Optimizing State NAEP:

Issues and Possible Improvements

Ina V.S. Mullis
Boston College

Commissioned by the NAEP Validity Studies (NVS) Panel
May 1997

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1791 Arastradero Road
PO Box 1113
Palo Alto, CA 94302
Phone: 415/493-3550
Fax: 415/858-0958

Acknowledgments

The author is especially indebted to Richard Jaeger, as well as to Albert E. Beaton, John Dossey, and Paul LeMahieu for their insightful reviews. Thanks also are given to the remaining members of the NVS Panel for their comments and suggestions, and to Fran Stancavage, in particular, for her support in this endeavor.

Contents

Tables

Abstract

NAEP has conducted the state assessments in 1990, 1992, 1994, and 1996. From the time that Congress authorized NAEP's state assessment program, there has been considerable energy expended to evaluate its quality. Now, after the fourth round of state assessments, it is clear that the program is generally successful. Care should be taken to retain the essential benefits of this important national resources, which permits states to compare to national trends and each other. Participation in state NAEP began at a high level (38 states and 2 jurisdictions) and has increased with 44 states and 3 jurisdictions participating in 1996. According to a survey of state testing directors, NAEP has considerable credibility as a highly valid and reliable source of information.

Despite the high regard for state NAEP, today's environment of limited federal and state resources has led to level funding for NAEP and intense scrutiny about how best to optimize all aspects of NAEP including the state component. Because the state component can account for nearly half the budget devoted to NAEP cooperative agreements, the idea of considering how to reduce effort and maximize utility is a good one for state NAEP. This paper addressed the following topics in relation to making state NAEP more efficient.

- *There is need to examine how to reduce the burden for states.* In sum, the primary way to reduce burden is through less testing. Even though combining the state and national samples or eroding the sample sizes within the states do not appear to be fruitful ideas for now, research needs to continue towards developing more efficient assessment and sampling procedures. However, the major way to significantly reduce burden most likely will remain conducting state assessments on a relatively infrequent schedule and keeping the number of subjects and grades to a reasonable level. The main challenge will be to maximize the information gained from those assessments.

- *There is a need for a stable assessment schedule.* A consistent schedule of regular state assessments would facilitate participation, maximize the use of NAEP as part of a state's own assessment program, provide ongoing trend data to monitor reform efforts, and make it worthwhile for states to link their own assessments to state NAEP. Considering the resource intensive nature of state NAEP, the schedule should be a manageable one, commensurate with the burden currently required by states.

- *The greatest need is to promote the use of state-NAEP data.* This could involve devoting greater attention to how best to link state assessment and NAEP results, developing more timely and user-friendly reports, and working with the states themselves and other organizations to more effectively address the data needs of different NAEP audiences. NAEP should consider developing a state capacity for special reporting. Promoting use will promote the participation and support necessary for the continued success of state NAEP.

Introduction

Since its inception in 1969, the National Assessment of Educational Progress (NAEP) has monitored our nation's educational progress through periodic assessments in a variety of curriculum areas. During the nearly three decades of NAEP assessments, a number of important innovations have been made in the methods used, but the fundamentals have remained essentially the same. In the early 1990s, however, there was a dramatic increase in NAEP's scope when Congress asked it to also begin providing results at the state level. Initially authorized by Congress on a voluntary and trial basis, the goals of the NAEP state assessments were to:

- Allow comparisons between trends in an individual state's performance and national performance.

- Allow states to be compared directly to one another on an independent measure.

Congress authorized trial state assessments, commonly referred to as the TSA, in 1990 and 1992. Since then, state assessments also were conducted in 1994 and 1996 for a total of four rounds of data collection.

The subjects and grades covered in the state assessments since the inception of the program are listed below:

- 1990 - Mathematics, Grade 8

- 1992 - Mathematics, Grades 4 and 8; Reading, Grade 4

- 1994 - Reading, Grade 4

- 1996- Mathematics, Grades 4 and 8; Science, Grade 8

Essentially, the approach has been to capitalize on the development effort required for the national program by making some of the assessments available for administration by states. State assessments in a given subject area, however, dramatically increase the magnitude of the resources and time required for that assessment by virtue of the sheer number of students involved. There also are political considerations as well as many technical issues to address in maintaining high quality comparative data across the participating states and the nation.

The magnitude of state NAEP is illustrated by table 1, showing the numbers of students and schools participating in the national and state components of NAEP's 1996 assessment. It can be seen that the number of participants in state NAEP substantially exceeds those in the national assessment.

