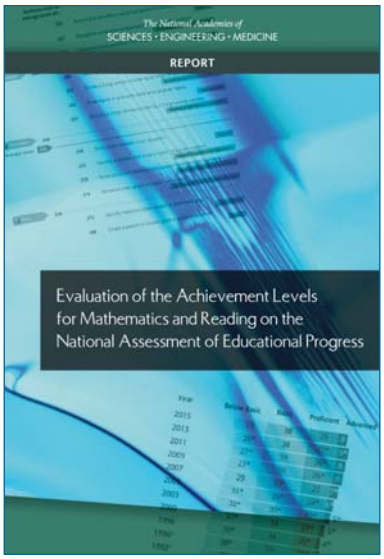
Report

April 2017

HIGHLIGHTS

Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress



The National Assessment of Educational Progress—often referred to as the Nation's Report Card or NAEP—has provided policy makers, educators, and the public with reports on the academic performance and progress of the nation's students since 1969. For more than two decades, results have been reported in achievement levels, noting what percentage of students fall into the categories Basic, Proficient, and Advanced. Achievement-level reporting is intended to make the results easier to understand—to provide a concrete way to explain what students know and can do at each of the levels.

In 2014, the National Academies of Sciences, Engineering, and Medicine appointed a committee to evaluate the extent to which the achievement levels in mathematics and reading are **reasonable, reliable, valid, and informative to the public** and to make recommendations about ways that the setting and use of achievement levels can be improved. The committee's report, *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress* (2017)

- finds that **additional work is needed to evaluate the validity of the achievement levels** and the extent to which they are aligned with each other and with other elements of NAEP.
- acknowledges that achievement-level reporting can be a useful way to communicate assessment results, but they identified many instances of misinterpretation and misuse of NAEP achievement levels. **Clear, accessible guidance on appropriate and inappropriate interpretations is essential.**
- recommends that along with interpretive guidance, NAEP provide research findings that document that these uses are valid and appropriate.
- recommends that to enhance understanding of the results, research should be conducted to find links between NAEP performance and real-world outcomes—a step that would make the results more meaningful and useful to the assessment's many audiences.

BACKGROUND: NAEP AND ACHIEVEMENT LEVELS

NAEP is given periodically in a variety of subjects—mathematics, reading, writing, science, the arts, civics, economics, geography, U.S. history, and technology and engineering literacy—to representative groups of students across the country. Scores are not reported for individual students. Instead, average scores are reported for the nation and for specific groups of students—for example, by state, gender, race and ethnicity, and socioeconomic status.

Since 1983, the results have been reported as "scale scores"—average scores on a scale ranging from 0 to 500. Over time, there was increasing interest in having the results reported in a way that policy makers and the public could understand

and use to examine students' achievement in relation to high, world-class standards. In response to that interest, three achievement levels were established—Basic, Proficient, and Advanced—and since 1992 NAEP reports show the percentage of students whose scores fall into each achievement level, as well as the percentage of students who score Below Basic. Scale scores continue to be reported as well. The box below shows the policy definitions of the achievement levels, and Figure 1 shows an example of achievement-level reporting.

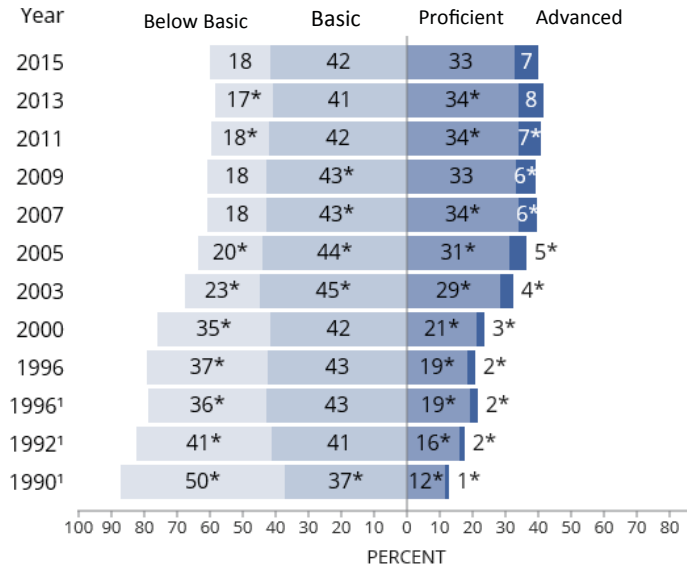


Figure 1 Percentage of students scoring at each achievement level in fourth-grade NAEP mathematics: 1990 through 2013.
Source:https://www.nationsreportcard.gov/reading_math_2015/#mathematics/acl?grade=4.

SETTING NAEP ACHIEVEMENT LEVELS

The first step in developing achievement levels is to set standards, a process that involves determining "how good is good enough." For example, what must a child be able to know and do in order to be considered a "proficient" fourth-grade reader? When the standards were set, feedback was sought from a wide range of experts and stakeholders—educators, administrators, subject-matter specialists, policy makers, parent groups, professional organizations, and the general public.

