

**Adopted: August 2, 1996**



## **National Assessment Governing Board**

### **Redesigning the National Assessment of Educational Progress**

#### **Policy Statement**

##### *Foreword*

*This policy statement was adopted in 1996, at a time when Congress had codified National Education Goals targeted for accomplishment by the year 2000. It was the expectation that the National Assessment of Educational Progress (NAEP) would be a primary means for monitoring progress in achieving the goal addressing student achievement and this expectation is reflected in the policy below. The National Education Goals legislation is no longer in effect and has been superseded by other national policies, the No Child Left Behind Act of 2001 (NCLB) being the most germane. Therefore, the references to National Education Goals in this policy statement are no longer relevant.*

*Under NCLB, state level participation in assessments in reading and mathematics in grades 4 and 8 became mandatory. Participation is required on a biennial basis, affecting costs and technical design. However, the overall intent and impact of the policy—to clarify purpose, define the audience, set forth limitations, maintain quality and integrity, and bring efficiencies to the design of the assessment—remain in effect and continue to guide the policy setting and operations of the NAEP. (Foreword added August 2007.)*

#### **A Better Way to Measure Educational Progress in America**

An effective democracy and a strong economy require well-educated citizens. A good education lays a foundation for getting a good job, leading a fulfilling life, and participating constructively in society.

But is the education provided in your state and in America good enough? How do our 12th graders compare with students in other nations in mathematics and science? Do our eighth grade students have an adequate understanding of the workings of our constitutional democracy? How well do our fourth grade students read, write, and compute? NAEP is the

only way for the public to know with accuracy how American students are achieving nationally and state-by-state.

The NAEP tests at grades 4, 8, and 12. By law, it covers 10 subjects, including reading, writing, mathematics, and science. NAEP has performance standards that indicate whether student achievement is “good enough.” NAEP is not a national exam taken by all students. In fact, only several thousand students are tested per grade, comprising carefully drawn samples that represent the nation and the participating states. Since its first test in 1969, the NAEP has earned a trusted reputation for its quality and credibility. That reputation must be maintained.

NAEP is unique because of its national, state-by-state, and 12th grade results. State and local test results cannot be used to provide a national picture of student achievement. States and local schools use different tests that vary in many ways. The results cannot simply be “added up” to get a national score nor can state scores on their different tests be compared. The National Assessment Governing Board believes that 12th grade achievement is important to monitor at the national level, because the 12th grade marks the end of elementary and secondary education, the transition point for most students from school to work, to college, or to technical training. NAEP is the only source of nationally representative data at the 12th grade. College entrance tests such as the ACT and the SAT are taken only by students planning on higher education; the results do not represent the achievement of the total 12th grade class. And to date, virtually no state-based assessment program tests 12th graders.

While there is much about NAEP that is working well, there is a problem. Under its current design, NAEP tests too few subjects, too infrequently, and reports achievement results too late—as much as 18 to 24 months after testing. Testing occurs every other year. During the 1990s, only reading and mathematics will be tested more than once using up-to-date tests and performance standards. Six subjects will be tested only once and two subjects not at all during the 1990s.

Why is NAEP testing so few subjects and fewer subjects now than years ago? Over the years, NAEP has become increasingly complex. Its quality and integrity have led to a multitude of demands and expectations beyond its central purpose. Meeting those expectations was done with good intentions and seemed right for the situation at the time. However, additions to NAEP have been “tacked on” without changing the basic design, driving up costs and reducing the number of subjects that can be tested.

For example, where a single 120-page mathematics report once sufficed, mathematics reporting in 1992 consisted of seven volumes totaling almost 1,800 pages, not including individual state reports. Also, there are now two separate testing programs for reading, writing, math, and science. One monitors trends using tests developed during the 1970s; the other reflects current views on instruction and uses performance standards to report whether achievement is good enough.

The current NAEP design is overburdened, inefficient, and redundant. It is unable to provide the frequent, timely reports on student achievement the American public needs. The challenge is to supply more information, more quickly, with the funding available.