Table 1— Approximate Number of Students and Schools in NAEP's 1996 Assessment

	Students	Schools
National*	100,000	1,500
State**	350,000	10,000

* Includes two subjects at three grades plus special studies.
** Includes two subjects at grade 8 and one subject at grade 4. Approximately 2,500 students per grade, per subject, per state and 100 schools per grade, per state.

Favorable Reviews for the NAEP State Assessments

In general, a series of congressionally mandated evaluation studies conducted by the National Academy of Education (NAE) have led to favorable reviews for state NAEP (Bohrnstedt, Glaser, & Linn, 1992, 1993, 1996). Despite the voluntary nature of state NAEP, there has been high interest from the start, with 38 states and 2 jurisdictions participating in 1990. The NAE found the 1990 TSA was carried out successfully and generally with a high degree of validity. The NAE report on the 1992 TSA found strong support for the TSA in the states. A survey of state testing directors conducted for the NAE as part of the NAE's 1994 TSA evaluation effort revealed that NAEP has considerable credibility and is thought to be a highly valid and reliable source of information (DeVito, 1996). NAEP's ambitious frameworks are considered forward looking, and innovative assessment approaches are used. For example, despite considerable controversy over the methods used, the idea of setting performance standards is viewed as an innovation supportive of the goals of the general reform effort. By 1996, participation in the program had even grown somewhat from its initial high level, with 44 states and 3 jurisdictions participating.

It appears that the states use NAEP information for a variety of purposes. Primarily, consistent with the original intentions of Congress, the state data provide an **externally developed reference point** that can be used to:

- Make comparisons to national performance, overall and for subgroups, and
- Make comparisons to other states.

These comparisons provide a general indicator of achievement for state policy makers. The data also provide a basis for arguing for more rigor in curriculum and standards, examining curricular strengths and weaknesses, helping to validate state testing programs, and studying item formats.

Becoming More Efficient

Notwithstanding the favorable reception of the state NAEP program, it remains an enormous undertaking for the Federal government and for the states themselves. Today, in an environment of limited Federal resources, the whole of NAEP, including national and state components, is under close scrutiny by the National Center for Education Statistics (NCES) and the National Assessment Governing Board (NAGB) to reduce costs while maximizing dependability and timeliness (KPMG Peat Marwick LLP, 1996; NAGB, 1996). Even if availability of resources was not a problem, the idea of making cost reductions that do not lower the quality of the program is a good one because it can provide resources to make improvements. By becoming more efficient, state NAEP may be able to do more with the resources that are available.

Understandably, an effort involving as many students and schools as the state NAEP program is very resource intensive. Although costs can vary substantially according to the number of subjects and grades assessed at the state level, an effort involving two subjects at two grade levels could require nearly half of the funds available for NAEP cooperative agreements.

Because of the magnitude of the state assessments, it is important that ongoing efforts to redesign NAEP consider the demands of the state as well as the national component. Such considerations can be complicated by the uneven relationship between the two components, with the state component representing a large proportion of the effort needed, but the national component providing the foundation for NAEP. Even though state NAEP requires a high proportion of the total resources available for NAEP, it uses the materials and procedures developed for the national program. Maintaining a high degree of comparability between the approaches and methods underlying the national and state components has been fundamental to the state program achieving its goals. Therefore, in making changes to national NAEP, any aspects that are to be made available as part of the state program need to be replicable on a very large scale. This need for comparability and large scale replicability goes beyond the instruments and data collection procedures to embrace the full range of assessment activities, including designing comparable samples, obtaining high degrees of participation, endless quality control steps, complex analysis, and a substantial reporting effort.

States also are operating in an environment of reduced resources, and the demands of participation in this voluntary program are substantial. From one perspective, the states have the opportunity to benefit from the Federal investment in NAEP. However, participation in NAEP entails a considerable administrative burden. Participating states must provide sampling information, recruit schools to participate, provide the personnel necessary to implement the data collection activities, and engage in numerous scheduling and record keeping activities.

Since state NAEP is resource intensive for both the Federal government and the participating states, it is important to keep an eye on how state NAEP fits into the redesign effort, identify efficiencies in conducting the state component itself, and most of all think of ways to promote extended use of the information from this important national resource. The remainder of this paper briefly describes a number of areas that could be examined from the view of optimizing NAEP's state component, including sampling, scheduling, content coverage, and data use.

