NAEP and No Child Left Behind: Technical Challenges and Practical Solutions

Catherine M. Hombo

Theory Into Practice, Volume 42, Number 1, Winter 2003, pp. 59-65
(Article)

Published by Ohio State University College of Education

Catherine M. Hombo

NAEP and No Child Left Behind: Technical Challenges and Practical Solutions

The National Assessment of Educational Progress (NAEP) is the only nationally representative and continuing assessment of what U.S. students know and can do in various subject areas. The No Child Left Behind Act of 2001 requires that all U.S. states, jurisdictions, and territories submit plans to the Secretary of Education that demonstrate that the state has adopted challenging academic content and student achievement standards. As part of that plan, states and school districts that receive Title I funds must participate in NAEP assessments in reading and math at grades 4 and 8 every two years. The purpose of NAEP has always been to survey what students know and can do and to monitor changes over time. Now, NAEP has a new role, to act as a serious discussion tool in evaluating results of state assessments and in providing a common base for comparison between states. This role brings new challenges and opportunities for the NAEP program.

What is NAEP?

The National Assessment of Educational Progress (NAEP) is a survey of student achievement for the United States, measuring academic progress at both the national and state levels. NAEP assesses what students know and can do in a variety of subject areas, including reading, mathematics, science, writing, history, geography, and the arts.

From its first administration in 1969 until 1989, the goal of NAEP was to measure academic achievement at the national level, and to measure trends in academic progress. Unlike many state assessment programs, NAEP did not evaluate the performance of individual students; indeed the authorizing legislation specifically prohibits revealing information about individual students or schools. NAEP is a survey, the goal of which is to report on the achievement of the nation and important demographic subgroups, and to track trends in educational progress over time.

NAEP is federally mandated, and the U.S. Congress and president must renew the authorizing law periodically. NAEP is directed by the National Center for Education Statistics, part of the federal Department of Education. An independent board, the National Assessment Governing Board (NAGB), provides policy direction, selecting subjects to be assessed and developing the frameworks that form the basis of every NAEP assessment. In addition to developing frameworks, NAGB also develops achievement levels for each subject area NAEP assesses. The achievement levels are intended to illustrate "appropriate performance standards for each grade in each subject area to be tested

Catherine M. Hombo is a research scientist at the Educational Testing Service, Princeton, NJ.

THEORY INTO PRACTICE, Volume 42, Number 1, Winter 2003

under the National Assessment" (Improving America's Schools Act, 1994). The percentage of students achieving at or above these levels, labeled *Basic*, *Proficient*, and *Advanced*, adds to the interpretation of student performance on NAEP.

State NAEP

In 1990, NAEP administered the first Trial State Assessment (the term *trial* was dropped beginning with the 1996 state assessment). For the first time, a goal for NAEP was to report on the academic achievement of individual states, and to track academic progress in individual states.

State NAEP fulfills a unique role in the U.S. educational system. The United States has a long tradition of local control over schools. States and local school boards determine school curricula and how achievement is evaluated. This means that there is no consistent testing policy, and different states (or even districts) evaluate their students, teachers, and schools differently. Some states use one of a variety of commercial tests while others have developed their own state assessments. Because the assessments are not comparable, comparisons across states cannot be validly made. Moreover, measurement of academic progress within some states can be difficult, because the states may change tests from time to time. NAEP provides the missing common measure of student achievement so that state-to-state comparisons can be made. The assessment frameworks developed by NAGB take into consideration broad national input from a variety of stakeholders and experts. This process assists NAGB in identifying the content and structure for each assessment.

In addition, NAEP supplies national statistics for federal government planning and to inform decisions regarding policy planning. Important goals of NAEP are to provide a broad measure of achievement in different subject areas, to report on factors related to achievement, and to measure trends in academic progress at the national and state level over time. To achieve these goals, NAEP uses data from students, teachers, schools, and other sources.