To meet this challenge, the NAEP design must be changed, building on its strengths while making it more efficient. The design of NAEP must be simplified. The purpose of the NAEP must be sharply focused and its principal audience clearly defined. Because NAEP cannot do all that some would have it do, trade-offs must be made among desirable activities. Useful but less important activities may have to be reduced, eliminated, or carried out by others. NAEP must “stick to its knitting” in order to be more cost-effective, reach more of the public, provide more information more promptly, and maintain its integrity.

## **NAEP Redesign**

To provide the American public with more frequent information in more subjects about the progress of student achievement, changes must be made in the way that NAEP is designed and the results are reported. These changes are described in this policy statement. Undergirding these changes is an explicit statement of the purposes, objectives, audiences, and limitations of NAEP.

While change is in order, many current policies should continue. For example, reliability, validity, and quality of data will remain hallmarks of NAEP. The sample of tested students will be as representative as possible, using policies and procedures that maximize the number of students included who are disabled or are of limited English proficiency. And reporting on trends over time will remain a central commitment of NAEP.

The intent of this policy statement is to guide current operations of NAEP, the development of new requests for proposals for contracts for conducting NAEP, and the activities and structure of the Governing Board. Contracts for current operations extend through assessments to be conducted in 1998. New contracts would cover assessments as early as 1999 and thereafter.

## **Purpose and Objectives of NAEP**

The purpose of NAEP is stated in its legislation:

**“...to provide a fair and accurate presentation of educational achievement in reading, writing, and the other subjects included in the third National Education Goal, regarding student achievement and citizenship.”**

Thus, the central concern of NAEP is to inform the nation on the status of student achievement. The Governing Board believes that this should be accomplished through the following objectives:

**1. To measure national and state progress toward the third National Education Goal and provide timely, fair, and accurate data about student achievement at the national level, among the states, and in comparison with other nations;**

**2. To develop, through a broadly inclusive process, sound assessments to measure what students know and can do as well as what students should know and be able to do; and**

**3. To help states and others link their assessments with NAEP and use NAEP data to improve education performance.**

The specific changes in the design of NAEP described below are discussed in relation to these objectives.

## **The Audience for NAEP**

The primary audience for NAEP results is the American public, including the general public in states that receive their own results from NAEP. Reports should be written for this audience. Results should be released within 6 months of testing. Reports should be understandable, free of jargon, easy to use, and widely disseminated. Although more comprehensible, direct, and useful, the reports will not trade accuracy for simplicity. The tradition of high quality of NAEP reports will be continued, with no erosion of validity and reliability. NAEP questions and samples of student work that illustrate performance standards are likely to receive heightened prominence in reports.

Principal users of NAEP data are national and state policymakers and educators concerned with student achievement, curricula, testing, and standards. NAEP data will be available to these users in forms that support their efforts to interpret results to the public, to improve education performance, and to perform secondary analysis.

## **Limitations: What NAEP Is Not**

NAEP is intended to describe how well students are performing, but not to explain why. NAEP only provides group results; it is not an individual student test. NAEP tests academic subjects and does not collect information on individual students' personal values or attitudes. Each NAEP test is developed through a national consensus process. This national consensus process takes into account education practices, the results of education research, and changes in the curricula. However, NAEP is independent of any particular curriculum and does not promote specific ideas, ideologies, or teaching techniques. Nor is NAEP an appropriate means, by itself, for improving instruction in individual classrooms, evaluating the effects of specific teaching practices, or determining whether particular approaches to curricula are working.

**OBJECTIVE 1: To measure national and state progress toward the third National Education Goal and provide timely, fair, and accurate data about student achievement at the national level, among the states, and in comparison with other nations.**

## **Assess all subjects specified by Congress: reading, writing, mathematics, science, history, geography, civics, the arts, foreign language, and economics**

The gap must be closed between the number of subjects NAEP is required to assess and the number of subjects it can assess at the national level under the current design. By law, NAEP is required to assess 10 subjects and report results and trends. In order to chart progress and report trends, subjects must be assessed more than once. However, during the 1990s only reading and mathematics will have been assessed more than once using up-to-date tests and performance standards to report how well students are doing.