Efficiency and Sample Sizes

To date, regardless of the source of the recommendations, the modifications made to state NAEP since the initial 1990 trial have increased the comprehensiveness and demands on any given subject area assessment at any given grade, primarily by requiring more inclusive samples (e.g., reporting for private as well as public school students) and involving more of them (e.g., district-level NAEP and international links). A number of pending suggestions continue the pressure toward expanding each individual subject area assessment (e.g., reporting for IEP and LEP students and adopting more stringent guidelines for sample participation rates).

More recently, however, attention has been drawn to the burden represented by the extremely large number of students involved in state assessments. Concerted energy has been given to examining how to reduce the sample sizes for those participating states that also are sampled for national NAEP; especially small states. The small states are hit particularly hard in relation to the large states, and the problem is exacerbated when the number of subject areas is increased. Keith Rust (1996), Bruce Spencer (1996), and others have written on these issues. At this time, however, many obstacles remain in terms of deriving national estimates from state samples. It is technically possible to design samples to do this, either by drawing state samples and supplementing them to obtain a national sample or vice versa. However, there are operational considerations that have no immediate solutions. For example, there are differences in administration procedures between the national and state NAEP assessments that have been addressed by equating the two samples. Using only the state administration procedures would jeopardize links to national trend data. Also, because the state assessments are subsets of the national assessments there are differences in content and operational approaches between the two NAEP components. Finally, since participation by states is voluntary, as is participation by the districts within some of the states, last minute withdrawal by states or districts can occur. In fact, state withdrawal can occur subsequent to data collection and analysis, but prior to reporting. To rely on the same samples for both national and state purposes would require eliminating the flexibility provided by the present approach of maintaining separate samples. That is, the same administration procedures, content, and participation rules would need to apply equally to the states and the nation, all of which would involve costs in training and quality control (or for an extremely large professional data collection

staff) not to mention extremely unpopular changes in the NAEP legislation to curtail state's flexibility, and the need to find some way to equalize motivational factors across testing sessions.

Possibly some reductions could be made in state sample sizes, particularly for small states, but this comes at the cost of the quality of the data (Rust, 1996). Another idea for savings involves rethinking the effort expended on collecting data for private schools within the state samples. The biggest savings, however, would come from careful scheduling of the state assessments. Because each one is so costly, the major cost savings come from conducting state assessments on a less frequent schedule and keeping the number of subjects and grades (Forsyth, Hambleton, Linn, Mislevy & Yen, 1996; Rust, 1996).

There are, of course, many trade-offs involved in the state assessment sampling issue. Large sample sizes and frequent assessments are burdensome for NAEP, the states, schools, and most of all the students. However, large sample sizes and high participation rates are necessary to maintain high quality data. More frequent assessments are necessary if policy makers want regular information about trends. States observe that more frequent information about more subject areas is useful for monitoring improvements, particularly in light of the on-going emphasis on educational reform. States also find the disaggregated data for demographic subpopulations very useful, particularly for gender, type of community, and race/ethnicity. This enables policy makers to make judgments about the relative effectiveness of their educational approaches for different subpopulations of students. Interest is growing in having information about even more subpopulations of students (e.g., districts, IEP students, advanced science students) to provide a basis for informed decision-making. Yet, for each targeted subpopulation, there is a high probability that the NAEP samples will need to be increased. Therefore, reductions or increases in sample sizes must be viewed in terms of the information gained or lost as well as in terms of cost and burden.

Assuming that the demand for high quality data about educational achievement will continue to grow, it is important that research continues about the most efficient sampling approaches for state NAEP. For example, different strategies might be used in large versus small states or in states that have state assessments versus those that do not. The research about burden reduction, however, should not be confined solely to issues of sampling methods. For example, exploring ways to improve the precision of proficiency estimates also might contribute to decreasing sample size (e.g., conditioning on state assessment scores at the student level).

As a related idea, NAEP also could explore ways to capitalize on the large sample sizes used for the state assessments. Whether in one subject area or multiple subject areas, frequent or infrequently, when state assessments do occur, it may be worth considering combining the data from the separate samples to obtain improved estimates for both states and the nation. This idea, also proposed by Rust, would have the possibility of optimizing the use of the data collected for state NAEP because the increased sample sizes

would permit more fine-grained analysis at the national and regional levels. Currently, however, it is difficult to work out the issues related to sample weighting and equating on the rigorous schedules required by NAEP.

A Stable and Manageable Schedule of State Assessments

One of the most important issues facing state NAEP appears to be the need for a stable schedule of assessment administrations. Just as NAEP needs stability from the states for efficient procedures, the states need stability from NAEP. For example, based on his survey of state testing directors, DeVito (1996) observed:

> "The most prevalent and deep seated concern of state education agency personnel has to do with the schedule of NAEP, particularly the TSA component. To date, there has been no consistent schedule of TSA content areas of grades."