New Legislation, New Roles

In 2002, as part of the No Child Left Behind (NCLB) Act of 2001, each state, territory, and jurisdiction receiving Title I funds must submit a plan to the U.S. Department of Education (USDOE). This plan must demonstrate that the state has adopted challenging academic content and student academic achievement standards to be used by the state and its local educational agencies. A statewide accountability system must be implemented that verifies the adequate yearly progress of all public schools under the direction of the state and its local agencies toward state academic achievement goals. The states must provide evidence to the Secretary of Education that the assessments used in the measurement of progress are of acceptable technical quality. States that do not submit a plan approved by the USDOE may lose some federal Title I education funds.

As part of the state plans, states and school districts that receive Title I funds must participate, if selected, in NAEP assessments in reading and math at grades 4 and 8 every two years. Until this change, state participation in NAEP had been voluntary. NAEP is still voluntary for students, and participation is still voluntary for schools and districts in all subjects other than reading and mathematics.

Until recently, the purpose of NAEP has always been to survey what students know and can do, and to monitor changes over time. Now, NAEP has a new role: to act as a serious discussion tool in evaluating results of state assessments, and in providing a common base for comparison between states. As NAEP assumes this role, it is increasingly important for practitioners and state education officials to understand what NAEP is, how it is administered, how the data are analyzed, and how scores are reported (see Appendix A for a detailed description of NAEP procedures).

Changes to NAEP in NCLB

NAEP student data are collected according to a multi-level, multi-stage sampling framework. Through these techniques, a representative sample of students is selected to respond to a subset of NAEP items. In addition to student cognitive item responses, NAEP collects background data from students, teachers, and schools for use in the analysis, as well as data from other sources such as the U.S. Census or state achievement test scores. Minority students and students attending private

schools are oversampled in NAEP to provide a sufficiently large sample for the subgroup analysis and reporting mandated by the U.S. Congress. The student sample for a state is drawn to be representative of the state, while the national sample is drawn to be representative of the nation as a whole. Prior to the 2002 assessment, the national and state NAEP samples were independent. In 2002 the sample design was altered so that the national sample is now a subsample of the combined state samples.

Prior to 2002, state school personnel collected the state assessment data. Starting in 2002, administrative changes were authorized so that a federal contractor coordinates and administers NAEP. The federal government pays the full cost of administering NAEP, relieving the states of the necessity to provide staff for this activity. In addition, having a single contractor responsible for all NAEP field data collection provides continuity and consistency throughout the assessment.

New Challenges Posed by NCLB

The advent of the NCLB Act and NAEP's role in that legislation bring new opportunities and new challenges. As NAEP begins this effort, new issues must be addressed. Some of the specific challenges currently being investigated are described briefly below.

Changes in population

The new NCLB regulations regarding participation in NAEP have changed significantly, and this has implications for the program. Any change in the population being measured may have implications for the validity of the measure and for trends and must be carefully evaluated. A concern that has received attention in recent years is the inclusion in NAEP of students with disabilities and students with limited English proficiency (LEP). The program has devoted substantial resources to careful study of the best way to increase participation of special needs students in NAEP (Mazzeo, Carlson, Voelkl, & Lutkus, 2000), and such research continues to date. As an example of the potential impact of NCLB on the student population being assessed, it may be that greater numbers of students with disabilities and LEP students, all of whom are required to take an assessment in their home state, may also participate in NAEP assessments in the future. Also, the tie between required NAEP participation and receipt of Title I funding may mean that the participation of districts and schools receiving those funds will change at a rate different from districts or schools who do not receive such funds.

Changes in assessment behavior

Another challenge is related to the effect of student motivation. NAEP has traditionally been a low-stakes assessment, in which scores for students were not produced and there were no consequences associated with the assessment results. NAEP's new role in NCLB does not change the stakes for students at all, but the increased attention and publicity may create the impression that the assessment is now higher stakes than before.

A related concern is the possibility of teaching to the test. This concern is raised when tests have high-stakes consequences for teachers and students. Indeed, many state assessment programs have faced this issue when implementing examinations that students must pass in order to graduate from high school. The design of NAEP is an asset in alleviating these concerns. Because of the way NAEP items are assigned to test booklets, it is not possible to know in advance with what items, or even in what subject area, a student will be assessed. In addition, students are sampled across classrooms in a school, so preparation of an entire class or grade for NAEP would be difficult and not useful for students not selected for the assessment.