Some have suggested that a solution is to combine into a single assessment several related subjects (e.g., reading and writing and/or history, geography, civics, and economics). Under such an approach, assessment data would be reported using both an overall score and sub scores for the respective disciplines. Although such an approach has the appeal of reducing the number of separate assessments, its feasibility, desirability, and costs are unknown. Also, such an approach has far-reaching implications for the test frameworks that guide the development of each assessment and for reporting results. These implications must be considered carefully. For the immediate future, subjects will continue to be assessed separately. However, the Governing Board is committed to providing the public with more information as efficiently as possible. The Governing Board will consult with technical experts and education policymakers, in conjunction with the development of assessment frameworks, to determine the feasibility, desirability, and costs of combining several related subjects into a single assessment.

- **NAEP shall be conducted annually, two or three subjects per year, in order to cover all required subjects at least twice a decade.**
- **NAEP shall assess all subjects listed in the third National Educational Goal—reading, writing, mathematics, science, history, geography, civics, the arts, foreign language, and economics—according to a publicly released schedule adopted by the Governing Board, covering 8 to 10 years, with reading, writing, mathematics, and science tested more frequently than the other subjects.**
- **The Governing Board shall consult with technical experts and with education policymakers, in conjunction with the development of assessment frameworks, to determine the feasibility, desirability, and costs of combining several related subjects into a single assessment.**

## **Provide NAEP results for states**

In 1988, testing at the state level was added to NAEP as a trial, with participation strictly voluntary, subjects and grades specified in law, and an independent evaluation required. Previously, NAEP had reported only national and regional results. For the first time, the information was relevant to individuals in states who make decisions about

education funding, governance, and policy. As a result, states now are major users of NAEP data.

Participation was strong in the first state-level assessment in 1990 and has grown to include even more states. In 1996, 44 states and 3 jurisdictions participated in the mathematics assessments at grades 4 and 8 and the science assessment at grade 8. The independent evaluation concluded that the trial state assessments produced valid and reliable data. The evaluation report recommended, and Congress agreed, that state-level assessments, with continued evaluations, be included in the 1994 reauthorization of the NAEP.

Currently, NAEP draws a separate sample to obtain national results in addition to the samples drawn for individual state reports. Keeping the schools drawn for national samples completely partitioned from the state samples increases costs and creates additional burdens on states, particularly small states. Options should be identified for making the national and state samples more efficient and less burdensome. For example, it may be possible to reduce the current state sample size of 100 schools to a smaller number (e.g., 65–75) without a great loss in precision.

States participate in NAEP for many reasons, including to have an unbiased, external benchmark to help them make judgments about their own tests and standards. NAEP data are used to make comparisons to other states, to help determine if curriculum and standards are rigorous enough, to develop questions about curricular strengths and weaknesses, to make state-to-international comparisons, and to provide a general indicator of achievement.

There is a strong interest among states to participate in NAEP to get state-level information at grades 4 and 8 in reading, writing, mathematics, and science. The level of interest in participating in NAEP varies with respect to the other subjects (i.e., history, geography, civics, economics, the arts, and foreign language) and at grade 12, where state officials say that obtaining cooperation from high schools and 12th grade students is difficult.

Some states, however, would like to be able to use NAEP tests in the other subjects and at grade 12. Such use of NAEP tests would be conducted as a service, with the reporting of results and maintenance of data under the control of the state. States will be able to use NAEP tests if they adhere to requirements to protect the integrity of the NAEP program and pay the additional costs. At the present time, states that participate in the NAEP to get state level information at grades 4 and 8 in reading, writing, mathematics, and science provide in-kind support to cover the cost of in-state coordination and test administration. The NAEP program covers the majority of costs, including test development, sampling, analysis, and reporting. States that wish to use NAEP tests in other subjects and at grade 12 would pay for much of these additional costs.

States are active partners in the NAEP program. States help develop NAEP test frameworks, review test items, and assist in conducting the tests. The NAEP program is effective, to a great degree, because of the involvement of the states.

