MAY 2012

NAEP:
LOOKING AHEAD

LEADING ASSESSMENT INTO THE FUTURE

Recommendations to the Commissioner
National Center for Education Statistics

NAEP: Looking Ahead
Leading Assessment into the Future

NCES Initiative on the Future of NAEP

The National Assessment of Educational Progress (NAEP) has undergone a series of notable changes in the past decade. The NAEP program has expanded to meet new demands. All 50 states, the District of Columbia, the Department of Defense schools, and (on a trial basis) 21 urban districts are now participating in the mathematics and reading assessments at grades 4 and 8. In addition, thirteen states are participating in trial state 12th-grade assessments in reading and mathematics. NAEP is also reporting in record time to ensure that the findings are highly relevant upon release. Technology has taken on a bigger role in the development and administration of NAEP, including computer-based tasks in the science and writing assessments. These are just a few of the major developments; the program has grown and matured in almost all respects.

There is also growing interest in linking NAEP to international assessments so that NAEP scores can also show how our nation's students measure up to their peers globally. Additionally, there is increasing interest in broadening assessments in the subject areas to incorporate college and career readiness, as well as what are often called "21st-century skills" (communication, collaboration, and problem-solving).

The National Center for Education Statistics (NCES), which administers NAEP, is dedicated to moving the program forward with its upcoming procurement cycle which will take the program to 2017. Under the leadership of NCES Commissioner Jack Buckley, NCES convened a diverse group of experts in assessment, measurement, and technology for a summit in August 2011. These experts discussed and debated ideas for the future of NAEP. NCES convened a second summit of state and local stakeholders in January 2012. Participants at both gatherings were encouraged to "think big" about the role that NAEP should play in the decades ahead.

NCES assembled a panel of experts from the first summit, chaired by Edward Haertel, an expert in educational assessment, to consider and further develop the ideas from the two discussions and make recommendations on the role of NAEP in the future—10 years ahead and beyond. Based on summit deliberations and their own extensive expertise, the panel developed a high-level vision for the future of the NAEP program, as well as a plan for moving toward that vision.

This paper contains the panel's recommendations to the NCES Commissioner. NCES will consider these recommendations in their mid- and long-range planning for the program.

Panel Members

Edward Haertel (chair)
Jacks Family Professor of Education
School of Education
Stanford University

Russell Beauregard
Research Scientist & Director of Design
Education Market Platforms Group
Intel Corporation

Jere Confrey
Joseph D. Moore Distinguished University Professor
College of Education
North Carolina State University

Louis Gomez
MacArthur Chair in Digital Media and Learning
Graduate School of Education and Information Sciences
University of California, Los Angeles

Brian Gong
Executive Director
National Center for the Improvement of Educational Assessment

Andrew Ho
Assistant Professor
Graduate School of Education
Harvard University

Paul Horwitz
Senior Scientist and Director
Concord Consortium Modeling Center
Concord Consortium

Brian Junker
Professor
Department of Statistics
Carnegie Mellon University

Roy Pea
David Jacks Professor of Education
School of Education
Stanford University

Robert Rothman
Senior Fellow
Alliance for Excellent Education

Lorrie Shepard
Dean and Distinguished Professor
School of Education
University of Colorado, Boulder

1. The Landscape of National Assessment

For more than four decades, educators, parents, policymakers, researchers, and the general public have been well served by reports from the National Assessment of Educational Progress (NAEP), also known as The Nation's Report Card. NAEP has provided the best available information about the academic achievement of the nation's students in relation to consensus assessment frameworks, maintaining long-term trend lines reaching back over this entire span of time. In addition to reporting at the national level, NAEP has offered achievement comparisons among participating states for more than two decades, and since 2003, all states have participated in the NAEP mathematics and reading assessments at the fourth and eighth grades. More recently, NAEP has also reported achievement for selected large urban school districts.

In addition to characterizing the achievement of fourth-, eighth-, and twelfth-grade students in a variety of subject areas, NAEP has also served to document the often substantial disparities in achievement across demographic groups, tracking both achievement and achievement gaps over time. Because no rewards or sanctions are directly tied to NAEP performance, and because NAEP reporting does not reach below the state or district level, NAEP is a "low-stakes" assessment. This has made it uniquely valuable as a point of reference when interpreting the score gains typically seen on the high-stakes tests used directly as tools for educational reform.

In addition to describing educational achievement, NAEP has furthered deliberation as to the scope and meaning of achievement in mathematics, reading, and other subject areas. NAEP assessments are aligned to ambitious assessment frameworks developed by a thoughtful process to reflect the best thinking of educators and content specialists. These frameworks have served as models for the states and other organizations to follow. Finally, NAEP has also served as a laboratory for innovation, developing and demonstrating new item formats, as well as statistical methods and models now emulated by large-scale assessments worldwide.

1.1 A Changing Environment, More Ambitious Expectations

As we look to the future, NAEP will be called upon to do all that it has historically done and more. We see at least four major trends to which NAEP must respond. First, NAEP must provide value as a nationally representative assessment when it is likely that other assessments will also provide information about student achievement that may be aggregated and compared across districts, states, and even at the national level. Forty-six states and the District of Columbia have adopted the new Common Core State Standards (CCSS) in English language arts (ELA) and mathematics. Two federally funded state consortia are developing assessments aligned with the CCSS for general education students in grades 3-8 and high school—the SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Two more state consortia are developing ELA and mathematics assessments linked to the CCSS for students with severe cognitive disabilities – the Dynamic Learning Maps Assessment Consortium (DLM) and the National Center and State Collaborative Assessment Consortium (NCSC). Yet another, the World-Class Instructional Design and Assessment (WIDA) consortium, is developing English language proficiency assessments. It is not entirely clear how NAEP's role may change with the advent of these new assessments. However, we can anticipate that many of these consortium tests will become "high-stakes" as they are used for accountability purposes. In response, educators will shift their focus toward preparation for these new accountability tests. If NAEP remains a low-stakes assessment program aligned to frameworks that reach beyond the confines of the CCSS, then it will be well positioned to provide uniquely valuable information about the extent to which other learning is maintained or declines as curriculum and instruction evolve toward the CCSS. History suggests that even for ELA and mathematics content included in the CCSS, achievement trends shown on NAEP will likely differ from those seen on the high-stakes tests themselves.

Second, NAEP will be called upon to assess a broader set of learning outcomes. Our educational system is challenged as never before to prepare our

students for a changing world. With increasing urgency, educators, policymakers, and business leaders are demanding that schools equip students with a broader range of skills, including the ability to evaluate critically the trustworthiness of information from different sources. Changing workplaces will require fluency with new, technology-based knowledge representations. As a nation, we are pressing toward the goal of readying all students for college or careers by the time they exit high school, striving for universal attainment of achievement levels historically reserved for the "academic track." The CCSS address some of these competencies. There is also increasing attention to a range of non-cognitive attitudes and dispositions that are critical to bringing students into fuller partnership with the act of learning and preparing them to collaborate effectively. Many of these new aspirations are captured by the phrase "21st-century skills." A precise definition of that term is elusive, but would include proficiencies and dispositions that cut across subject domains such as research, innovation, communication, and persistence after failure; and might also include group collaboration and problem-solving skills. These cross-cutting skills are best assessed as part of subject-matter assessments, not as separate, stand-alone domains. NAEP can provide leadership in constructing carefully considered assessment frameworks, informed by the best available expertise, that include these new proficiencies, specify their relationships with more traditional academic areas, and show how they can be assessed.

Third, NAEP will be called upon more than ever to serve as a leader in assessment innovation as new technologies become available for administering, scoring, and reporting assessment results. Hand-held and tablet devices are already ubiquitous in schools, and in a few more years, the traditional paper-and-pencil testing platform may be all but obsolete. The arrival of these new technologies coincides with new demands for innovative item formats, assessment designs, and psychometric models that can be used in NAEP and also inform large-scale assessment programs worldwide.

Fourth, sharing of both data and instructional resources on a very broad scale, across states and perhaps even across nations, will pose new opportunities and new challenges as NAEP evolves in response to a changing assessment data infrastructure. NAEP can be a leader in providing information not only from its own assessments, but in showing the value of merging information across multiple sources. NAEP could assume a role in helping stakeholders interpret achievement data in the context of school information from longitudinal surveys, or in providing explicit linkages between state, national, and international assessment data. We anticipate that NAEP may be called upon to develop new models for sampling students and conducting assessments to capture academic learning in out-of-school settings, as the popularity of home schooling and online learning increases. In these ways, NAEP may serve as a source of information not addressed by any other nationally representative assessment, and provide new methods and ideas for the measurement field.