The lack of a stable schedule for state NAEP influences data quality because it affects participation at both at the state level and for schools within states. First, the states need advance warning in order to budget the necessary resources to participate in the state NAEP program. In order to plan successfully, the states need information about the scope of the offerings in terms of the number of subjects and grades. Second, advance planning greatly facilitates in recruiting schools and maintaining high participation rates.

A consistent schedule would enable states to capitalize on state NAEP in particular ways, optimizing its use in relation to their own state assessment programs. Providing the state and NAEP frameworks coincide, states might use NAEP to augment state funded assessment programs or to monitor specific reform efforts. For example, several states, including Colorado, had planned to use state NAEP as part of their efforts in monitoring progress on the State Systemic Initiative projects funded by the National Science Foundation. However, science was not included in the early rounds of the state program as originally projected. DeVito (1996) summarized the issue:

> "States have learned not to count on particular grades or subjects to help fill a void in the state assessment role. To show leadership on this issue, NAEP staff should develop a basic assessment plan for the future and stick to it. It may be better to have a less ambitious, yet well articulated structure that can be counted on than to have an unclear potential schedule that will likely not be implemented."

Beyond improving the quality of NAEP by facilitating planning and participation, it is likely that a regular schedule would dramatically increase use of state NAEP results. The states themselves and other organizations gradually would come to rely on regularly available data to address public issues of accountability and make policy decisions. This reliance could, in turn, lead to a desire for an increase in the supply of state NAEP data and the availability of the resources to provide it.

Keeping to a schedule of regular, frequent assessments for state NAEP, however, presents quite a challenge during times of fiscal uncertainty at the Federal level. Funding needs to be secure for the development, data collection, and reporting phases necessary to maintain the schedule both within and across the particular assessments included in the cycle. Nevertheless, from a cost-benefit point of view, it is crucial to try and set NAEP on a regular track, including state NAEP.

It is equally important that the schedule envisioned for state NAEP be realistic and manageable in terms of state burden and available resources. In its redesign document, NAGB recommends that reading, writing, mathematics, and science at grades 4 and 8 be given priority for state-level assessments.

Considering the need to curtail the states' burden, it is worth noting that by keeping the special needs of the state assessments in mind while redesigning NAEP, the burden on states can be minimized. For example, even if national NAEP moves to annual assessments, this need not be the case for state assessments. State assessments can remain on an every other year schedule by confining the program to a reasonable number of subject areas.

An example of such a schedule is shown in table 2. Like the currently proposed NAGB schedule, this example has two subject areas assessed every other year at the state level. Also, as in the current NAGB proposal, the state assessments are conducted only at grades 4 and 8, and not all are three grades. Although this schedule primarily tries to keep the state component to a manageable level, it also contains other aspects of the NAEP redesign under discussion. For example, whether it be through existing methods or new methods, there is a "core" portion of each subject area assessment that is a subset or portion of the comprehensive assessment. This subset can be used for measuring trends. The idea is that the core portion could be reassessed and reported with relatively little new or redevelopment effort, and that the analysis and reporting procedures would be already in place. To further reduce state burden, the core portions of assessments also can be made available for use in state NAEP. That is, states not wishing to participate in the comprehensive assessment might participate only in the core portion.

Table 2— Example Assessment Schedule

Year	Comprehensive Assessment†	Core Only for Trends	Core Only for Trends	Focused or Specialist
1	**MATHEMATICS***	**SCIENCE***	Reading	Advanced Mathematics/ Science
2	U.S. and World History	Geography	Reading	Arts
3	**READING***	**WRITING***	Mathematics	—
4	Economics	Civics	Reading	Foreign Languages
5	**SCIENCE***	**MATHEMATICS***	Reading	Advanced Mathematics/ Science
6	Geography	U.S. and World History	Reading	Arts
7	**WRITING***	**READING***	Mathematics	—
8	Civics	Economics	Reading	Foreign Languages

†Includes a core component that can be readily reused to measure trends.
*State NAEP–grades 4 and 8 only; shown in boldface type

The three different kinds of assessments involved in the example schedule are described below:

1. **Comprehensive assessments** are very ambitious in scope, similar to the 1996 mathematics and science assessments. A new framework is developed as the foundation of each comprehensive assessment, and many of the assessment items are newly developed. However, each new comprehensive assessment maintains a link to the previous trend assessment in that subject area. Each comprehensive assessment includes coverage of several subareas and extensive questionnaires. One comprehensive assessment is conducted annually, but for any given subject area the comprehensive assessment is on an eight-year cycle. The one comprehensive assessment conducted each year alternates between being conducted only at the national level and at both the national and state levels. Therefore, a comprehensive assessment is available every other year at the state level. The comprehensive assessments at the state level include mathematics, reading, science, and writing.