Changes in procedures

Assessment administration conditions have always been standardized in NAEP, but the implementation of contractor administration of all sessions is new. Before 2002, school personnel presented state assessment sessions, although contractor personnel randomly monitored a proportion of the state sessions as well. No meaningful differences have been observed between the monitored and unmonitored sessions in the past. Beginning in 2002, the state assessment sessions were administered by contractor personnel. It is not expected that this change in administration conditions will have much impact, but it is a change that will be monitored.

NAEP has traditionally been a voluntary assessment, and continues to be so for students. Under the NCLB, parents of students selected into a NAEP sample must be informed before the assessment is administered that their child is not required to participate in the assessment, is not required to complete the assessment, and is not required to respond to any item on the assessment. Schools and students selected for the sample do not all participate. The data from the nonparticipants may not be missing at random, but instead may in some way be related to the variables of interest in the assessment. Ignoring this could bias the assessment results.

Utilization of NAEP results

The NCLB calls for NAEP to be used as a serious discussion tool to provide a context for state assessment scores. However, the exact role and uses are still being defined. Similarly, adequate yearly progress, and how to assess it, are both subjects of ongoing discussion. This is a policy issue that NAGB has devoted serious study to in the past year. A report, "Using the National Assessment of Educational Progress to Confirm State Test Results" (NAGB, 2002) discusses these concerns in more detail.

A specific example of the issues involved is closing the gaps between specified subgroups of students. This raises new concerns and issues. Holland (2002a, 2002b) has studied some of the technical issues involved in measuring gaps and measuring the closing of gaps. The interested reader is referred to those reports for a more detailed discussion on this topic.

Conclusion

NAEP has played a vital role in assessing and reporting on the state of education in the United States for more than 30 years. NAEP assessment and analysis techniques are complex. They have been developed explicitly so that the scale score and trend measure results are as precise and reliable as the current state of research in teaching and learning, content areas, psychometrics, and statistics can make them. Techniques developed in NAEP are used in a variety of other settings in educational measurement, and NAEP is widely considered to be the gold standard of educational achievement survey assessment. Under No Child Left Behind, NAEP's role is increasingly important, and state NAEP scores will become the focus of more attention and discussion. Technical issues inherent to the new role are currently being investigated and resolved. As this emphasis on NAEP scores increases, teachers and practitioners will want to become more informed about the procedures and processes that go into a NAEP report.

APPENDIX A
NAEP Procedures

Under the No Child Left Behind Act, NAEP will play an ever more prominent role in the evaluation of adequate yearly progress in the states, jurisdiction, and territories receiving federal Title I assistance. It is critical that these stakeholders have a clear understanding of NAEP procedures, including test development, scoring, sampling, data collection and analysis, and reporting.

Test development

All NAEP assessments are developed to represent content-area frameworks. These frameworks are developed by the National Assessment Governing Board (NAGB) with broad input from stakeholders, including teachers, subject-matter specialists, testing experts, and interested members of the general public. The resulting frameworks specify content, outcomes, and item formats. Each subject area assessed in NAEP also has a test development committee that selects NAEP items from a pool written by teachers and content specialists. The test development committees meet quarterly. To ensure that NAEP assessments continue to remain relevant to current educational practice, subject-area frameworks are reviewed and revised approximately every 10-12 years. If appropriate, a new framework is developed for a subject area. When this occurs, the old NAEP trend lines are discontinued and new ones are started.

NAEP items undergo extensive scrutiny, review, and pretesting before operational use. Items must survive reviews by test specialists, editorial reviews, fairness reviews, small-scale pilot tests, and full-scale national field tests. Additional reviews by state curriculum and testing personnel are employed for state NAEP subjects. Statistical

analyses determine if the item is functioning as intended, and differential item functioning (DIF) analyses are completed in order to remove any unfair items from the operational assessments.