For all these reasons, it is a propitious time for NCES to undertake this review of the future of NAEP. We hope that our reflections here will be of value in guiding NAEP toward the middle of the 21st century.

1.2 Organization of this report

The body of this report is divided into four sections, numbered two through five, with a final section serving as the summary and conclusion. Section two, "NAEP as The Nation's Report Card," takes up the structure and organization of the assessment itself. This includes the workflow organization through which assessment frameworks are developed, items are written, samples are drawn, tests are administered, analyses are run, and reports are written and disseminated. It also includes the basic structure of NAEP—what units are sampled and how the assessment data collection is organized. In recognition of the historic value of various special studies conducted as part of NAEP, this section highlights the value of an "Innovations Laboratory" expanding upon this traditional NAEP function.

Section three, "NAEP's Assessment Frameworks and Learning Outcomes," takes up the questions of what content and skills NAEP should cover, and of how decisions about NAEP coverage should be made. Within the overall structure of periodic assessments focused on academic subject areas, there is much latitude for changes in the specific knowledge and skills measured. Over time, for

example, there may be increased focus on obtaining, evaluating, and using information, with relatively less emphasis on factual or procedural knowledge. This section also takes up the implications of the CCSS for NAEP assessment frameworks in mathematics and reading.

The implications of emerging technologies, including tablet and hand-held devices, are discussed throughout the report, but these topics are treated most extensively in section four, "NAEP and New Technologies." New technology is changing what is learned, where it is learned, and how it is learned, as well as changing how learning can be assessed. While we anticipate that NAEP will continue to rely primarily upon "stand-alone" assessments – tasks given to students specifically for the purpose of assessment – new technology platforms will enable expanded use of items that rely on more than brief, static texts to frame and pose questions. Interactive technology may enable adaptive scaffolding to engage students in more complex tasks, posing successive questions or offering hints contingent upon students' previous responses, as well as targeting some portions of the assessment to just those students who have had experience with relevant learning technology. All in all, with the demise of paper-and-pencil tests, we will also see diminished reliance on small, independent, self-contained items each yielding only one or a few bits of information. The section on emerging technologies also discusses the potential for fuller inclusion of students with disabilities as well as English language learners, capitalizing on the greater flexibility of digital platforms to present information and record responses in ways adapted to individual student needs. Section four also addresses the potential of new education data systems to connect NAEP with greatly expanded sources of achievement information.

Finally, section five takes up the matter of "NAEP Reporting and Use." NAEP findings are released in several forms for different audiences. When a major report is released, press packets help to shape media coverage. Separate reports document detailed findings for participating states and districts. The web-based NAEP Data Explorer application enables anyone to construct customized tabulations. NAEP also makes special, restricted-use data files available to qualified researchers for further analysis. Apart from achievement per se, technical reports document NAEP's methodology and findings concerning new item types and other innovations. For more than two decades, and especially since the passage of the No Child Left Behind Act of 2001 (NCLB), NAEP findings have been presented and interpreted primarily with respect to achievement levels defining Basic, Proficient, and Advanced performance. The "Reporting" section examines the strengths and weaknesses of this approach and proposes alternatives. Additional specific reporting recommendations are also presented, including greater reliance on technology-based interactive tools.

1.3 Notes of Caution

Going forward, we expect that NAEP will continue to serve as the most authoritative source of information concerning patterns and trends in the academic achievement of American youth, and also as a model of excellence and innovation in large-scale assessment. It will continue to serve as a trustworthy, low-stakes benchmark test against which to judge the effectiveness of various large-scale educational reforms. It will also evolve to measure an expanded range of learning outcomes using new technologies. At the same time, it is important that the NAEP program continues to focus on those things it can do well. NAEP reports have been, and will continue to be, primarily descriptive. NAEP is not principally a research program, and its design is not conducive to supporting causal inferences as to "what works" in education.

As much as we would like to see more attention to learning within classrooms, we acknowledge that NAEP seems poorly suited for the large-scale assessment of educational processes. Without fundamental design changes that allow for the tracking of individual students or schools over time, NAEP will continue to provide only a series of snapshots of achievement. In this report, we recommend some exploration of the feasibility of longitudinal linking at the student level, as well as adaptive testing according to students' learning contexts and their patterns of item responses. However, we do not know what conclusion such explorations would reach, and no timeline is

proposed for the design changes these new functionalities would entail.

A continuing challenge as NCES, the National Assessment Governing Board (NAGB), and the NAEP contractors strive to expand the range of learning targets assessed by NAEP is the inherent limitation of using a common set of test questions for students taught using various curricula. Rich content knowledge is the medium through which students' reasoning, problem-solving, and critical thinking are applied and demonstrated. Even after full adoption of the CCSS, it will probably not be realistic to expect all students to have read the same novel or Shakespearean play, or to have studied the same biological organisms or ecosystems. If students could be assumed to have covered the same specific content, then test questions could challenge them to work with that content to demonstrate their reasoning or problem-solving ability. However, to the extent that students across the nation have studied different things, test questions must instead be self-contained. Each item must assume whatever content knowledge it requires or else package that content within the test question itself along with any accompanying text, pictures, or other displays of information. While it is certainly valuable to be able to assimilate and work with content quickly, that is not the same as working with complex knowledge one has already mastered. There may be no completely satisfactory resolution to this dilemma although we do offer some suggestions for further exploration. It may well be that, given the organization of curriculum and instruction in U.S. schools, NAEP assessments simply cannot be strongly connected to specific antecedent instruction. If this is so, then it may be that some valued forms of complex learning simply cannot be assessed in a nationally representative survey program.

2. NAEP as The Nation's Report Card

2.1 Overview

Ongoing production work in NAEP—item development, assessment design, school and student sampling, test booklet production and distribution, data collection, scoring, analysis and reporting, and software development for computer-based testing—is conducted by an alliance of contractors. This complex enterprise is supplemented by a set of research and development (R&D) "special studies," conducted and closely coordinated by the alliance, involving such topics as new item development, assessment design, and data collection.

NAEP has functioned well as a suite of complex survey modules conducted as assessments of student achievement in fixed testing windows, using largely paper-and-pencil testing methods. The complexity of NAEP evolved by necessity, to address its legal and policy reporting requirements and the complex sampling of items and students needed to make reliable and valid inferences at the subgroup, district, state, and national level for stakeholders, ranging from policymakers to secondary analysts, without creating an undue burden on students and schools.

This same complexity, however, makes NAEP a very difficult ship to turn. The development of new assessment frameworks takes years and standard setting for interpreting student performance is under heavy scrutiny. Seemingly innocuous changes in the underlying survey and psychometric models can take years to understand and validate, and more years before they become part of NAEP operations. To some extent this is understandable, given the many stakeholders that NAEP serves, and the need for the results for one stakeholder to be consistent with those reported to another. However, NAEP is entering an era in which it must become more nimble in order to maintain its role as the preeminent source of inferences about what students know and can do.

The national assessment ecosystem has grown up around NAEP, not as a consequence of it. Rather, that ecosystem is the result of other standards and accountability efforts at the state and federal levels. Notable among these efforts are NCLB, which established a new role for NAEP as a benchmark for mandated state accountability assessments, and the Race to the Top competition funded by the American Recovery and Reinvestment Act of 2009, which underwrote state consortia now developing elaborate accountability assessment programs aligned to the CCSS. In addition, there is a growing array of tools that can be used to learn what students know and can do, including innovations in test item and task types (often technology-enabled) to assess "21st-century skills" and new statistical and psychometric methodologies such as educational data mining and learning analytics. The panel strongly endorses NAEP's continued and essential involvement in this ecosystem.

In order to thrive within and contribute to this environment, NAEP will need to maintain a distinct, well-defined, and useful role. Indeed, there is general agreement that NAEP needs to be updated and expanded, not overhauled or abandoned. In considering NAEP's evolving role, it is valuable to consider NAEP's current structure. In this section we review that current structure, and we recommend a review of NAEP's R&D processes and the creation of an "Innovations Laboratory" in which to organize NAEP-related R&D.