2. Each comprehensive assessment has a **core** component for measuring **trends. Trend assessments** are brought forward as **subsets** of comprehensive assessments so that relatively little new development is needed. Trend assessments are streamlined to provide overall indicators of performance for a curriculum area (no subareas) and results for basic demographic groups, including gender, race/ethnicity, socioeconomic level, and private/public schools. School and

teacher questionnaires are abbreviated, if not eliminated entirely. Trend assessments stand on their own in years when a comprehensive assessment is not feasible or desirable.

The core components are incorporated into the next redevelopment of a comprehensive assessment in that subject area to maintain constant monitoring of trends. Biennially, together with the comprehensive assessment, a trend assessment is conducted at the state level. These state trend assessments also are in mathematics, reading, science, and writing, paired with the comprehensive assessments in a way to provide states regular information in these four curriculum areas.

3. The purpose of **specialist or focused assessments** is to measure achievement in subject areas studied by only some students, and where there is no expectation that the entire student population would or should have these skills. To be meaningful, the tests need to be given to special target populations where there is reason to believe that the students would have the special knowledge and skills being assessed. This type of assessment is particularly useful in the arts, foreign languages, and the advanced areas included in international assessments.

   Focused assessments also provide opportunities for a special limited study of a subject, including special topics within subject areas or new assessment methods. Regular use of focused assessments will help ensure that NAEP remains a leader in assessment methods. These types of assessments, however, are not routinely conducted at the state level. They might be made available to states at the states' own cost.

Table 3 shows another perspective of how the example schedule works across assessment cycles. The states receive trend information in each of mathematics, science, reading, and writing on a four-year cycle. They would be able to report trends every other year in either mathematics and science or reading and writing at grades 4 and 8. For one of the two subjects being biennially reported, the results could be based on a comprehensive assessment. This schedule provides for regular flow of information to the states, with a burden commensurate to that currently involved.

Table 3— Alternative View of Example Assessment Schedule to Illustrate State Assessment Cycle

Subjects	Years 1	2	3	4	5	6	7	8
Math	Comp.		Trend		*Trend		Trend	
History		Comp.				Trend		
Reading	Trend	Trend	*Comp.	Trend	Trend	Trend	*Trend	Trend
Economics				Comp.				Trend
Science	*Trend				*Comp			
Geography		Trend				Comp.		
Writing			*Trend				*Comp.	
Civics				Trend				Comp.

*Denotes state assessments given every two years beginning with year one.

The example schedule has the following features designed to optimize both national and state NAEP.

Nationally, three subjects are assessed every year, with the comprehensive assessment receiving the greatest redevelopment effort. A comprehensive development effort is expected in one subject or another on an annual basis.

1. State assessments are conducted every other year, and in only two subject areas. If one or both of the subjects is assessed using a core approach, the state burden remains similar to its current level.

2. Not all subjects need to be assessed at all three grades for either state or national NAEP. In particular, state assessments only are conducted at grades 4 and 8.

3 Should additional funding become available, flexibility exists to expand the assessments and conduct additional special focused assessments.

Comprehensiveness of Content Coverage and Background Questionnaires

From the beginning, comprehensiveness has been a hallmark of each assessment included in the TSA. Broad content coverage has been stressed in the NAEP frameworks and specifications for item development. More recently, the ability to maintain broad content coverage has become more challenging as the types of items specified move increasingly

toward longer and more complicated tasks. These tasks use a disproportionate amount of assessment time requiring either longer testing sessions or larger sample sizes (or both) to maintain content coverage.

The states applaud assessment innovations or the direction of more performance-based approaches and the call for more cutting-edge assessment tasks remains strong. Again, however, in an environment of level funding such as that currently faced by NAEP, there are trade-offs to consider. The more elaborate each single assessment is required to be, the fewer assessments it is possible to conduct. It will take longer to score, analyze, and report the information. Also, the more extensive and detailed the available information, the less likely it will be to find the resources to mine the data in-depth and to effectively disseminate such detailed information to the relevant NAEP audiences.

