rests in large part on a very few indicators of achievement: the NAEP, the SAT (however inappropriately), and a small number of

international studies. Similarly, where do the proponents of current reforms obtain information that the reforms of the 1980s were insufficient and need to be replaced with new initiatives, including national testing? Again, the NAEP. While many state and local tests have shown sizable gains in recent years, the NAEP, the integrity of which was protected because no one taught to it, has shown only very slight improvements (e.g., Linn and Dunbar, 1990).

The importance of maintaining NAEP or some other vehicle for monitoring aggregate performance gradually has become more widely acknowledged over the past year. For example, the final report of the National Council on Education Standards and Testing calls for maintaining a separate system of tests such as NAEP for monitoring national trends. Nonetheless, other uses are still proposed for NAEP.

Evaluation and accountability

The movement to use NAEP not just for monitoring, but also for evaluation and accountability, is already well underway.

The principal step in this direction so far was the initiation of state comparisons using NAEP ("state NAEP" for short). This began with the Trial State Assessment (TSA) of grade-8 mathematics in 1990 and is scheduled by law for gradual expansion. The motivations for state NAEP were undoubtedly diverse, but the desire to judge the quality of state's educational programs and to hold policymakers or educators accountable for differences in performance was prominent among them. For example, the Alexander-

James report, which was perhaps the most influential call for state NAEP, argued:

> Today state and local administrators are encountering rising public demand for thorough information on *the quality of their schools,* allowing comparison with data from other states and districts and with their own historical records. Responding to *calls for greater accountability*, state officials have increasingly turned to the national assessment for assistance (Alexander and James, 1987; emphasis added).

The interpretations given to the first TSA results when they were released last June were varied, but many observers used them to evaluate the quality of school systems. Even though observers suggested a variety of non-educational causes as well, they offered a potpourri of putative educational causes of between-state differences. A report on interpretations by policymakers and leading media noted the following explanations of between-state differences: ability grouping; the shortage of minority teachers; poorly focused curricula; "rigid, centrally controlled schools;" television; parental involvement; the proportion of two-parent families; the proportion of students in large cities; and the proportion of students in poverty (Jaeger, 1992). One state education department of education official immediately announced an intent to scrap the state's general math curriculum.

Several of the most important recent proposals are straightforward extensions of the precedent set by state NAEP. For example, the National Assessment Governing Board (NAGB) recently proposed removing all restrictions on state NAEP and rescinding the statutory ban on reporting of local districts' performance on NAEP. The Administration's now dormant

request to use NAEP as an interim national achievement test would have extended the trend yet further, to the level of individual students.

NAEP as Rosetta stone

A third major proposed use of NAEP is to be the national benchmark that would clarify the meaning of scores on other tests. These proposals would in effect use the NAEP as a substitute for publisher's norming samples or as a national anchor test.

One proposal that gained considerable prominence recently would have created "NAEP-like" tests that could be used to provide scores on individual students, while the real NAEP would be reserved for use at the aggregate level. Immediately before being appointed to his current position as Secretary of Education, for example, Lamar Alexander proposed to establish a center at the University of Tennessee, of which he was then President, that would produce NAEP-like tests that states and districts could use to score individual students. Later, the National Center for Education Statistics began exploring the potential for federal funding of such an effort. The phrase "NAEP-like" could have a variety of disparate meanings, but the intent was that the tests should be similar to NAEP in ways that would facilitate the interpretation of individual's scores in terms of national NAEP results. In effect, NAEP would have served as a substitute for the norms offered with current commercial norm-referenced tests. Both of these specific initiatives seem to be at least dormant.

Other proposals would use NAEP to benchmark states' results on their own assessments. The Kentucky legislature, for example, recently

established a legal requirement that the state's assessments be linked to NAEP. Initially, this will be accomplished by an equating study using TSA results and special supplementary NAEP samples. Other states are considering linkage as well, and NAGB is considering possible changes in policy to facilitate, and perhaps evaluate, linkages of this sort.

A more unusual proposal would use NAEP to set *international* norms for performance. ETS has recently proposed that data from the second International Assessment of Educational Progress be used to estimate the proportion of students in foreign countries who score above each of the *a priori* achievement levels--basic, proficient, and advanced--that NAGB has recently established for the reporting of NAEP results. In theory, that would permit states to compare their performance, not only with the national norms currently provided by national NAEP, but also with norms for a large number of diverse countries.