2.2 Basic Assessment Structure

NAEP currently consists of four separate survey modules—national NAEP, state NAEP, NAEP Trial Urban District Assessment (TUDA) and the long-term trend (LTT) survey—as well as occasional special studies (which will be discussed separately below). National NAEP assesses many subjects, including mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history, following assessment frameworks developed by NAGB. Nearly all subjects are assessed at grades 4, 8, and 12. Mathematics and reading are assessed every two years, and other subjects are assessed less frequently. Additionally, not all grades are assessed each time. State NAEP

began in 1990 and reports results in four subjects—mathematics, reading, science, and writing—usually only for grades 4 and 8. When possible, the sample for national NAEP is a subset of the state NAEP sample in participating states. NAEP TUDA is a multiyear study of the feasibility of a district-level NAEP in selected urban districts that is supported by federal appropriations authorized under NCLB. The first TUDA took place in 2002, with six urban districts participating in the reading and writing assessments. TUDA has taken place in every odd-numbered year since then, using a representative sample of schools and students in each district that can also be included in state and national NAEP results. The LTT assessments are given at the national level only, and use different target populations and sampling frames than the other NAEP survey modules. LTT is conducted every 4 years, and reports trend results in mathematics and reading going back to the early 1970s. For all NAEP assessments, test items are presented in matrix-sampled blocks to reduce student testing time; consequently scores are not reliable or valid at the individual or school level and are not reported. Aggregated results at the state, regional and national levels are available through primary reports published by NCES, as well as the *NAEP Data Explorer* and other customizable reporting tools available online.[1] Disaggregated data are available to secondary researchers under a restricted-use data agreement with NCES.

The modular structure of NAEP is effective in reporting on populations in different jurisdictional units with both cross-sectional and trend components, and should remain intact. Trend results are reported in two ways by comparing results in different years: on the cross-sectional National, State, and TUDA surveys, and in the LTT results. As reported by Beaton and Chromy (2010), these trends are generally similar, but sometimes differ in meaningful ways. All NAEP assessments aim to get as close to 100% participation as possible, via supports for English language learners and accommodations for students with disabilities consistent with their regular school instruction. As is discussed in Section 5.4, there has been research on extrapolations from results for students assessed to the full population. In addition, motivation and engagement have been persistent concerns, especially for 12th-grade NAEP test-takers.

The panel has discussed several refinements of this basic design – such as consolidating trend data/reports, reporting results at the individual or school level, and using technology to improve the current matrix sampled item block design. It does not take a strong position, but does recommend consideration of ways to consolidate or combine long-term trend and main NAEP data collections. The long-term trend is a valuable component of NAEP. By maintaining alternative definitions of subject matter achievement, it enriches interpretations of main NAEP trends. There may be ways to preserve the long-term trend lines while reducing costs, simplifying the structure of the NAEP program, and at the same time improving the accuracy of some important kinds of long-term trend interpretations.

With regard to frameworks, one avenue that may be worth exploring is a process for more frequent updates to assessment frameworks and the items representing those frameworks. Changes would be deliberate and incremental, designed to keep NAEP up to date with educational policy and research. This would be somewhat similar to how updates are made to the market basket used for the Bureau of Labor Statistics' Consumer Price Index (CPI). Rather than expending effort maintaining a deeply psychometrically valid link across years, the CPI is based on a gradually changing set of goods and services that is updated to reflect current spending habits of U.S. consumers. Similarly, one might consider periodic updates to a set of tasks sampling an evolving mix of knowledge and skills that reflects current educational policy and research about what is important that students know and can do (discussed further in Section 3.2.3).

The panel does not believe that redesigning NAEP to provide results at the individual or school level is a worthwhile endeavor. A primary difficulty would be dramatically increased complexity and cost for the survey. In addition, individual scores are already provided by state assessments. NAEP can and should maintain its role as an independent benchmark for state assessments and other policy initiatives. Any move toward further disaggregation in NAEP reporting would be likely to increase the perceived stakes attached to NAEP assessments,

[1] http://nces.ed.gov/nationsreportcard/about/naeptools.asp

potentially compromising its function as a neutral audit/benchmark.

The possibilities for technology-based improvements in NAEP are many, especially since NAGB has determined that NAEP mathematics and reading will both be online assessments as of 2017. One particularly low-hanging fruit is replacing static test item blocks in the current matrix-sampled design with a branching structure, so that the second block of items can be tailored to student proficiency as estimated from the first block. This may improve precision of measurement generally, and it should make it possible to reach farther above or below grade level than the current NAEP can do. This two-stage adaptive testing scheme was piloted in NAEP's 2011 Mathematics Computer Based Study (MCBS) at grade 8. The MCBS was conducted in part in anticipation of the NAGB 2017 date for online assessment, to inform the relative merits of static versus adaptive models for online assessment. An adaptive model has the potential to reduce measurement error, especially away from the middle of the proficiency distribution, and to improve student engagement by administering items better tailored to students' individual ability levels.

2.3 Innovations Laboratory

2.3.1. Introduction

The evolution of NAEP over the past four decades has been informed by many R&D efforts, undertaken by various individuals or agencies at different times. Some were necessitated by unforeseen technical challenges, such as the adoption of item response theory (IRT) estimation to a matrix-sampled assessment design, or the 1986 Reading anomaly. Others were prompted by new policy questions or proposed score interpretations, including some early uses of performance assessment in NAEP, a pilot study offering students a choice of passages to read, or a study using hierarchical linear modeling (HLM) to examine charter school performance. Many were undertaken simply to improve the accuracy, speed, and efficiency of the NAEP assessment development, data collection, scoring, and reporting.

Two special studies currently underway aim to understand what is feasible by way of technology-based improvements for NAEP. The MCBS has already been mentioned. Another, the Knowledge and Skills Appropriate (KaSA) mathematics special study, considers blocks of new items specifically developed at each grade to better measure the knowledge and skills of lower ability students while still conforming to the content distribution specifications of the NAEP mathematics framework. And the NAEP Writing assessment is already computer-based as an operational assessment, as of 2011.

More broadly NAEP conducts special studies in many areas, to investigate new modes of assessment, provide context to NAEP achievement results, and link NAEP to other assessments. Current and recently active special studies include the Achievement Gaps Studies, the Charter School Pilot Study, the High School Transcript Study, Mapping State Proficiency Standards, Measuring Status and Change in NAEP Inclusion Rates of Students with Disabilities, the National Indian Education Study, the Oral Reading Study, Student Achievement in Private Schools, and the Technology-Based Assessment Project.

It is not only the NAEP program itself that has benefited from these efforts. To cite just three examples, NAEP R&D has informed the design and reporting of large-scale assessments around the world; the mandate for achievement level reporting in NAEP has prompted an enormous investment in standard setting methodology; and analysis and reporting tools developed to make NAEP exercises and NAEP findings more broadly accessible are serving as models for other programs. In short, NAEP is of value not only as a trusted source of information about student achievement but also as a model of state-of-the-art assessment technology and a source for psychometric innovation and research of the highest quality.

2.3.2. Scope of NAEP research and evaluation

Perhaps the best known technical innovation originating with NAEP was the introduction of the matrix-sampled Balanced Incomplete Block (BIB-spiral) design and its associated multiple-imputation (also known as plausible values) methodology, which enables IRT estimation of group achievement distributions with sparse data at the level of individual respondents. This model has been emulated in numerous large-scale assessments. But numerous smaller, less well known efforts have also been of value both within and beyond the NAEP program. As is discussed in Section 5.4, research on full population estimates in the context of NAEP has clarified the magnitude of estimation bias due to exclusions and accommodations, and ways of reducing that bias. The NAEP Validity Studies Panel has conducted a study comparing trends using main NAEP versus long-term trend NAEP (Beaton & Chromy, 2010); a feasibility study of two-stage testing in NAEP (MCBS); investigations of the cognitive processes engaged by NAEP items or relating NAEP scores to other performance measures; studies of linking or equating between NAEP and other assessments; and studies of the validity of accommodations, among other topics.[2] Other investigations have probed the reasons for non-response on NAEP, the effect of monetary rewards on performance, and even the degree to which policymakers and educators are able to understand NAEP executive summary reports.

NAEP has spurred research by other groups and individuals, as well. A keyword search of National Center for Research on Evaluation, Standards, and Student Testing (CRESST) reports at http://www.cse.ucla.edu/products/reports.php turns up dozens of examples. A visit to the NCES website (http://nces.ed.gov) turns up still more examples, such as the NAEP-TIMSS linking study, research relating states' proficiency standards to the NAEP scale, and a report on computer-based writing assessment. A search with the keyword "NAEP" at the ETS Research Reports website (http://www.ets.org/research/policy_research_reports/ets) brings up a list of over 200 reports on such topics as bias in weighted random effects model estimators, evaluation of methods to compute complex sample standard errors in latent regression models, and effects of administration mode (computer-based versus paper-and-pencil) on eighth-grade students' mathematics test performance. Finally, some independent research has been undertaken in the context of external evaluations of NAEP, as reported in the National Research Council's volume *Grading the Nation's Report Card* (1999) and earlier evaluations carried out by the National Academy of Education (Glaser, Linn, & Bohrnstedt, 1992, 1993, 1994, 1996, 1997).