Through this process, a set of achievement levels was adopted for each subject area and grade that included "achievement-level descriptors"—a description of the knowledge and skills necessary to perform at the Basic, Proficient, and Advanced levels—and "cut scores," the minimum scores needed to attain each achievement level.

NAEP's first standard settings were conducted for the 1992 mathematics and reading assessments. It is important to point out that while standard setting can be done in a systematic, carefully controlled way, it is still a subjective process that relies on the judgments of trained experts. Independent evaluations of these standard settings highlighted numerous concerns. As a result, Congress stipulated that until an evaluation determined that the achievement levels are reasonable, reliable, valid, and informative to the public, they were to be designated as "trial"—a provisional status that still remains, 22 years later.

NAEP ACHIEVEMENT-LEVEL POLICY DEFINITIONS

There are three achievement levels for each grade assessed by NAEP (4, 8, and 12): Basic, Proficient, and Advanced. The following definitions apply to all subjects and all grades assessed.

Basic. This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Proficient. This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Advanced. This higher level signifies superior performance.

Source: https://nces.ed.gov/nationsreportcard/achievement.aspx.

FINDINGS FROM THE EVALUATION

The committee examined the available documentation of the 1992 standard settings in reading and mathematics and the evaluations of them, as well as the body of research on standard setting that has accumulated over the past two decades.

Overall, the committee found that the process was well documented, providing the types of evidence called for in the Standards for Educational and Psychological Testing in place at the time and currently. However, the evidence that resulted from that analysis raises questions about the extent to which the achievement levels are reasonable, reliable, valid, and informative to the public.

RELIABILITY EVIDENCE

Reliability indicates the extent to which judgments about where cut scores should be placed are consistent when standard setting is repeated—for example, with different panelists, with different test questions, and on different occasions.

During the 1992 standard setting, there was considerable variation among panelists' judgments about where cut scores should be placed. The sources and impact of this variation were not fully addressed before achievement-level results were released to the public and have not yet been fully resolved.

VALIDITY EVIDENCE

Validity refers to the extent to which test results mean what they are intended to mean and can legitimately be used in the way they are intended to be used. In the context of standard setting, validation usually consists of demonstrating that the proposed cut score for each achievement level corresponds to the descriptor, and that the achievement levels are set at a reasonable level, not too low or too high. The committee reviewed results from studies designed to provide content-related validity evidence and criterion-related validity evidence.

For each grade and subject tested by NAEP, there is a content framework—a detailed description of what students know and should be able to do at that grade. The framework serves as the basis for developing the test questions, the achievement-level descriptors, and the cut scores. For valid inferences to be drawn about student achievement, all of these elements should be aligned. This alignment provides content-related validity evidence.

Since the original standard setting, there have been changes to the frameworks, the pool of test questions, the assessment itself, and the achievement-level descriptors. Some research has been conducted to evaluate the alignment between the cut scores and the achievement-level descriptors and to make appropriate adjustments, but more of these alignment studies are needed.

EVIDENCE THAT THE ACHIEVEMENT LEVELS ARE REASONABLE

One way to evaluate the extent to which NAEP achievement levels are reasonable is to examine their correspondence to other measures of the same or similar skills. These kinds of studies can provide criterion-related validity evidence.

The committee compared NAEP achievement-level results with those for two international assessments—the Programme for International Student Assessment (PISA) and the mathematics benchmarks on Trends in International Mathematics and Science (TIMSS)—as well as for the advanced placement (AP) exams in reading and math. These studies show that the NAEP achievement-level results are generally consistent with those from other assessments.

Specifically, the percentage of students scoring at the proficient and advanced levels on NAEP are generally consistent with the percentage of U.S. students scoring at the reading and mathematics benchmarks on PISA and TIMSS and at the higher levels for the AP exams. These studies also show that significant numbers of students in other countries score at the equivalent of the NAEP advanced level.

EVIDENCE THAT THE ACHIEVEMENT LEVELS ARE INFORMATIVE

The committee was unable to find any official documents that provide guidance on the intended interpretations and uses of NAEP achievement levels, beyond brief statements in two policy documents. The committee was also unable to find documents that specifically lay out appropriate uses and the associated research to support these uses.

The committee found a disconnect between the kind of validity evidence that has been collected and the kinds of interpretations and uses that are made of NAEP's reported results. That is, although the committee found evidence for the integrity and accuracy of the procedures used to set the achievement levels, the evidence does not extend to the uses of the achievement levels—the way that NAEP audiences use the results and the decisions they base on them. Without appropriate guidance, misuses of NAEP data are likely, and the committee found numerous types of inappropriate inferences.

Achievement-level reports cannot be informative to the public if the public does not know how to interpret them. Research is needed to articulate the intended interpretations and uses of the achievement levels and collect validity evidence to support these interpretations and uses. Studies are also needed to identify the ways audiences are actually interpreting and using the data and evaluate the validity of each of these. This information should be communicated to users, with clear guidance on substantiated and unsubstantiated interpretations.