Because it is useful to them, and because they invest time and resources in it, states want a dependable schedule for NAEP testing. With a dependable schedule, states that want

to will be better able to coordinate NAEP with their own state testing program and make better use of NAEP as an external reference point.

- **NAEP state-level assessments shall be conducted on a reliable, predictable schedule according to an 8- to 10-year plan adopted by the Governing Board.**
- **Reading, writing, mathematics, and science at grades 4 and 8 shall be given priority for NAEP state-level assessments.**
- **States shall have the option to use NAEP tests in other subjects and at grade 12 by assuming a larger share of the costs and adhering to requirements that protect the integrity of the NAEP program. However, the Governing Board shall seek ways to make such use of NAEP tests attractive and financially feasible.**
- **Where possible, changes in national and state sampling procedures shall be made that will reduce burden on states, increase efficiency, and save costs.**

## **Vary the amount of detail in testing and in reporting results**

More subjects can be assessed if different strategies are used. Currently, each time NAEP is conducted, it uses a similar approach, regardless of the nature of the subject or the number of times an assessment in a subject has been administered. This approach is locked in through 1998 under current contracts. Under this approach, a larger number of students is tested in order to provide not just overall results, but fine-grained details as well (e.g., the achievement scores of fourth-grade students whose teachers that year had five hours or more of in-service training). NAEP also collects “background” information through questionnaires completed by students, teachers, and principals. The questionnaires ask about teaching practices, school policies, and television watching, to name a few. Data analyses are elaborate. Reports are detailed and exhaustive, involving as many as seven separate reports per subject. Although NAEP has been praised for this thoroughness, the cost of this thoroughness is that fewer subjects are assessed, assessments occur less frequently, and reports take longer to produce.

The different strategies needed might include several approaches to testing and reporting, all of which should be designed in ways that maintain NAEP’s commitment to providing valid and reliable data of high quality. For example, these approaches could take the form of “standard report cards,” “comprehensive reports,” and special, focused assessments.

A standard report card would provide overall results in a subject with performance standards and average scores. Results for standard report cards could be reported by sex, race/ethnicity, socio-economic status, and for public and private schools, but would not be broken down further. This may reduce the number of students needed for testing and may reduce associated costs. Generally, subcategories within a subject (e.g., algebra, measurement, and geometry within mathematics) would not be reported. However, data from NAEP would continue to be available to state and local educators and policymakers for additional analysis.

Comprehensive reports, like the current approach, would be an in-depth look at a subject, perhaps using a newly adopted test framework, many students, many test questions, and ample background information. In addition to overall results using performance standards and average scores, subcategories within a subject could be reported. Results would be reported by sex, race/ethnicity, socio-economic status, and for public and private schools, and might be broken down further as well. In some cases, more than one report may be issued in a subject. Comprehensive reporting in a particular subject would occur infrequently, perhaps once in 10 years, but under a planned schedule of assessments.

Special, focused assessments on timely topics also would be conducted. They would explore a particular question or issue and may be limited to particular grades. Generally, the cost would be less than the cost of a standard report card. Examples of these smaller-scale, focused assessments include: (1) assessing subjects using targeted approaches (e.g., 8th grade arts), (2) testing special populations (e.g., in-school 12th graders versus out-of-school youth), and (3) examining skills and knowledge across several subjects (e.g., readiness for work).

The use of background surveys also would be varied. The three kinds of background surveys—student, teacher and principal questionnaires—would not necessarily all be employed each time a subject is assessed. Instead, the use of such surveys would be limited and selective, with reports of results focused on a core of background questions addressing the most essential issues. Also, background surveys used for standard report cards in a particular year would be designed to complement, rather than duplicate, background surveys used for comprehensive reports in the same year.

- **NAEP testing and reporting shall vary, using standard report cards most frequently, comprehensive reporting in selected subjects about once every ten years, and special, focused assessments.**
- **NAEP results shall be timely, with the goal being to release results within 6 months of the completion of testing for standard report cards and within 9 months for comprehensive reports.**

## **Simplify the NAEP design**

The current design of NAEP is very complex and, in fact, has grown more complex over the years. Here are just three examples of this complexity. (1) No student takes the complete set of test questions in a subject and as many as 26 different test booklets are used within each grade. Scores are calculated using sophisticated statistical procedures. (2) Students, teachers, and principals complete separate background questionnaires and may submit them for scoring at different times. Data from the questionnaires are used in calculating results of the assessments. (3) Current requirements for data analysis demand that test scores be calculated for every background variable collected by the NAEP before any report can be produced. This lengthens the time from data collection to reporting and adds significantly to cost.