A longstanding small grants program supporting secondary analyses of NAEP data was discontinued after 2008, but the Institute of Education Sciences (IES) program on Statistical and Research Methodology in Education continues to support work on new item response models and new approaches to marginal likelihood estimation, for example. Special studies involving new item development, assessment design, and data collection are undertaken as needed, usually by NAEP contractors. NCES tries to set NAEP research priorities formally about once every five years, and some study proposals are evaluated against these priorities.

The panel recommends a systematic survey of the organizational structures through which NAEP research has been carried out, including the NAEP Validity Studies Panel, the IES program on Statistical and Research Methodology in Education, the NAEP Design Analysis and Reporting contractor, the Education Statistics Support Institute Network (ESSIN), the discontinued NAEP Secondary Analysis Grants program, and perhaps other mechanisms not listed here. The survey would not aim to classify or evaluate research products. Its primary goal would be to examine the range of mechanisms for initiating projects, evaluating and funding proposals, reviewing and disseminating findings, and otherwise managing NAEP R&D, with an eye toward identifying gaps and redundancies.

[2] Visit http://www.air.org/reports-products/index.cfm?fa=viewContent&content_id=890 for a list of NAEP Validity Studies Panel reports since 1995.

2.3.3 Proposal for NAEP Innovations Laboratory

The panel envisions an "Innovations Laboratory" with an expanded assessment R&D budget, under which a portfolio of research studies would be managed. The Innovations Laboratory would support studies in the tradition of past NAEP R&D, but would also include innovative research as yet unimagined, but essential to keep NAEP at the forefront of innovation and best practice. It would serve as a point of access for vetting new ideas from different sources, and would support both in-house and third-party studies. In addition to generating new knowledge for the improvement of NAEP itself, the Innovations Laboratory would serve as a hub for dissemination of technical innovations, including new statistical methods, technological advances, innovative item types, and more.

Various organizations of Innovations Laboratory initiatives would be possible, but four categories illustrating a range of possibilities might be: (1) investigating and assuring the validity of intended inferences from NAEP; (2) improving NAEP processes to balance among reducing respondent burden, shortening reporting time, increasing precision, and reducing costs; (3) expanding the range of achievement constructs NAEP can validly assess; and (4) enabling NAEP to serve new purposes.

The first of these four, concerning validity, is perhaps most important, both historically and for the future. Among other topics, validity studies might comprise research on content framework development and test specifications, item development, item response processes (as with cognitive labs), student motivation, effects of accommodations for English language learners and students with disabilities, sampling weights and non-response bias, estimation methodology, calculation and reporting of standard errors, and the clarity of NAEP reports.

The second of these categories, improving NAEP processes, would encompass many of the more technical investigations carried out by the NAEP contractors. Examples of past work in this category include the integration of national and state samples, refinements to sampling designs to incorporate available achievement data within states, the standardization of NAEP booklet designs across subjects to enable more efficient simultaneous administrations, and pre-equating using field test data to shorten the time between administration and reporting.

Work in the third category, expanding the range of constructs NAEP can assess, would include investigations of new item types, as well as new models for targeted sampling of examinees. Computer-based test administration opens the door to branching items or testlets to probe students' reasoning in ways a single paper-and-pencil item cannot. It facilitates scaffolding to determine what students can accomplish given additional support. Adaptive testing may also make it possible to administer more advanced or more complex items or tasks to just the subset of students for whom they are appropriate, perhaps as a function of specific instructional technologies they have used or other features of their prior instruction. Computer-based test administration expands the possibilities for testing accommodations. In addition, it enables new ways of capturing students' performance, including the sequence of moves in complex problem spaces as well as response latencies.

The fourth category, enabling NAEP to serve new purposes, would include work on linkages between NAEP and other large-scale assessments. The "audit" function of NAEP was not envisioned by the original architects of NAEP, but emerged as high-stakes testing evolved into an instrument of educational policy. Linking score scales on NAEP to those from other assessments might enable much more refined investigations of score inflation, capitalizing on changes in linking functions over time. It might also include exploring ways to provide context for NAEP results. In the 1980s,"Opportunity to Learn" was a revolutionary idea that provided an important new lens through which to examine achievement results. Today it may be possible, through mechanisms similar to the NAEP contextual variables, to provide some general educational background or curricular context for NAEP results.

The panel would like to emphasize that a valuable function of the Innovations Laboratory will be in identifying new research priorities and envisioning possibilities proactively

3. NAEP's Assessment Frameworks and Learning Outcomes

3.1 Background and History

Assessment frameworks are conceptual, overview documents that lay out the basic structure and content of a domain of knowledge and thereby serve as a blueprint for assessment development. Typically, assessment frameworks, for NAEP and for other large-scale assessments, are constructed as two-dimensional matrices of content strands and cognitive processes. For example, the current NAEP mathematics framework includes five content areas: number properties and operations; measurement; geometry; algebra; and data analysis, statistics and probability. These are assessed at different levels of cognitive complexity, which include mathematical abilities such as conceptual understanding, procedural knowledge, and problem-solving. In geography, the content areas include: space and Earth places; environment and society; and spatial dynamics and connections. The levels of the cognitive dimension consist of knowing, understanding, and applying.

NAEP Assessment Frameworks are developed under the auspices of the Governing Board through an extensive process involving subject matter experts, who consider how research in the discipline and curricular reforms may have shifted the conceptualization of proficiency in a given knowledge domain. The development process also requires multiple rounds of reviews by educators, policy leaders, members of the public, and scholars. It is expected that assessment frameworks will need to be changed over time. However, the decision to develop new frameworks is approached with great caution because measuring change requires holding the instrument constant. Introducing new frameworks—while providing a more valid basis for the assessment—could threaten one core purpose of NAEP, which is to monitor "progress." In the past, when relatively minor changes have been made in assessment frameworks, as judged by content experts, trend comparisons over time have been continued and bridge validity studies have been conducted to verify that conclusions about gains have not been conflated with changes in the measuring instrument or redefinition of the construct being assessed.

When more profound changes occur in the conceptualization of an achievement domain, then a new framework is essential, and correspondingly the beginning of a new trend line. The adoption by nearly all states of the CCSS in English language arts and literacy and mathematics and the new Science Education Framework developed by the National Research Council (NRC) could be the occasion for a substantial enough change in conceptualization of these domains that new NAEP frameworks and new trend comparisons are warranted. Still, the future of NAEP—as a statistical indicator and as an exemplar of leading-edge assessment technology—requires great care and attention to the implications of new trend comparisons rather than merely acceding to the hoopla surrounding the new standards.

In the history of NAEP, few changes have been made in the assessment frameworks for reading and for mathematics. The old frameworks in these two core subjects, begun in 1971 and 1973 respectively, were replaced in the early 1990s, and then again in 2009 for reading. The old assessments have been continued on a less frequent cycle and are referred to as long-term trend NAEP. The 1990's mathematics framework and 2009 reading framework guide the present-day assessments, referred to as main NAEP. While NCES has been careful to insist that the old and new frameworks measure different things and therefore cannot be compared, the existence of the two trends provides a critically important example to illustrate how changing the measure can change interpretations about educational progress (e.g., see Beaton & Chromy, 2010). The earlier assessments focused much more on basic skills. Reading passages were generally shorter compared to today's NAEP and did not require students to demonstrate so wide a range of reading skills or answer extended-response questions. In mathematics, long-term trend NAEP had a greater proportion of computational questions and items asking for recall of definitions, and no problems where students had to show or explain their work. In a 2003 study, researcher Tom Loveless complained that the new NAEP mathematics assessment exaggerated progress in mathematics during the 1990s because gains on the basic skills test over the same period were much

smaller (when compared in standard deviation units of the respective tests). Because the two assessments are administered entirely separately, Loveless then had to rely on comparisons based on the less than satisfactory item-percent-correct metric to try to track progress in subdomains of the test. A more recent study using more sophisticated methods has largely confirmed his general conclusions, but that same study has highlighted the technical challenges of comparing trends for two assessments administered under such different conditions (Beaton & Chromy, 2010).

3.2 New Approaches for Assessment Frameworks

3.2.1 Designing frameworks and assessments to evaluate directly the effects of changing domain definitions

NAEP cannot be a research program and in particular cannot be structured to investigate the effectiveness of various instructional interventions. However, it can and should be attentive to the ways that shifting definitions of subject matter competence can affect claims about progress or lack of progress (cf. Section 3.2.3). In the CCSS context, it will be especially important to pay attention directly to potential differences between consortium-based conclusions and NAEP trends. Taking this on as a role for NAEP continues its important function as a kind of monitoring instrument. For example, when some state assessment results have shown remarkable achievement gains and closing of achievement gaps, achievement trends for the same states on NAEP have helped to identify inflated claims. These disparities might exist because of teaching-the-test practices on state tests (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998), state content or achievement standards that do not rise to NAEP levels (Bandeira de Mello, Blankenship, & McLaughlin, 2009), exclusion of low-performing students on NAEP, or lower motivation on NAEP. More direct linking by carefully accounting for the consortium frameworks within new NAEP frameworks, would allow NAEP to act somewhat like an external monitor for CCSS assessment results. While the current NAEP frameworks do cover many of the same skills as the CCSS, they can be enhanced with some shifts in content.