A balance must be maintained between the more efficient multiple-choice and short-answer questions and the more interesting and content valid-extended tasks. In light of the plans for the performance-based arts assessment and the extraordinary energy given to hands-on and in-depth tasks in the 1996 science assessment, there is concern that NAEP may be placing too much emphasis on the less efficient performance-based tasks. It is important to recognize, however, that it is essential for NAEP to continually improve the content validity of the assessments. From the perspective of subject matter specialists and other NAEP audiences, NAEP's reputation as a high-quality program is highly dependent on prominent use of innovative assessment tasks. Also, the appropriate mix of item types is highly dependent on the subject area being assessed, with performance areas like writing, the arts, and performing scientific investigations requiring performance-based assessment approaches.

There also is a question about the cost effectiveness of the extensive questionnaire information currently being collected by NAEP. Not all NAEP users would reduce the scope of contextual information collected and few recommend eliminating it entirely, but most agree that the information has the potential to be of much greater use to practitioners than it currently is.

Based on the survey of state test directors, DeVito (1996) recommends retaining only the background questions that can be reasonably validated and packaging the results "to encourage insightful conversation that may inform educational reform efforts in the state and the nation, rather than their current reporting mode as appendices to the NAEP reports." In its redesign document, NAGB (1996) notes that the questionnaires asking about teaching practices, teacher preparation, school policies, homework, and television watching—to name a few topics—lead to data analyses that are elaborate, extensive, and complex and reports that are detailed and exhaustive. The Peat Marwick (1996) review found the questionnaires to be well thought out and carefully and clearly worded. However, of two recommendations for technical modifications made in that report, one was "careful delineation and prioritization of the purposes of the NAEP followed by

refinement of the background questionnaires." The NCES plans currently under discussion, include building consensus on a core set of background variables to be collected at various grades and with various subjects (Forgione, 1996).

While the cost savings may not be dramatic (Peat Marwick, 1996), honing the background questionnaires—at least for the state assessments—would dramatically impact the burden on principals and teachers. The scope of the person effort expended on questionnaires in the state assessments is enormous, if one considers that nearly 10,000 principals and as many as 150,000 teachers (approximately 3 to 5 teachers per school per grade) could be spending approximately 20 minutes apiece completing these questionnaires. Empirical studies of the NAEP data and of users of NAEP data should be conducted to determine which contextual variables are most useful. For example, an analysis could be done of recent NAEP reports and secondary analyses of NAEP data to see which variables have been used to date. The survey planned by NCES to collect information from states and other constituents about NAEP implementation, issues, and options also will provide valuable insight into which of the currently reported variables are the most useful to educational decision makers, and which variables they would most like to see included in NAEP analyses. Even with judicious pruning, reporting its extensive background questionnaire data will continue to present a particular challenge for state NAEP. There is a general sense that more analysis could be done with the data and that this information has the potential to be much more useful to practitioners than it currently is.

NAEP as the Norm: Linking State Assessments to NAEP

From a more traditional testing perspective, in a schedule similar to the example given, the NAEP assessments administered every other year at the state level could provide excellent norming samples. In a time of strict fiscal accountability, it may be appropriate for NAEP to give concerted thought on how to best capitalize on this situation. A direct approach, however, is ill-advised. In the direct application of the norm-referenced testing approach, the NAEP trend assessments in all four subject areas would be made available for ongoing administration in intervening years at state option and cost so that states could monitor trends in mathematics, science, reading, and writing on an annual basis if they so desired. Unfortunately, if this approach was successful, it would substantially erode the integrity of NAEP. Security would be a major problem and states using the same NAEP trend assessments year after year undoubtedly would become susceptible to the "Lake Woebegon" effect plaguing commercial test publishers. Particularly in medium- to high-stakes testing environments, lack of security and direct teaching to the NAEP test would lead to a situation in which all states eventually would be performing above the nation. Essentially, for all intents and purposes, NAEP would be ruined.

As an alternative with the same benefits to states and little risk to NAEP, states could link their own state assessments to state NAEP by giving their own assessments to the same students participating in state NAEP. The notion of linking has been raised since the inception of state NAEP, but the technical challenges have yet to be completely overcome. One difficulty is obtaining the two sets of scores for the same students. Another major difficulty arises in trying to link tests with differences in content, item format, and motivational levels. The quality of the results based on the linking is highly dependent on a high degree of congruence between the two measures. However, if individual state assessments (or even parts of them) were more closely aligned with state NAEP, then these difficulties might be reduced.