The design became more complex, in part, because NAEP's purposes and audiences had proliferated and the amount of background information collected had expanded. Specifying the purposes, audiences, and limitations of the NAEP, as well as providing for varied means for testing and reporting, will result in opportunities for simplifying the NAEP design.

- **Options shall be identified to simplify the design of NAEP.**

## **Simplify the way NAEP reports trends in student achievement**

From its beginning in 1969, monitoring achievement trends has been a central mission of NAEP. Monitoring long-term trends in educational achievement, both for the population as a whole and for significant subgroups, is a capacity unique to the NAEP and should be continued as a central mission. However, as NAEP approaches its third decade, it must address the problem of how to assess trends in achievement when curricula continue to evolve and change. An assessment in a subject must be kept stable to monitor trends. However, stable assessments may not reflect important changes in curricula. Over time, there develops a legitimate concern about the relevance of the content of the assessment versus the ability to track change in achievement.

As a solution to this problem, since 1990, NAEP has reported achievement trends using two unconnected assessment programs. The tests, criteria for selecting students, and reporting are all different. The first program, “the main NAEP,” tests at grades 4, 8, and 12 and covers ten subjects. The assessments are based on a national consensus representing current views of each subject. Performance standards are used to report whether student achievement on NAEP is “good enough.” The schedule of subjects to be assessed in the main NAEP is unrelated to the schedule of subjects under the second testing program.

The second assessment program reports long-term trends that go as far back as 1970. Only four subjects are covered: reading, writing, mathematics, and science. The assessments are based on views of the curricula prevalent during the 1970s and have not been changed. Testing is at ages 9, 13, and 17 except for writing, which tests at grades 4, 8, and 11. Trends are reported by average score; performance standards are not used. The long-term trend program has been valuable for documenting declines and increases in student achievement over time and a decrease in the achievement gap between minority and nonminority students.

It may be impractical and unnecessary to operate two separate assessment programs. However, it also is likely that curricula will continue to change and that current test frameworks may be less relevant in the future. The tension between the need for stable measures of student achievement and changing curricula should be recognized as a continuing policy matter for NAEP, requiring efficient and balanced design solutions. Among the factors to consider are: (1) setting a standard period of time for a long-term trend (e.g., 15–20 years) using a particular “metric” in a subject; (2) providing for overlapping administrations of old and new assessments and “bridge” studies to determine whether the new can be linked to the old assessment; and (3) periodic administration of older

assessments (e.g., once every 10 years once a new trend line has been established so that it would be possible to compare performance in 2010 with that in 1970 on the old trend line and with that in 1990 on a new trend line).

- **A carefully planned transition shall be developed to enable “the main NAEP,” to become the primary way to measure trends in reading, writing, mathematics, and science in the NAEP program.**

## **Use performance standards to report whether student achievement is “good enough”**

In reporting on “educational progress,” the NAEP has, until recently, only considered current student performance compared to student achievement in previous years. Under this approach, the only standard was how well students had done previously, not how well they should be doing on what is measured by NAEP. Although this approach has been useful, it began to change in 1988 from a sole focus on “where we have been” to include “where we want to be” as well.

In 1988, Congress created a nonpartisan citizen’s group—the Governing Board—and authorized it to set explicit performance standards, called achievement levels, for reporting NAEP results.