"21st-century skills" aren't actually new in this century, but it is a relatively new idea (beginning in the 1990s) that these reasoning skills should be more broadly attained and expected of all students. More importantly, it is indeed new that policy leaders would move toward a view of learning that calls for reasoning and explaining one's thinking from the earliest grades, in contrast to outmoded theories of learning predominant in the 20th century that postponed thinking until after the "basics" had been mastered by rote. In addition, the CCSS firmly ground reasoning, problem-solving, and modeling in relation to specific content, not as nebulous generalized abilities. While there is widespread enthusiasm for designing new assessments that capture these more rigorous learning goals, we should note that promises like this have been made before. In the case of the current NAEP mathematics assessment, item developers acknowledge that the proportion of high complexity items actually surviving to the operational assessment is much smaller than is called for in the NAEP Mathematics Framework, and a validity study at both grades 4 and 8 found that the representation of high-complexity problems was seriously inadequate at grade 8, especially in the Algebra and Measurement strands (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007).

Good intentions to measure "higher order thinking skills" are often undermined for three interrelated reasons. First, test questions at higher levels of cognitive complexity are inherently more difficult to develop. Because the dimensions of the task are intended to be ill-specified, such problems are often perceived to be ambiguous. But as soon as the item developer provides clarifying parameters, the challenge of the problem is diminished. Second, because "21st-century skills" involve applying one's knowledge in real world contexts, prior experience with particular contexts (or lack thereof) can create very large differences in performance simply because students unfamiliar with the context are unable to demonstrate the intended content and reasoning skills. In fact, application or generalization can only be defined in relation to what is known to have been taught. This is the curriculum problem that haunts large-scale assessments like NAEP that seek to be curriculum independent. Finally, well

designed items can fail on statistical criteria if too few students can do them.

These are all cautionary tales. They do not imply that NAEP should be less ambitious in developing new assessment frameworks that reach as far as possible in representing these higher levels of subject matter proficiency. But they do suggest a hedging-one's-bets approach that does not discard old frameworks wholesale in favor of the new. Rather, as mentioned previously, some conscious combination of old and new would create an assessment better equipped to track progress over time. Later we discuss Innovations Laboratory studies like those NAEP has used historically to explore the feasibility of new assessment strategies. However, we should emphasize that studies of innovative assessment strategies that tap complex skills should not merely be new assessment formats administered to random samples of students. Rather, in recognition of the fact that opportunities to learn particular content and skills may affect whether an assessment looks psychometrically sound, studies should be undertaken with carefully selected populations where relevant opportunities to learn can be established. This will help determine whether more advanced performance can be accurately documented to exist within the parameters of the new standards.

3.2.2 Standing subject-matter panels

To aid in this process, provide substantive oversight, and ensure meaningful interpretation of trends, we elaborate a recommendation for the future of NAEP previously made by a National Academy of Education Panel, which called for standing subject-matter committees. We recommend an expanded role whereby standing committees of subject matter specialists would review field test data, for example, and call attention to instances when after-the-fact distortions of the intended domain occur because more ambitious item types fail to meet statistical criteria. These committees would also have a role in ongoing incremental updates to content frameworks. They might include at least one member with psychometric expertise to aid in formulating technical specifications. The role of these committees is further described in Section 6.1.3.

3.2.3 Dynamic assessment frameworks and reporting scales

As just explained in Section 3.1, NAEP assessment frameworks have historically been held fixed for a period of years and then changed. It might be added that historically, NAEP item pools have been constructed according to test specifications derived from assessment frameworks. NAEP reporting scales, in turn, have reflected the resulting mix of NAEP items. Periodic small revisions to assessment frameworks have been made while maintaining trend lines; major breaks requiring new trend lines have occurred only rarely. With standing subject-matter panels, assessment frameworks for each subject-grade combination might be adjusted more frequently, defining a gradually changing mix of knowledge and skills, analogous to the Consumer Price Index (cf. Section 5.3). At the same time, item pools might be expanded somewhat, including everything in the assessment framework but also covering some additional material. Assessment frameworks would still define the intended construct underlying NAEP reporting scales, but not all items in the NAEP exercise pool would be included in the NAEP reporting scales. For example, content required to maintain long-term trend NAEP, to assure sufficient representation of the CCSS, or to improve the linkage to some other assessment could be introduced into the pool without affecting NAEP reporting scales. With somewhat broader exercise pools, alternative construct definitions could be investigated in special studies. The panel assumes that broader exercise pools, supporting modestly different construct definitions, will increase the value of NAEP by highlighting distinctions among achievement patterns under different construct definitions. Of course, there would still be one main NAEP reporting scale for each subject/grade combination. Clarity in communicating NAEP findings would remain a priority.

Different assessment frameworks may imply different definitions of the same broad subject area achievement construct (e.g., "reading" or "mathematics"), and achievement trends may differ depending on the construct definition chosen. Incremental changes in assessment frameworks and the corresponding set of items on which NAEP reporting scales were based would afford local (i.e., near-term) continuity in the meaning of those scales, but over a period of decades, constructs

might change substantially. This was seen by the panel as a potential strength, but also a potential risk. Policymakers and the public should be aware of how and when the construct NAEP defines as "reading," for example, is changed. Not every small, incremental change would need to be announced, but it would be important to establish and to enforce clear policies concerning the reporting of significant changes in assessment frameworks, so as to alert stakeholders when constructs change and to reinforce the crucially important message that not all tests with the same broad content label are measuring the same thing. As small content framework adjustments accumulate over time, standing committees, using empirical studies, would need to determine when the constructs measured have changed enough to require establishing new trend lines.

Dynamic frameworks would balance dual priorities of trend integrity and trend relevance. As an analogy, the Consumer Price Index (CPI) tracks inflation by deliberately conflating two concepts: change in the cost of a fixed basket of goods and change in the composition of the basket itself. As time passes, an increase in the cost of a product that is no longer relevant should contribute less to estimated inflation. By adopting dynamic frameworks, NAEP would similarly conflate increases in student proficiency with a change in the definition of proficiency itself. Although this conflation may seem undesirable, it may be the best way to balance desires for both an interpretable trend and a relevant trend.

3.2.4 Learning progressions as possible guides to assessment frameworks

Learning progressions or trajectories represent descriptions of how students' knowledge, skills, and beliefs about the domain evolve from naïve conceptions through gradual transformations to reach proficiency with target ideas at high levels of expertise over a period of years (Heritage, 2008). They entail the articulation of intermediate proficiency levels that students are likely to pass through, obstacles and misconceptions, and landmarks, of predictable importance as students' knowledge evolves over time. Empirical study of learning progressions highlights the key roles of instruction, use of tools, and peer interactions in supporting learning. Because the process of evolving understanding can take multiple years, learning progressions bridge formative and summative assessment.

A learning progression can provide much more information than a typical assessment framework. A learning progression ideally specifies both what is to be learned as well as how that learning can take place developmentally over time. It often integrates content and cognition. It includes not only the learning targets but also common less-than-ideal states that many students pass through. It is ordered developmentally. It provides a domain-based interpretation of development or growth that is useful to educators. The 2009 NAEP Science Framework already contains a section on learning progressions; however, learning progressions may offer guidance for the development of future NAEP assessment frameworks, especially in mathematics.

Learning progressions are closely entwined with instructional decisions regarding the sequencing of key concepts and skills. In the Netherlands, for example, the related constructions are referred to as "learning-teaching trajectories." However, few empirically supported "learning progressions" as yet exist, and developing more has proven challenging. In addition, because of NAEP's role as a curriculum-independent monitor, it may be more difficult to develop assessment frameworks that are entirely built as a collection of learning progressions. More likely some particular sequences, if proven to be valid across curricula, could be embedded within more general assessment frameworks.

4. NAEP and New Technologies

4.1 Introduction

Changing technology, especially advances in digital technology, will affect virtually all aspects of NAEP, both directly and indirectly. NAEP will be called upon to take advantage of advances in digital media to assess in new ways those constructs already reflected in NAEP assessment frameworks, and also to assess new constructs, beyond those measured in the past. The challenges regarding how to use technology will become particularly noticeable over the next 10 years. As access to technology increases, we may within that time reach the day when each student carries a device with interactive content rather than textbooks. In this section, we highlight four major areas of innovation and change. Some of these changes have already reached the stage of large-scale pilots in leading countries.[3]

First, new technologies for creating, representing and communicating knowledge underlie new capabilities for teaching, learning, and assessment. Several emergent technologies will be described for representing knowledge as well as new modalities and interfaces through which people can interact with those representations.