As research in the area of linking becomes more widespread, it is entirely possible that even more methodological challenges will emerge. NCES presently is conducting research with four states to study the methodology required to link their individual state assessments to NAEP. Also, even after results are obtained, they need to be interpreted with care and monitored across time. For example, the linkings might not hold up over time if state assessments are closely tied to a state curriculum different from that assessed by NAEP and there is considerable teaching to the test.

Despite the many hurdles, research in this area should continue. If such linking could be accomplished successfully, states then could re-administer their own assessments to monitor trends in intervening years and have the additional capability of comparing their results to NAEP. Availability of results would not depend on NAEP, but on the states themselves, increasing the likelihood of fast turn-around time. Using the example schedule as an illustration, the links between individual state assessments and state NAEP could be updated every four years.

The point made in the Design/Feasibility Team Report to NAGB (1996) is well taken, that "NAEP should not be in the business of policing and certifying linkages between NAEP and other assessments. The best way to support these efforts would be to provide clear discussions and outlines of procedures for valid linking approaches, and examples to use as models." However, the states would seem different than external audiences because they are integral to the NAEP effort. If, as is the case for some states, individual state assessments and state NAEP provide conflicting results for state audiences, why is this? Do the differences relate to the content of the tests, the formats used, or the samples? Taking into account the concentrated efforts toward educational reform taking place in a number of states, it probably would behoove both NAEP and the states to learn as much as possible from each other. Tightening the coherence between NAEP and the states will be a challenge, but in the long run it would provide increased credibility and utility for NAEP. Even though the links between individual state assessments and state NAEP would never be perfect, the benefits may outweigh the concerns.

Need to Promote the Use of State Data

The utility issue is crucial to the continued success of state NAEP. Promoting use promotes participation and this, in turn, increases the likelihood of continued support for the program. Providing a basis for linking state assessment results is only one way to promote use of the state NAEP data. NAEP needs a multi-faceted approach to encourage widespread and correct usage of its data, while minimizing erroneous conclusions. Another way to help reach this goal is by providing timely, informative reports and other useful materials such as the frameworks and items. But, the concept needs to be enlarged by working more closely with users and staging more mediated encounters with the NAEP data, either through technology or structured events.

Two major issues center on the NAEP reports of state assessment results. The first is the length of time taken after data collection to produce the reports and the second is the overall utility of those reports. Apparently most state directors felt that six months or less after data collection should be the goal for reporting results (DeVito, 1996). Given that the assessments are conducted in February and March, the state testing directors felt an effort should be made to release the results in September/October or at least prior to the end of the calendar year in which the assessment was conducted. NAGB also supports releasing NAEP results within six months of the completion of testing. Whether this goal is feasible remains to be seen, since the most time consuming part of the process is the NCES-NAGB review/revision stage—as much as one year out of a two-year process (Peat Marwick, 1996). The Peat Marwick report (1996) also raised serious concerns about the efficiency of two-tier Report Cards, stating that the costs "seem to outweigh the advantages."

The second suggestion for increasing the use of state NAEP reports involves making them more user-friendly. The DeVito (1996) survey found a preference for reports and materials prepared specifically for use by classroom teachers, principals, superintendents, and local school boards stating that: "Less detailed, targeted pieces should be produced for different audiences to increase the usefulness of the information."

To help educational policy makers understand the utility of the state NAEP data, it appears that NAEP needs to take better account of the different constituencies that have different needs for NAEP data—national legislators, for example, or local school boards and the general public. NAEP has a major responsibility for providing better information to the public about state-by-state comparisons, and needs to continue its progress in working towards more timely and user-friendly approaches to meet this obligation.

It does, however, seem unrealistic for the states to expect the Federal government to assume responsibility for creating articles and pieces pertinent to a variety of audiences within each state. Perhaps it is time to consider more shared reporting responsibilities between NAEP and the states participating in state NAEP. The federal role might be one

of providing initial training and staff development. Encouraging increased dissemination and use of NAEP results within states, would benefit both the states and NAEP, generally. Some states already prepare materials including state NAEP results.

On a pilot basis, NAEP might consider working with several states to produce a series of short publications entitled "Conversations with the States." One goal would be to produce user-friendly pieces suitable for dissemination to the public nationally and within each state. Another goal would be to develop some pieces targeted toward particular audiences, for example, teachers or school boards. Thus, the "conversation" topics might vary, with some cutting across states and other having more relevance within a particular state. Similarly, some might be for specialized audiences and others for the general public.