The achievement levels describe “how good is good enough” on the various tests that make up NAEP. Previously, it might have been reported that the average mathematics score of fourth graders went up (or down) 4 points on a 500-point scale. There was no way of knowing whether the previous score represented strong or weak performance and whether the amount of change should give cause for concern or celebration. In contrast, NAEP now also reports the percentage of students who are performing at or above “basic,” “proficient,” and “advanced” levels of achievement. Proficient, the central level, represents “competency over challenging subject matter,” as demonstrated by how well students perform on the questions on each NAEP test. Basic denotes partial mastery and advanced signifies superior performance on NAEP. Using achievement levels to report results and track changes allows readers to make judgments about whether performance is adequate, whether “progress” is sufficient, and how NAEP standards and results compare to those of other tests, such as state and local tests.

First employed in 1990, the achievement levels have been the subject of several independent evaluations and some controversy. Information from these evaluations, as well as from other experts, has been used over the last 6 years to improve and refine the procedures by which achievement levels are set. Although the current procedures may be among the most comprehensive and sophisticated standard-setting procedures used in education, the Governing Board remains committed to improving the process and to the continuing conduct of validity studies.

- **The NAEP shall continue to report student achievement results based on performance standards.**

## Use international comparisons

Looking at student performance and curriculum expectations in other nations is yet another way to consider the adequacy of U.S. student performance. NAEP is, and should be, a domestic assessment. However, decisions on the content of NAEP tests, the achievement standards, and the interpretation of test results, where feasible, should be informed, in part, by the expectations for education set by other countries, such as Japan, Germany, and England. Although there are technical hurdles to overcome, consideration of such qualitative information can be used to good effect. In addition, NAEP should promote “linking” studies with international assessments, as has been done with the Third International Mathematics and Science Study, so that states that participate NAEP can have state, national, and international comparisons. This, in turn, should take into account problems in making international comparisons truly comparable, such as differences in the samples of students tested, differences in the curricula, and differences in the translated test questions.

- **NAEP test frameworks, test specifications, achievement levels, and data interpretations shall take into account, where feasible, curricula, standards, and student performance in other nations.**
- **NAEP shall promote “linking” studies with international assessments.**

## Emphasize reporting for grades 4, 8, and 12

An aspect of the NAEP design that needs reconsideration is age versus grade-based reporting. At its inception, NAEP tested only by age. Current law requires testing both by age (ages 9, 13, and 17) and by grade (grades 4, 8, and 12). Grade-based results are generally more useful than age-based results. Schools and curricula are organized by grade, not by age. Grades 4, 8, and 12 mark key transition points in American education. Grade 12 performance is particularly important as an “exit” measure from the K–12 education system. Grades 4, 8, and 12 are specified for monitoring in National Education Goal 3. Age-based samples may be more appropriate with respect to international comparisons and, given high school dropout rates, would be more inclusive for age 17 than for grade 12 samples, which are limited to youth enrolled in school. However, assessing the knowledge and skills of out-of-school youth may properly fall under the purpose of another program, such as the National Adult Literacy Survey.

Although grade-based reporting is generally preferable, there is a problem about the accuracy of grade 12 NAEP results. At grade 12, a smaller percentage of schools and students that are invited actually participate in testing than is the case with fourth and eighth graders. Also, more 12th graders fail to complete their tests than do fourth and eighth graders. In addition, when asked, “How hard did you try on this test?” and “How important is doing well on this test?” many more 12th graders than 4th or 8th graders say that they didn’t try hard and that the test wasn’t important. Low participation rates, low completion rates, and indicators of low motivation suggest that the NAEP may be underestimating what 12th graders know and can do.

One possible reason for low response and low motivation is that schools and students receive very little in return for their participation in the NAEP beyond the knowledge that they are performing a public service. They do not receive test scores nor do they receive other information from the NAEP that teachers and principals might wish to use as a part of the instructional program. This should be changed. The NAEP design should use meaningful, practical incentives that will give school principals and teachers a greater reason to participate and students more of a reason to try harder. The underlying idea is clear: if principals and teachers see direct benefits, they are more likely to agree to participate in the NAEP. Students may be more likely to take the assessment seriously if they see that their teachers and principals are enthusiastic about participating. Without practical incentives, even at grades 4 and 8, the willingness of district and school administrators and staff to participate in NAEP may diminish over time.