Second, technology is transforming both the goals of education and the means of reaching those goals. New tools for knowledge acquisition, representation, and interaction are changing the nature of the workplace, and young people today must be educated for future participation in that changing world. Technology is also supporting new kinds of learning activities. Students are engaging with new kinds of supports for learning, including 24/7 ubiquitous information access, interactive social media, and immersive interactive learning experiences. Some formal learning is already happening outside of school, and where and when, as well as what and how and why learning happens, will likely continue to evolve.

Third, technology is transforming assessment. As we note below, old constructs are being measured in new ways, and new constructs (such as collaboration and computational thinking) are being assessed as well. Most obviously, the use of interactive digital platforms for test administration greatly expands the range of options for presenting questions, for capturing student responses, and for adaptively tailoring questions to students' achievement levels, access needs, or to their learning environments. More profoundly, technology may over time transform the familiar relationship between instructional activities and assessment activities.

With regard to tests themselves, when they are administered using computers or newer digital devices, the potential range of item formats, branching, and immersive nature of test items is vastly expanded. Far beyond mere "electronic page-turners," computer-based testing should make possible items that offer much richer contexts to support reasoning and problem-solving, including multimedia presentations and immersive environments in which examinees can, for example, run simulations, construct data representations, or write and edit, and where prior responses can influence the sequence of questions posed.

Branching items or testlets can scaffold students' performance, providing hints as needed, with that assistance then factored into the scores assigned. With adaptive feedback, students might be guided through multistep problems in a way that enabled recovery from early errors so that later steps could be attempted. Beyond the design of individual items (including testlets), computer-based testing can be adaptive at the level of test item or item block selection, posing more challenging questions to higher achieving students and conversely.

In addition to selecting items or item blocks according to individual learners' proficiencies, adaptive testing may also enable efficient targeting of specific assessment contents or item formats to

[3] Large-scale investments by governments to provide a notebook computer or tablet to each student are becoming more common. Examples of major investment programs include Portugal, Argentina, Brazil, South Korea, Turkey, and China among many more. An important part of the motivation of these investments is reconceptualization of the educational content and embedded assessments that define 21st-century learning.

just those students for whom they are appropriate. Adaptive testing in this sense might be used to focus on students with access to particular kinds of digital learning tools, or to focus on students engaged in particularly rich forms of instructional activities (e.g., problem-based learning). Today in some schools, learners have new computational tools for modeling and simulation that vastly expand the classroom instructional tasks that are possible. Further, modern science learning environments require learners to engage in higher level problem representation tasks and self-monitoring tasks. It will also be worthwhile to consider new emphases on teaching and assessing computational thinking (NRC, 2012). These observations lead us to ask: Are current assessments up to the job of monitoring and capturing the genuine range of complexity of the tasks that students engage in as they take part in modern learning environments? It is a challenge, with a standardized assessment but no standardized curriculum, to measure adequately the educational attainments of that significant minority of students with access to much more sophisticated instructional resources. If NAEP testing were adaptive not only to individual learners' prior test responses but also to their learning environments, then it would be possible to pose appropriately rich and challenging questions for that significant minority without burdening other examinees with questions to which they were unprepared to respond.[4]

Increasingly, computers will enable students to communicate their understanding by drawing graphs, through their interactions with simulation software, and via speech recognition and physical gestures. Use of new media technologies in assessment may also improve student motivation with items that are more interesting and engaging. Among other approaches, assessments might offer immersive experiences and input modalities familiar to users of video games and social media. Finally, computer technology can transform data collection and scoring, improving efficiency, reducing reliance on human scoring of constructed responses, and shortening turnaround time for reporting.

In assessment, as with learning and instruction, technologies are especially powerful in enabling fuller educational participation by students with disabilities, and for providing scaffolding tailored to the particular needs of English language learners and other significant student subgroups. Scaffolding and adaptive testing can improve measurement precision for students far below or far above the modal grade level. With flexible menus of available affordances, "universal design" approaches can offer individualized testing accommodations. Responding by voice or gesture may offer an empowering alternative to conventional response modes for some subgroups of students with disabilities.

In addition to the potential impacts of technology on the design and administration of stand-alone tests, some mention should also be made of technology's potential impact on the relationship between instructional activities and assessment activities. Technology has the potential to capture data unobtrusively during the course of students' ongoing learning activities. If such data are used to describe student learning or to diagnose learning difficulties, then instruction and assessment become seamlessly integrated. Curriculum embedded assessments[5] can already provide deeper elicitation and representation of students' expertise than most standardized tests, because standardized tests are only weakly related to the specific prior instruction students have received or to the particular classroom activities they have engaged in. New learning technologies promise to expand the capabilities of embedded assessment enormously. As students interact with word processors, online simulations, or immersive micro-environments, fine-grained records of their activities can be captured and analyzed. Note that the panel is not recommending consideration of embedded assessments for NAEP in the near-term. Rather, as

[4] In order to draw an efficient sample of participating schools, adaptive testing in this sense would require prior information about where specific new technology was being used for instruction. That information might come from technology vendors or from state data warehouses, for example.

[5] "Embedded assessment" refers to assessment that is inherently part of another activity, such that engaging in that activity provides useful assessment information and/or that the activity inherently requires use of the assessment information. Often, embedded assessments serve the purpose of "formative assessment," which may be defined as "a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (CCSSO FAST-SCASS, 2008, http://www.ccsso.org/Documents/2008/Attributes_of_Effective_2008.pdf).

these types of assessments come into use for instructional purposes, it may be valuable to monitor these developments and to investigate whether or in what ways these innovations can inform large-scale (curriculum independent) assessments.

Fourth, technology is transforming the infrastructure of education data and the timely use of information that could better serve instructional improvement. State longitudinal data systems are making it feasible to track individual students' progress over time, and the new Common Education Data Standards and the Shared Learning Infrastructure (see Section 4.4.2) promise to simplify greatly the sharing of information across state lines. Such large-scale data systems are expected to serve such purposes as recommending learning resources predicted to be well matched to learner needs based on prior data. Massive data warehouses may also make it possible for NAEP to incorporate achievement or other information from external sources in ways not yet envisioned.[6] Nearer-term, existing databases may enable stronger linkages between NAEP and other large-scale assessment programs, including international assessments.

In this section, we briefly describe changing knowledge representational technologies and user interface modalities, and their relevance both for assessing learners' abilities to perform at a high level and for providing "background information [that] serves the additional purpose of enriching the reporting of NAEP results by examining factors related to academic achievement in the specific subjects assessed" (National Assessment Governing Board, 2010, p. 47). We consider how technologies can both measure old assessment constructs in new ways, and measure new assessment constructs. Our aim is to point the way toward a differentiated, strategic approach whereby NAEP might thoughtfully respond to these changes. These are topics to be prioritized as R&D issues for the NAEP Innovations Laboratory; almost none has reached the stage of readiness for immediate implementation. That said, while acknowledging the substantial research investment required, we believe that some of the technologies already being used by consumers, as well as some technologies on the near horizon, may overcome constraints that have previously limited the ability of assessment regimes to flexibly determine progress in educational achievement and in the prior learning environments that are contributing to that progress.

[6] One possibility worthy of further exploration might be use of state assessment data and demographic information to project NAEP scale score distributions for schools or districts.

4.2 New Ways of Representing and Interacting With Knowledge

Two primary technology development vectors are especially relevant to NAEP's concerns. The first vector has to do with 'knowledge representations'—those technologies that people use to express and to display knowledge and information. The second vector has to do with the ways people produce or interact with knowledge representations—from handwriting, to keyboard-and-mouse typing and clicking, to more naturalistic interactions using touch, gesture, drawing and voice. Each is important as they are increasingly changing the learning contexts in which educational progress is being attained and will be measured.

4.2.1 Knowledge Representations (KR)

Among the symbolic systems that are used to represent knowledge, NAEP has long used paper-based written language text representations such as sentences and lists; mathematical symbol systems including equations, data tables and graphs such as the Cartesian coordinate system; scientific diagrams such as those for wiring a circuit; and geographic and map representations of the physical world. These KR are used to express assessment items, and students need to understand them and produce their own KR to meet the test demands.

Dynamic KR, displayed by computer and inviting learner interactions to explore or solve problems with them, are game-changing for assessment as well as for learning. Mislevy et al. (2010) provide a fruitful analysis of how assessment tasks can be structured around the knowledge, relationships and uses of domain KR, and of how technology can provide more dynamic assessment task design supports and automated task construction and scoring. Cisco Systems' computer network simulation assessments of design and troubleshooting, with over a million students trained, is offered as a particular mature example (also see NRC, 2011a, pp. 35-37).