Beginning with perhaps three states, NAEP could work with policy makers and practitioners within each of the individual states participating in the pilot. NAEP, perhaps in conjunction with the Council of Chief State School Officers (CCSSO), could help state education agency personnel use the NAEP data to develop publications for particular use within the context of that state. The pilot states would be responsible for providing individuals within their state to participate in the project and for publishing the materials developed for their own individual state. These materials would be for the state's own use and not be subject to NCES review; the latter involves a lengthy process and would slow down the publication schedule significantly. With some planning, the "within state" materials could be ready for simultaneous release with the initial state NAEP reports.

The pieces developed in working separately with the three pilot states could then become models for use in other states. To facilitate this idea, NAEP might even consider using these "model" pieces to conduct a workshop for states on how to develop shorter targeted pieces for within state use. Based on the work with the pilot states, NAEP would develop a heightened sense of how its data can be better used at the state and local level. NAEP could highlight specific uses of state NAEP within the pilot states in developing brochures and pamphlets for dissemination to both general and targeted audiences across the nation. Such concrete examples of the benefits of state NAEP data would illustrate the importance of this extraordinary program.

Besides collaborating with individual states, there are other organizations with which NAEP could work to help promote the use of state NAEP results. Groups that would likely be interested in working to improve the utility of state NAEP data include the Council of Chief State School Officers, the Council of Greater City Schools, and the National Governors Association. Technology could be used to hold teleconferences sponsored by these organizations as well as to provide policy briefs electronically and engage in electronic conversations about them. One or more of these groups, might develop a consultation service on the use of state NAEP data in state-level decision making about education.

Teacher's organizations also might be interested. For example, working with the National Council of Teachers of Mathematics (NCTM) to create packets for schools that provide individualized information on their performance on specific released items might be useful and might not violate confidentiality. Naturally, the degree of precision associated with such school estimates would need to be explored, but the data might serve as a springboard for developing staff training and information materials. The NCTM could provide advice on which items to feature and provide commentary for teachers about why achievement on the items was important to success in mathematics. The NCTM could also provide information about how to improve performance in the areas represented by the items, if performance was lower than desired by the district or school.

Final Thoughts

Even though the widespread participation in state NAEP attests to the high regard for the program, greater attention to dependability and coherence could substantially increase its utility. Although some refinements in procedures for individual assessments may be in order according to recommendations included here and elsewhere, the primary theme seems to be a greater need to carry through, specifically, in the areas of schedule and dissemination. Everything considered, state NAEP may require a disproportionate amount of resources for the payoff received.

Certainly the quality and integrity of NAEP cannot be jeopardized, and it must continue in the forefront of innovative assessment approaches. Without this foundation of excellence, decision makers and practitioners simply will not use the results. Since, however, state NAEP is receiving generally high marks for credibility, the emphasis in improving this program needs to be on stepping back and looking at broad-based issues. The intent should be to maintain high quality, while trying to increase utility and keeping a strict eye on feasibility. That is, how can state NAEP make the most of its resources? How can it be optimized?

Promoting more use of the state NAEP data would appear vital to the continued success of the program. Promoting use equals promoting participation equals promoting support. At least in the short term, it is worth examining the idea of expending proportionately less of the state NAEP resources on data collection and proportionately more on disseminating information about the many uses of the program.

References

Bohrnstedt, G., Glaser, R. and Linn, R. (1993) *The trial state assessment: Prospects and realities*. The third report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: 1992 trial state assessment. Stanford, California: National Academy of Education.

Bohrnstedt, G., Glaser, R. and Linn, R. (1992) Assessing student achievement in the states. The first report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: 1990 trial state assessment. Stanford, California: National Academy of Education.

DeVito, P.J. (1996) Issues relating to the future of national assessment of education progress: The state perspective. Prepared for the annual meeting of the American Educational Research Association, New York, April.

Forgione, P. (1996). Draft Statement of the Commissioner. NCES Plans for Implementing NAEP Redesign and the Future of NAEP.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., and Yen, W. (1996). Design/Feasibility Team Report to the National Assessment Governing Board. Washington, DC: National Assessment Governing Board.

KPMG Peat-Marwick LLP and Mathtech, Inc. (1996). A review of the National Assessment of Educational Progress: Management and methodological procedures. Study conducted for the U.S. Department of education, National Center for Education Statistics. Washington, DC: Author.

National Assessment Governing Board (1996). Policy Statement on Redesigning the National Assessment of Educational Progress. Washington, DC: Author.

Rust, K. (1996). "Sampling Issues for Redesign" memorandum to Mary Lynn Bourque, NAGB, May 9, 1996.

Spencer, B. (1996). Combining state and national NAEP. A paper prepared for the evaluation of state NAEP conducted by the National Academy of Education.