IS A NEW STANDARD SETTING NEEDED?

The cut scores for grades 4 and 8 in mathematics and all grades in reading were set more than 24 years ago. Since then, there have been many adjustments to the frameworks, item pools, assessments, and achievement-level descriptors, but there has been no effort to set new cut scores for these assessments. While priority has been given to maintaining the trend lines, it is possible that there has been "drift" in the meaning of the cut scores, making it questionable whether inferences about trends are valid.

The committee concluded that there is evidence to support conducting a new standard setting for all grades in reading and mathematics. However, setting new cut scores would interrupt the NAEP trend line at a time when many other contextual factors are in flux—such as changes in the way NAEP results are used by states and districts—and when other changes are being considered for NAEP—such as a digital-based assessment.

In the short term, the committee notes that most of the significant arguments in favor of a new standard setting can be addressed by revising the achievement-level descriptors; that is, by continuing to follow the same cut scores but ensuring the descriptions are aligned with them.

In the long term, the committee recommends a thorough revision of the achievement-level descriptions that are informed by a suite of education, social, and economic outcomes important to key audiences. They envision a set of descriptions that correspond to a few salient outcomes, such as college readiness or international comparisons.

USERS NEED MORE GUIDANCE ON HOW TO USE ACHIEVEMENT LEVELS

Currently, guidance on how to interpret NAEP achievement level results is provided to users in an inconsistent and piecemeal way. Some audiences receive considerable guidance just prior to a release of the results, but for audiences that obtain most of their information from the Web-

site or hard-copy reports for the general public, interpretive guidance is hard to locate.

Actions are needed to improve the interpretation and use of NAEP reports, maintain the validity and usefulness of NAEP data, and ensure the currency of the NAEP achievement levels. The first step is to develop more concrete guidance for users on appropriate and inappropriate interpretations of achievement levels, to avoid NAEP's audiences attaching their own understandings to them. Moreover, the intended interpretations and uses of the achievement levels need to be articulated, and research is needed to collect validity evidence to support these interpretations and uses

In addition, the existing achievement-level descriptors may not provide users with enough information about what students at a given level know and can do. The descriptors need to be reviewed and revised to provide accurate and more specific information.

RESEARCH SHOULD SEEK LINKS BETWEEN NAEP SCORES, REAL-WORLD OUTCOMES

The notion of proficient (or basic or advanced) is abstract, and these terms are currently connected only to the assessment and the framework—not to real-world measures that hold value for the public. When a doctor is licensed or an accountant is certified, it is understood that the person has enough knowledge and skill to practice medicine or accounting. When someone is judged to be proficient in reading or mathematics, the obvious question is, "for what?"

It would be valuable for "proficient" to be linked to some real-world measures, such as relating 12th-grade reading and mathematics performance to college readiness, which would provide concrete meaning and connect the results to something the public values. Research is needed on the relationships between the NAEP achievement levels and measures external to NAEP, such as college readiness, being on track for a high school diploma (for 8th grade) and readiness for middle school (for 4th grade).

COMMITTEE ON THE EVALUATION OF NAEP ACHIEVEMENT LEVELS

CHRISTOPHER EDLEY, JR. (*Chair*), School of Law, University of California, Berkeley School of Law; **PETER AFFLERBACH**, Department of Teaching and Learning, Policy and Leadership, University of Maryland; **SYBILLA BECKMANN**, Department of Mathematics, University of Georgia; **H. RUSSELL BERNARD**, Institute for Social Science Research, Arizona State University, and Department of Anthropology, University of Florida; **KARLA EGAN**, National Center for the Improvement of Educational Assessment, Dover, NH; **DAVID J. FRANCIS**, Department of Psychology, University of Houston; **MARGARET E. GOERTZ**, Graduate School of Education, University of Pennsylvania (emerita); **LAURA HAMILTON**, Education Program, RAND Corporation, Pittsburgh, PA; **BRIAN JUNKER**, Department of Statistics Carnegie Mellon University; **SUZANNE LANE**, School of Education, University of Pittsburgh; **SHARON J. LEWIS**, Council of the Great City Schools, Washington, DC (retired); **BERNARD L. MADISON**, Department of Mathematics, University of Arkansas; **SCOTT NORTON**, Standards, Assessment, and Accountability, Council of Chief State School Officers, Washington, DC; **SHARON VAUGH**, College of Education, University of Texas at Austin; **LAURESS WISE**, Human Resources Research Organization, Monterey, CA; **JUDY KOENIG**, *Study Director*; **JORDYN WHITE**, *Program Officer;* **KELLY ARRINGTON**, *Senior Program Assistant.*

For More Information . . . This Report Highlights was prepared by the Board on Testing and Assessment based on the report *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress* (2017). The study was sponsored by the U.S. Department of Education. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of any organization or agency that provided support for the project. Copies of the report are available from the National Academies Press, (800) 624-6242; http://www.nap.edu or via the BOTA page at http://nas.edu/NAEP-AchievementLevels.

Division of Behavioral and Social Sciences and Education

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org