Recent forays into technology-enhanced performance assessments in science provide salient examples. WestEd's SimScientists project explores how simulations can play important roles in enriching learning and assessment in life science, earth science, and physical science. For example, learners using SimScientist simulations are challenged to explore interactively how to balance the proportions of algae, shrimp, and alefish to keep the food chain and environment healthy. The SimScientist environment employs multiple linked representations, such as model components under learner control, a runnable ecosystem model, graphs depicting changing quantities of model components, and summary data tables from graphed values of the running model. Students can be assessed in their design of experiments and controls, in their graph interpretations, in identifying functional relationships with model components (such as sketching a food web in an ecosystem), and so on. Many other examples are provided in the work of the TELS (Technology Enhanced Learning in Science) Center, which developed interactive computer models that can be used for learning and assessment in many major topics in the high school science curricula for biology, chemistry, and physics. Some good examples in practice from around the world for one-to-one e-learning include recent curriculum and integrated formative assessment software from Adaptive Curriculum (CeBit fromTurkey), Time to Know (Israel), and Sigong Media iScream (Korea). NAEP's Technology and Engineering Literacy framework[7] and assessment development for the NAEP Science Interactive Computer Tasks[8] offer a useful contrast, showing what is already envisioned within the constraints of NAEP's purposes and overall design. An excellent beginning has been made with the Problem-Solving in Technology-Rich Environments NAEP field study (Bennett, Persky, Weiss, Jenkins, et al., 2007; Bennett, Persky, Weiss, & Jenkins, 2010).

As these examples illustrate, curriculum-embedded assessments may offer a rich source of examples for new item formats and expanded construct definitions, even though their ties to a particular curricular program limit their direct applicability to

[7] See http://www.nagb.org/publications/frameworks/tech2014-framework/ch_z/appendix_d.html

[8] See http://www.caesl.org/conference2006/NAEP%20ICTs.ppt

NAEP. [9] Most embedded assessments are "curricular assessments" (unit tests, quizzes, and in-class assignments), designed to determine if students have grasped the enacted curriculum. These assessments are often the first to address new content topics, such as engineering, modeling, and simulations as they enter the curriculum. They will also include new constructs, reflecting important processes and practices of the disciplines—constructs like "meta-representational competence" (diSessa & Cobb, 2004) or "cognitive complexity." These do not reside in particular content but rather cross sub-fields and include such knowledge as how to successfully gather, represent, synthesize, act on, critique, and present information. However, even though these cross-cutting skills are called for in a broad range of situations, they can only be assessed in the context of situations with which students are already somewhat familiar.

Evolving uses of dynamic KR media are relevant for assessment in part because these media are already transforming the work and knowledge production practices of adults in society. To the extent that we neglect in our assessment practices how mature practitioners in the communities of practice for a knowledge domain do their work, we render our assessment results less predictive of success in work in these domains. Discoveries in the biological and physical sciences, mathematics, economics, engineering, history, journalism, medicine, health, even literary studies and humanities—among many other domains—are increasingly dependent upon new technologies for data capture and sensing, KR, simulation and modeling, including large-scale data-mining and interactive data visualization and analysis. A clear explication of this point concerning the need for K-12 education to promote these technology-associated practices of science and engineering is provided in the new Framework for K-12 Science Education (NRC, 2011a, see Chapter 3). NAEP should investigate ways to incorporate these new KR technologies, so as best to inform policymakers and educational leaders about the progress of our nation's young people in acquiring the full range of reasoning and analytic skills demanded by the economy they will join and grow.

[9] Some tasks might be common enough across a range of popular curricula that they could be adapted directly.

4.2.2 User interface modalities

The problem of modality and bandwidth has been around for many years (see 1992 NAEP 'Reading Assessment Redesigned'). It is a large point that NAEP constructs are likely to be under-representing learner knowledge by defining proficiency largely in terms of text-based assessment media while the world beyond the test admits of reasoning with gesture, sketching, voice, and other modalities. We may have better opportunities to measure accurately the full range of students' knowledge and reasoning if we can open up the assessment environment to provide for more multi-modal and dynamic problem representations and for more naturalistic interactive modes for students to demonstrate their competencies. In our increasingly pluralistic and multi-lingual society, these new modes of expression in assessment may be particularly important in assuring that all learners are able to demonstrate what they know and are able to do with suitable affordances.

While human-computer interaction (HCI) technologies first demonstrated in laboratories often take 10-15 years to mass-market adoption, several of the user interface modalities below are already in consumer products, making their application to educational purposes feasible in the near future. These technologies are making HCI simpler, increasing the range of what digital technologies are being used for, and significantly reducing the ages at which children can effectively interact with a range of digital devices. We are seeing large-scale initiatives for each child to own an educational digital device (one-to-one e-learning) by countries including Portugal, Argentina, Brazil, Turkey, and many others. In terms of priority, we see applications of gesture/touch, voice, and digital stylus input having nearer-term applicability in NAEP learning environments and learner assessments than visual recognition/augmented reality. Sketching is still further out in its widespread applications among these user interface modalities. Although the examples and possibilities we present below are primarily focused on education rather than assessment, NAEP needs to be aware of these HCI technologies, both to understand what experiences and habits students are bringing to the assessment and as possible sources of innovative task and assessment design.

4.2.2.1 Gesture and touch

In the past few years, with advent of mass-scale consumer user interfaces such as Nintendo's Wii and Microsoft's Kinect, kinesthetic interaction with computer games using whole-body and gestures as inputs has become an appealing interactive modality, and has been rendered in rich, immersive graphical environments that make possible forms of continuous (or sampled) assessment associated with competencies in playing these games. The potential exists to have K-12 learning and assessment be mediated by gestural interaction with KR and problem-solving tasks in the virtual worlds employed in these gaming environments. Smaller form touch screen user interfaces such as the iPad are also coming to be used for bringing interactive 3-D KR into e-textbooks in domains such as chemistry, with the potential for interactive assessments employing such a touch input modality. Construct definitions will need to evolve as kinesthetic performance comes to be accepted as part of what have historically been viewed as cognitive domains.

4.2.2.2 Voice

Voice recognition has been heralded for several decades as the coming naturalistic interface to computing. Finally there are mass-market applications of these capabilities, and they leverage the vast information databases and web services that have accumulated. For example, Siri on the iPhone 4S, launched in Fall 2011, is billed as a virtual personal assistant, and while exploiting voice recognition functionality, Siri is not a separate application but a fundamental part of the iOS5 system software, exploiting a variety of artificial intelligence capabilities. Unlike GPS devices or car entertainment systems, specific voice commands are not needed to control the software, since a natural language interface is used and users are encouraged to speak as they would to a person. Siri users have access in their requests to receive answers from such web services as Wikipedia, Wolfram Alpha, Yelp, and Google Maps, or to have their wishes fulfilled in terms of making calls, taking notes or sending email, playing music, setting reminders, adding or querying calendar events,

getting stock quotes or movie times, and retrieving many basic facts. As the software improves over time due to greater volumes of usage and learning from its limitations, Siri will be even more likely to have an answer for you. Many writers have been speculating about the implications of Siri for changing education. These include replacing reference books, less emphasis on memorization, freeing teachers to spend more time teaching children to apply and understand new learning, greater accessibility for learners with visual impairments or physical difficulties typing, facilitating learning new vocabulary and conducting spoken assessments of language learning.

4.2.2.3 Visual recognition, visual search and augmented reality

Advances in machine vision and machine learning have led to systems that sense the content of visual environments and provide associated information services that can be used for learning, and in principle, assessments. For example, Google Goggles provides image-based web search, text recognition and translation for cell phones and other mobile devices. WebCam Laboratory (an educational software company in Hungary) enables kids to track moving objects (pendulum or a ball rolling down a track) with a tablet webcam and graph the objects in real time, among other visual recognition functions for scientific exploration. ClassmateAssist from Intel Labs is a netbook-based application using computer vision and context-aware computing to assist students and teachers in effectively using mathematics manipulatives in kindergarten through second-grade classrooms (Beckwith, et al., 2010).

Similarly, augmented reality (AR) is being incorporated into smartphones (e.g., iPhone 3-4S, any Android, Symbian OS) by overlaying graphics, audio and other modalities onto real time environments being viewed through the phone. AR uses a variety of technologies including GPS antennae and other geolocation techniques, as well as 3G and Wi-Fi networks. AR apps often exploit digital compasses, accelerometers, and the smartphone cameras. Using these sensors enables inference of the user's approximate location and viewpoint, so that relevant information about the sensed environment can be overlaid onto the camera image for the user to see and interact with. Educators are enthused about AR as a mobile learning technology that can provide place-based information for learners, and assess learning in situations when relevant.