A FEW TECHNICAL ISSUES

Although these proposed new uses for NAEP may seem relatively straightforward, they actually raise a number of difficult technical issues. I will note four today. None of the three has received sufficient attention in the policy debate about national testing.

Linking a broad, matrix-sampled test to narrower, student-level tests

One of the strengths of NAEP is its breadth of coverage. One factor that makes this possible is its "BIB-spiraled" design--for simplicity, a variant of matrix sampling. Each student is administered only a fraction of the total

items in a domain such as mathematics, but the questions administered to each student nonetheless cover a variety of subdomains, such as measurement and algebra and functions. In addition, in recent assessments, each student has been tested in only one subject area. The total assessment therefore can be broad even if individual testing time is kept short and students are administered time-consuming tasks such as some performance assessments. Information for each student is limited to a single domain, and information about that domain--let alone the subdomains--is quite unreliable. Those disadvantages, however, are relatively unimportant when the assessment is designed, as NAEP is, to provide information only at the aggregate level.

Tests used to obtain scores for individuals, however, must meet a different set of standards, particularly when stakes are high. Unlike NAEP, such tests must include enough items in each domain to provide reliable and valid estimates of each student's performance. If each student is administered only a subset of the test, or if multiple forms are used for other reasons, the forms administered to different students must meet high standards of comparability--a constraint that NAEP does not face. Unlike NAEP, these tests must provide each student with a test of every subject area for which scores are desired.

As a consequence of these considerations, tests used to provide scores for individuals are likely to require much more time to administer than NAEP, but even with more testing time, such tests are likely to be narrower than a good matrix-sampled test such as NAEP. Given that, what does it mean to say that an individual-level test is "NAEP-like?" In what sense is

performance on such a test really comparable to NAEP? If two states use different "NAEP-like" individual tests, each of which is a subset of the NAEP's content, to what degree are they likely to measure the same things?

Linking "taught-to" tests to uncorrupted tests

We know that tests that are used for accountability tend to be taught to in ways that produce inflated scores (e.g., Cannell, 1987; Koretz, Linn, Dunbar, and Shepard, 1991; Linn, Graue, and Sanders, 1990). All other things being equal, narrower tests are likely to be more susceptible than others to inflation. NAEP, on the other hand, remains uncorrupted, at least for the moment; it is protected by its breadth, careful test security procedures, and--until recently--a lack of interest in teaching to it. This may help explain the fact that NAEP has shown relatively flat trends in performance in recent years, while many local and state tests have shown marked if perhaps illusory progress (e.g., Linn and Dunbar, 1990).

How are local and state tests, many of which will be corrupted if they are used for accountability, to be linked to the as yet uncorrupted NAEP? For example, if a state test used for accountability is equated to NAEP in the first year of its use, how will that equating relationship change over time as students are coached on the content of the state test but not on the content of the NAEP? To what extent will improvement on the state's test really indicate progress on the NAEP?

Keeping the NAEP itself uncorrupted

There is growing awareness in the policy community of the value of keeping NAEP itself uncorrupted, lest we lose our only frequently collected,

nationally representative, and uninflated indicator of the achievement of American youth. There is no agreement, however, about the precautions that must be taken. Now that states are ranked (and evaluated) by the media on the basis of NAEP scores, will the breadth of the NAEP and the test security procedures suffice to keep the test uncorrupted? What about the NAGB proposal to permit use of NAEP at the local level? At this point, we have no firm answers, and consequences of inadequate caution could be large indeed.

Erroneous "evaluations" based on insufficient data

The issue raised by using NAEP to evaluate programs is simple: neither NAEP nor most other large-scale assessment programs provide meaningful data on school effectiveness. The rankings of states need not indicate differences in effectiveness and provide no reliable information about the factors that influence achievement (Koretz, 1991).

One consequence is apparent in the interpretations of state NAEP noted earlier: people attribute differences in mean performance to whatever factors they find appealing. This risks substantial harm. Inadequate programs will often look good and will be emulated; effective ones will sometimes be scrapped. Teachers facing unrealistic evaluations based on cross-sectional data will continue to be tempted--as they have for a decade now--to find shortcuts, some illegitimate, for raising scores.