- **NAEP shall continue to test in and report results for grades 4, 8, and 12; however, in selected subjects, one or more of these grades may not be tested.**
- **Age-based testing and reporting shall be permitted when deemed appropriate and when necessary for international comparisons and for long-term trends, should the Governing Board decide to continue long-term trends in their current form.**
- **Grade 12 results shall be accompanied by clear, highlighted statements about school and student participation, student motivation, and cautions, where appropriate, about interpreting 12th grade achievement results.**
- **The NAEP design shall seek to improve school and student participation rates and student motivation at grade 12.**
- **The NAEP shall provide practical incentives for school and district participation at grades 4, 8, and 12.**

## **Use innovations in measurement and reporting**

NAEP has a record of innovations in large-scale testing. These include the early use of performance items, sampling both students and test questions, using standards describing what students should know and be able to do, and employing computers for such things as inventory control, scoring, data analysis, and reporting. NAEP should continue to incorporate promising innovative approaches to test administration and improved methods for measuring and reporting student achievement.

Technology can help improve NAEP reporting and testing. For example, reports could be put on computer disc, transmitted electronically, and made available on the World Wide Web. Test questions could be catalogued and made available online for use by state assessment personnel and classroom teachers. Also, NAEP could be administered by computer, eliminating the need for costly test booklet systems and reducing steps related to data entry of student responses. Students could answer “performance items” in cost-effective, computerized formats. The increasing use of computers in schools may make it

feasible to administer some parts of the NAEP by computer under the next contract for NAEP, beginning around the year 2000.

Other examples of promising methods for measuring and reporting student achievement include adaptive testing and domain-score reporting. In adaptive testing, each student is given a short “pre-test” to estimate that student’s level of achievement. Students are then administered test exercises that are in the range of difficulty indicated by the pre-test. Since the test is “adapted” to the individual, it is more precise and can be markedly more efficient than regular test administration. In domain-score reporting, a subject (or “domain”) is well defined, a goodly number of test questions are developed that encompass the subject, and student results are reported as a percentage of the “domain” that students “know and can do.” This is in contrast to reporting results using an arbitrary scale, such as the 0–500 scale used in NAEP.

- **NAEP shall assess the merits of advances related to technology and the measurement and reporting of student achievement.**
- **Where warranted, NAEP shall implement such advances in order to reduce costs and/or improve test administration, measurement, and reporting.**
- **The next competition for NAEP contracts, for assessments beginning around the year 2000, shall ask bidders to provide a plan for (1) conducting testing by computer in at least one subject at one grade, and (2) making use of technology to improve test administration, measurement, and reporting.**

**OBJECTIVE 2: To develop, through a broadly inclusive process, sound assessments to measure what students know and can do as well as what students should know and be able to do.**

## **Keep test frameworks and specifications stable**

Test frameworks spell out in general terms how an assessment will be put together. The frameworks also determine what will be reported and influence how expensive an assessment will be. Should eighth grade mathematics include algebra questions? Should there be both multiple-choice questions and questions in which students show their work? What is the best mix of such types of questions for each grade? Which grades are appropriate for assessment in a subject area? Test specifications provide detailed instructions to the test writers about the specific content to be tested at each grade, how test questions will be scored, and the format for each test question (e.g., multiple choice, essay, etc.).

Since 1989, the Governing Board has been responsible for developing test frameworks and specifications for NAEP. The Governing Board has done this through a broadly inclusive process, involving hundreds of teachers, curriculum experts, directors of

state and local testing programs, administrators, policymakers, practitioners in the content area (e.g., chemists for science, demographers for geography, etc.) and members of the public. This process helps determine what is important for NAEP to test, how it should be measured, and how much of what is measured by the NAEP students should know and be able to do in each subject.

The process of developing frameworks and specifications involves consideration of both current classroom teaching practices and important developments in each subject area for inclusion in NAEP. In order to ensure that NAEP data fairly represent student achievement, the test frameworks and specifications are subjected to wide public review before adoption and test questions developed for NAEP are reviewed for relevance and quality by representatives from participating states.