NAEP uses both objective and constructed response items in order to achieve a thorough and valid assessment of content. Performance assessment items range from short answer through extended constructed response, complex mathematics items, essays, and performance items such as science experiments and musical performances. These items make up a substantial proportion of every NAEP assessment.

Scoring

Scoring a NAEP assessment is an intricate business. Books are scanned so that the multiple-choice item responses can be machine-scored. Images of the student responses to the constructed-response items are scanned and presented via computer to human raters for scoring. In 2002, the operational reading assessment used more than 350 professional raters to score 150 items, for a total of more than 4 million student responses to constructed-response reading items.

The professional raters are carefully trained before being allowed to score operational student responses, and rating reliability is monitored within the assessment year through "second scoring" a percentage of the student responses. When subject assessments have a trend measure, cross-year reliability is also monitored through periodic rescoring of a set of student responses from the previous assessment(s). If the monitoring indicates that the scoring has fallen below reliability standards or is consistently shifting in one direction, item scores are discarded, scorers are retrained and scoring of the item is done over. These quality control measures over the constructed response scoring process, though challenging to complete, ensure that the data forming the basis of reported NAEP scale scores and trends are trustworthy.

Sampling

NAEP must operate under several program constraints. NAEP must cover broad content areas and report on subdomain proficiency within those content areas. The item pool used in any NAEP assessment must meet framework specifications about item types and subdomain representation. As a result, NAEP item pools for a single grade often contain between 100 and 200 items, and the aggregate testing time for a single student to complete the entire NAEP assessment item pool in a subject area is impractically long, typically several hours.

Therefore, all NAEP results in all assessments are based on samples, both of students and items. NAEP sample designs must meet three criteria: they must (a) be practical, (b) lead to efficient administration procedures, and (c) produce acceptable levels of precision for the target statistics. Simple random sampling is neither practical nor efficient, so NAEP uses multistage complex sampling procedures.

The student samples are drawn to be representative of the nation (in the national samples) or the state (in the state samples). Starting in 2002, in subjects where state assessments are given, the state and national samples have been combined so that the national sample is a subset of the aggregation of the state samples. Single grade samples range in size from approximately 8,000 for a national-only subject to 150,000 for a combined state and national assessment subject. State samples are approximately 2,500 students. Certain subgroups of interest, such as minority students or students attending private schools, are oversampled so that sufficient precision in the results for these subgroups may be obtained. Because the resulting samples are not simple random samples, they require the use of sampling weights in analysis.

NAEP assessments are usually administered to groups of 25-30 students, although exceptions are made for students who require accommodations to meaningfully participate in the assessment. Multiple subjects are assessed in the same session, as the books are designed with common sets of directions and timing. Item blocks and sections of background questions are separately timed, and the session administrator indicates when the time to complete a section is over. Testing is done at the students' school, and "intact classroom" samples are not taken.

In addition to the student samples, items are also sampled into test books. Because no student can realistically complete the entire assessment, the total item pool is divided into "blocks" of items. These blocks of items are selected to meet time

and content constraints. Blocks of items differ with respect to content coverage within subject area, item type, difficulty, and number of items. The blocks of items are assembled into test books according to a balanced incomplete block (BIB) design. The design is balanced in that each block of items appears equally often in the total set of books, each block appears in every position in the set of test books, and each block of items appears paired with each other block of items. The design is incomplete in that the conditions for position and occurrence of each other block with each other block are not crossed. Generally, this results in more than 20 test books per grade per subject.

Use of student and item samples has implications for analysis. Much of standard statistical theory and many existing software programs that implement statistical tests assume independent and identically distributed (i.i.d.) observations. NAEP data are not i.i.d.; observations are clustered within schools and geographically. Also, most simple data analysis procedures assume simple random sampling (SRS). By oversampling subgroups of interest, NAEP data violate simple random sampling assumptions. Thus, if such standard analysis procedures are used with NAEP data, estimates of simple descriptive statistics will be biased. Weights must be used in the analysis of NAEP data to correct for the effects of the oversampling. Estimating standard errors for NAEP target statistics is complex because using SRS-based formulas will underestimate the actual sampling variability in the complex sample. NAEP uses a jackknife procedure to achieve a more accurate estimate of the sampling variability and, thus, provides accurate statistical tests of the results.