The interpretations of state NAEP noted earlier were based on simple, unadjusted cross-sectional differences among states. There are many reasons why such comparisons cannot tell us anything meaningful about school effectiveness. Educational factors are confounded with non-educational

differences among states; the data provide no information about student growth in achievement; and the data do not indicate anything about students' educational histories (even which states they spent most of their school years in). Data about educational factors that influence achievement are sparse and are partially at the wrong level of aggregation; many of the important factors vary at the level of districts, schools, teachers, or even specific classes. Data were provided only for one grade and would include only three grades if the system were fully implemented, but rankings need not be consistent from grade to grade.

Nor would a more refined use of NAEP data would provide more meaningful information about school effectiveness. For example, why not adjust states' scores to control for confounding differences in demographics and other factors? One could, of course, but the results are unlikely to be meaningful. NAEP, like most large-scale assessment programs, does not provide enough of the needed information. What would be needed is sufficient data to estimate how students in one state would score if they were subjected to the educational policies of another state but remained the same in all other respects, and NAEP does not even approximate that level of detail. It includes, for example, no longitudinal data on educational experiences or outcomes and includes only weak information about non-educational variables that are known to influence achievement. Moreover, even if NAEP could reliably identify states with better educational programs, it cannot tell us which aspects of those programs matter.[1]

---

1 Experiences with the use of adjusted scores at the local and state level have generally been discouraging. Adjusted rankings are often found to be highly inconsistent, varying markedly from year to year and across grade levels, subject areas, and even the statistical methods used to adjust the scores (e.g., Frechtling, 1982; Guskey and Kifer, 1990;

CONCLUSION

The National Assessment of Educational Progress remains best suited for its original purpose: monitoring the performance of America's youth. In that function, unlike other roles that have been proposed for it in recent years, NAEP has no close substitute. It has been valuable as a tool for monitoring achievement in the past, and it can remain so if we maintain it properly. To use NAEP for other roles, however, raises difficult technical issues and, in some cases, could jeopardize NAEP's effectiveness in meeting its primary purpose.

Kippel, 1981; Mandeville and Anderson, 1987; Matthews, Soder, Ramey, and Sanders, 1981; Rowan and Denk, 1983). The causes of these inconsistencies have not been fully explored, but the inadequacy of background data and the lack of longitudinal data are likely to be important factors.

REFERENCES

Alexander, L., and James, H. T. (1987). *The Nation's Report Card: Improving the Assessment of Student Achievement.* (Report of the Study Group). Washington, D. C.: National Academy of Education.

Cannell, J. J., (1987). *Nationally Normed Elementary Achievement Testing in America's Schools: How All the States are Above the National Average,* Friends for Education, Daniels, W. V.

Frechtling, J. A. (1982). Alternative methods for determining effectiveness: convergence and divergence. Paper presented at the annual meeting of the American Educational Research Association, New York

Guskey, T. R., and Kifer, E. W. (1990). Ranking school districts on the basis of statewide test results: Is it meaningful or misleading? *Educational Measurement, 9* (1), 11-16.

Jaeger, R. (1992). *General Issues in Reporting of NAEP Trial State Assessment Results.* Paper prepared for the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project.

Kippel, G. (1981). Identifying exceptional schools. *NewDirections in Program Evaluation*, no. 11 (September), 83-100.

Koretz, D. (1991). State comparisons using NAEP: Large costs, disappointing benefits. *Educational Researcher, 20* (3), 19-21.

Koretz, D. M., Linn, R. L., Dunbar, S. B., and Shepard, L. A. (1991). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. In R. L. Linn (Chair), *Effects of High Stakes Testing on Instruction and Achievement*, symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 5, 1991.

Linn, R. L, and Dunbar, S. B. (1990). The Nation's report card goes home: Good news and bad about trends in achievement. *Phi Delta Kappan, 72* (2), October, 127-133.

Linn, R. L, Graue, M. E., and Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that "everyone is above Average. *Educational Measurement: Issues and Practice, 9* (3), 5-14.

Mandeville, G. K., and Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement, 24* (3), 203-214.

Matthews, T. A., Soder, J. B., Ramey, M. C., and Sanders, G. H. (1981). Use of districtwide test scores to compare the academic effectiveness of schools. Paper presented at the annual meeting of the American Educational Research Association, April.

Rowan, B., and Denk, C. E. (1983). *Modeling the AcademicPerformance of Schools Using Longitudinal Data: An Analysis of School Effectiveness Measures and School and Principal Effects on School-Level Achievement*. San Francisco: Far West Laboratory.