An important role of NAEP is to report on trends in student achievement over time. For NAEP to be able to measure trends, the frameworks (and hence the tests) must remain stable. However, as new knowledge is gained in subject areas and as teaching practices change and evolve, pressures arise to change the test frameworks and tests to keep them current. But, if frameworks, specifications, and tests change too frequently, trends may be lost, costs go up, and reporting time may increase.

- **Test frameworks and test specifications developed for NAEP generally shall remain stable for at least 10 years.**
- **To ensure that trend results can be reported, the pool of test questions developed in each subject for the NAEP shall provide a stable measure of student performance for at least 10 years.**
- **In rare circumstances, such as where significant changes in curricula have occurred, the Governing Board may consider making changes to test frameworks and specifications before 10 years have elapsed.**
- **In developing new test frameworks and specifications, or in making major alterations to approved frameworks and specifications, the cost of the resulting assessment shall be estimated. The Governing Board will consider the effect of that cost on the ability to test other subjects before approving a proposed test framework and/or specifications.**

## **Use an appropriate mix of multiple-choice and “performance” questions**

To provide information about “what students know and can do,” NAEP uses both multiple-choice questions and questions in which students are asked to produce their own answers, such as writing a response to an essay question or explaining how they solved a math problem. Questions of the latter type are sometimes called “performance items.” Both types of questions can vary in difficulty and the richness of information they provide, and may require students to demonstrate different kinds of skills and knowledge.

Performance items are desired because they provide direct evidence of what students can do. They range in length of test taking time from a short-answer or fill-in-the-blank format requiring about a minute of response time, to items requiring about 5 minutes of response time, to writing exercises that may allow 15 to 50 minutes response time. Although they may be desirable, performance items are more expensive than multiple-choice to develop, administer, and score. In addition, much larger proportions of students fail to respond to performance items, particularly as the amount of required response time increases.

Multiple-choice questions can be challenging and are desired because they are efficient in collecting information about student knowledge. However, multiple-choice questions are more subject to guessing than are performance items.

Currently, all students tested by NAEP are given both types of questions. Generally, about half the testing time is devoted to each type of question, but the amount of time for each differs based on the skills and knowledge to be assessed, as established in the NAEP test frameworks. For example, in a writing assessment, all students are asked to write their responses to specific exercises. In other subjects, the mix of multiple-choice and performance items varies. The appropriate mix of items for each subject should be determined by the nature of the subject, the range of skills to be assessed, and cost.

- **Both multiple-choice and performance items shall continue to be used in NAEP;**
- **In developing new test frameworks, specifications, and questions, decisions about the appropriate mix of multiple-choice and performance items shall take into account the nature of the subject, the range of skills to be assessed, and cost.**

**OBJECTIVE 3: To help states and others link their assessments with NAEP and use NAEP data to improve education performance.**

The primary job of NAEP is to report frequently and promptly to the American public on student achievement. The resources of NAEP must be focused on this central purpose if it is to be achieved. However, the products of NAEP—test frameworks, specifications, scoring guides, results, questions, achievement levels, and background data—are widely regarded as being of high quality. They are developed with public funds and, therefore, should be available for public use as long as such uses do not threaten the integrity of NAEP or its ability to report regularly on student achievement.

NAEP should be designed in a way that permits its use by others, while protecting the privacy of students, teachers, and principals who have participated in NAEP. This should include making NAEP test questions and data easy to access and use, and providing related technical assistance upon request. Generally, the costs of a project should be borne by the individual or group making the proposal, not by NAEP.

Examples of areas in which particular interest has been expressed for using the NAEP include linking state and local tests with NAEP and performing in-depth analysis on NAEP data. States that link their tests to NAEP would have an unbiased external benchmark to help make judgments about their own tests and standards and also would have a means for comparing their tests and standards with those of other states.

**NAEP shall develop policies, practices, and procedures that assist states, school districts, and others who want to do so at their own cost to link their test results to NAEP.**

- **NAEP shall be designed so that others may access and use NAEP test frameworks, specifications, scoring guides, results, questions, achievement levels, and background data.**
- **NAEP shall employ safeguards to protect the integrity of the NAEP program, prevent misuse of data, and ensure the privacy of individual test takers.**