Data analysis

After every assessment, NAEP releases and replaces approximately 30% of the items in the pool. As a result, across trend years in NAEP assessments, item pools are not identical. Blocks of items are released, and the replacement blocks do not have precisely the same characteristics as those released. Some items are unique to the current year, and some are common across years. NAEP analysis methods combine results across different test books within an assessment year, and produce results that are comparable across assessment years despite changes in the item pool. The results provide information about student proficiency in the subject area and content subdomains in terms of average scores and percentages of students at or above the Achievement Levels for the subject, in addition to providing information about trends over time.

NAEP achieves these goals using analysis techniques based on Item Response Theory (IRT) (e.g., Lord, 1980), and through the use of analysis methods developed by Mislevy (1991), Rubin (1987) and others. The IRT item parameter estimation is completed in a context in which no student responds to the entire item pool, but the analysis results must provide an estimate of group scores on the "whole" assessment. This requires tracking of the multiple book compositions, block positions within books, and the varied connections between blocks. Background variables for each student, including relevant teacher and school variables, must be matched to the correct students. Students do not respond to enough cognitive subject-area items to yield reliable individual scores; however, NAEP data collection and analysis procedures have been specifically designed to provide unbiased and reliable estimates of student population and subgroup scores.

Measuring trends across time requires that assessment results be reported on a common scale. NAEP scale metrics are set at the beginning of a subject trend line. Since the analysis results are in an arbitrary metric, typical of IRT, a scale score metric is selected that is thought to be easily interpreted by the general public. Trend measures are maintained by analyzing both the current and the previous assessment years' data together. The common item blocks in the two assessment years are estimated with a single set of item parameters, unless there is strong evidence of differential functioning across assessment years. In those cases (usually few) the item is estimated with two separate sets of item parameters, one for each year. At the end of the data analysis, the proficiency distribution for the previous assessment's data is reestimated. Once this process is completed, the reestimated proficiency distribution for the previous years' assessment data is transformed to the reporting metric. The same transformation is applied to the

current year's assessment results, placing both years' results in the common scale score metric. Further details of NAEP data analysis procedures are available in *The NAEP 1998 Technical Report* (Allen, Donoghue, & Schoeps, 2001).

NAEP reporting

NAEP results are reported to the general public, and so must be presented in as understandable a format as possible. The score reports must support a broad range of uses by researchers, educators, policy makers, politicians, and interested members of the general public. Scores are reported in print format, and a large amount of NAEP data is accessible on the NCES web site (http://www.nces.ed.gov). Special data tools have been developed to allow examination of NAEP results in detail, and released NAEP items (with response data information and sample student responses) are also available. All NAEP documentation, including detailed technical documentation of every NAEP assessment year, is available from the USDOE, and can be ordered from the web site listed above.

References

Allen, N.L., Donoghue, J.R., & Schoeps, T.L. (2001, August). *The NAEP 1998 technical report* (NCES 2001509). Washington, DC: National Center for Education Statistics. Available at http://nces.ed.gov/pubsearch/getpubcats.asp?sid=031

Holland, P.W. (2002a). Measuring progress in student achievement: Changes in scores and score-gaps over time. In *Using the National Assessment of Educational Progress to confirm state test results* (Appendix B, pp. 1-27). Washington, DC: National Assessment Governing Board.

Holland, P.W. (2002b). Two measures of change in the gaps between the CDFs of test score distributions. *Journal of Educational and Behavioral Statistics, 27*, 3-17.

Improving America's Schools Act, Public Law 103-382, 20 USC 9010 (1994).

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mazzeo, J., Carlson, J.E., Voelkl, K.E., & Lutkus, A.D. (2000, February). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES 2000473). Washington, DC: U.S. Department of Education.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177-196.

National Assessment Governing Board. (2002, March). *Using the National Assessment of Educational Progress to confirm state test results*. Washington, DC: Author.

Rubin, D.B. (1987). *Multiple imputations for nonresponse in surveys*. New York: John Wiley & Sons.