4.2.2.4. Sketching

In many knowledge domains, one can assess what learners know by asking them to draw a sketch of an environment, labeling its primary components, and using symbols such as arrows to depict forces or other ontological items and their relationships in the knowledge domain. Recent developments such as CogSketch—an interactive system for sketch-understanding—from Ken Forbus's group at Northwestern University show how an assessment tool using visual expressive media can function. They have been evaluating spatial scientific knowledge in Earth sciences (geological formations) and mechanical design as domains to explore what such sketch-understanding systems require, though we can readily imagine similar systems for optics, mechanics, cell biology and other spatial scientifically 'rich' knowledge domains where graphical depictions are a key knowledge indicator.

4.3 Technology, Learning Environments, and Instructional Tasks

Consider the importance of how today's learning tasks and their associated learning materials make increasing use of publisher and open e-textbooks, interactive media, web-based resources, and massive educational data warehouses, with a trend toward interoperable data formats across states. Interactive problems and multimedia visualization capabilities that go beyond static text, diagrams and equations lend themselves to enhanced capabilities for learners' use while studying, and possibly to more embedded assessments, capturing achievement data while students engage in learning activities. "Flipped classrooms" are having students watch web video lectures at home, with school providing more personalized teacher guidance and project-based learning. More children are being home-schooled with such resources. These trends together indicate that U.S. children are experiencing increasing learning environment variability.

In this section, we ask: Is it possible for NAEP results to be more sensitive to measured aspects of the

learning context? Classroom and learning task infrastructures are changing, as is the broader educational environment that children experience outside of school. A full accounting of the nation's educational progress demands attention to the taught curriculum in all the environments in which children are learning. Schools, of course, remain the first priority for assessments today, because they are where most designed learning is happening. A key principle that we would urge is that, when the learning and teaching infrastructure changes in significant ways, NAEP's approaches for measuring educational progress should co-evolve with those changes.

In general, the tasks set for students to measure their achievement must be neither completely familiar nor completely unfamiliar. A problem identical to one a student has seen before may elicit no more than rote recall, but faced with an utterly foreign task, a student may have no way to begin. One can conceive of a continuum in the degree of correspondence between assessment tasks and classroom activities. The ideal task for eliciting complex reasoning and problem-solving may be one that draws upon a rich base of prior knowledge and skills, but requires using that knowledge in new ways. Questions concerning matters that students have already studied can "drill deeper" than questions concerning matters newly introduced to them.

This notion of prior familiarity extends beyond content and procedural skills alone. A student accustomed to using a word processor may be expected to write better with that familiar tool than without it. A student accustomed to using a spreadsheet program or specialized software for graphing and data exploration may be expected to arrive at more insightful interpretations of quantitative data when these tools are available. Beyond such individual uses of technology, students who work in classrooms where peer collaboration and exchange of ideas is valued may perform less well than otherwise on tasks they are required to perform in isolation. Finally, some students, especially those from groups stereotyped as low-performing, may do better in the familiar motivational context of routine classroom activities than in the more stressful context of a formal assessment. All of us "work smarter" in familiar environments, equipped with the tools that make these environments "smarter workplaces."

Proficiency with academic tasks across a range of contexts is important, of course, but assessments offering impoverished contexts for student work may fail to reveal the full range of their capabilities.

Inferences about learner's knowledge depend on capabilities enacted in some range of environments. As learning contexts become more varied, it will become increasingly important to contextualize the learning competencies that NAEP aspires to measure in terms of how teaching and learning have occurred. Students may be expected to perform best in familiar environments, equipped with familiar tools for representing knowledge and for communicating their understanding. If the measurement of best performance, in this sense, is valued, then for both the measurement of old constructs in new ways with technologies and for the measurement of new constructs, instructional context becomes increasingly relevant.

This suggests an exploration of ways to extend NAEP's current infrastructure for adaptive testing, tailoring item selection to aspects of learning environments, as opposed to adaptive testing solely according to students' proficiency as estimated from prior item responses. This proposal necessarily implies that NAEP must either access existing information or else directly measure aspects of learning environments. We see this recommendation as in the spirit of the background data that NAEP already collects to examine "factors related to academic achievement in the specific subjects assessed."[10] It might imply special studies of student performance in particular kinds of

[10] "As stated in Governing Board policy, the collection of background data on students, teachers, and schools is necessary to fulfill the statutory requirement that NAEP include information whenever feasible that is disaggregated by race or ethnicity, socioeconomic status, gender, disability, and limited English proficiency. Background information serves the additional purpose of enriching the reporting of NAEP results by examining factors related to academic achievement in the specific subjects assessed. To satisfy the goal of enriching reports on student achievement in reading, background variables are selected to be of topical interest, timely, and directly related to academic achievement. The selection of variables about which questions will be developed may reflect current trends in the field, such as the use of technology in reading instruction or the extent to which students use the Internet as a reference tool" (NAGB, 2010, p. 47).

technology-equipped classrooms, for example, not for the purpose of evaluating that technology, but instead to characterize the range of proficiencies the population of students accustomed to such technology is able to demonstrate.

Learning with new e-textbooks that embed problem sets and multimedia interactives could provide NAEP with new data on changes in learning environments. For example, NAEP's evaluations of its reading assessment will be more meaningful if contextualized by empirical documentation of changes in learning environments as reading to learn in grades 4, 8, and 12 becomes supported by digital text reading technologies. The NAEP 2011 Reading Framework states that 12th grade students performing at the Advanced level should be able to analyze both the meaning and the form of the text and provide complete, explicit, and precise text support for their analyses with specific examples. Besides common e-text features like enabling unknown words to be spoken or defined at a touch, text analysis tools like the Visual Thesaurus VocabGrabber enable students to create a digital concordance of word uses by the author(s) in texts they are studying – which may materially improve their sense-making with complex literature. Similar issues may be raised for NAEP science and mathematics assessments concerning the learning technologies employed as students learn.

Another noteworthy aspect of the NAEP 5- to 10-year horizon in changing learning environments is that most students will come to school with a device that has computational power exceeding current smartphones and tablets—which could arguably form the technical backbone of a new assessment infrastructure if concerns of equitable access to the curriculum and instruction that they can provide can be addressed, as well as concerns with test standardization and test security. For NAEP assessments, sampling of both persons and occasions will be the central concern, although non-universal access would compromise sampling as well as equity. Keeping with the equity theme, an enhanced capability on NAEP's part to sense characteristics of learning environments could lead to more accurate policy analysis. With deeper contextual information, policymakers will have more nuanced insight into the potential instructional impact of various technologies. So armed, policymakers and others will be better able to make investments that might go some way to leveling the nation's education playing field.

Issues like the aforementioned place in sharp relief how technology platforms are non-neutral actors in the strategies we use to track educational progress. On the horizon are technologies and applications that may allow us to re-conceive assessment, in general and in regard to NAEP. Both would allow for innovative assessment items that are sensitive to these enduring problems of assessment and open up new ways to communicate to NAEP stakeholders using more 'active reporting' (see Section 5.6), such as interactive graphs. Knowledge refers to a capability to perform in some range of contexts. As technology transforms the contexts for teaching and learning, the constructs NAEP measures may need to change, as well.

4.4 Technology and Assessment

4.4.1 Measuring old constructs in new ways

At the center of modern learning sciences is the fundamental idea that learners think, reason, problem-solve and act at any point in time based upon a complex system of knowledge, strategies and dispositions. For virtually any knowledge domain that has been examined empirically, it has been found that the learner's cognitive system is occupied with conceptions that deviate from those taught in the discipline (variously called 'misconceptions' or 'alternative conceptions': e.g., Chi, 2005; NRC, 2000). It follows from this observation that accurate interpretations of students' performance on problem-solving tasks may require discerning the sorts of misconceptions in play as they attempt to apply their knowledge systems. In addition, one potentially powerful way to frame target constructs and describe achievement more meaningfully may be to examine directly the distribution of students' conceptions and misconceptions, especially where these can be arrayed in "learning progressions" characterizing typical trajectories whereby students attain mature understanding (see Section 3.2.4). Here, we conjecture that both expanded KR and newly expressive capabilities like voice, gesture, and sketch recognition technologies will make learner's conceptualizations much more visible. It is important as well that NAEP assessments that do

not employ such technologies may be missing opportunities for distinguishing alternative conceptions from more fully mature understanding. NAEP could be thoughtful about priorities for measuring old constructs in new ways by looking for guidance from research on the power of learning assessments using these new media, such as the work on simulations in science below. The larger point is that the assessment technology should match the domain/construct being assessed, and the types of behavior that will be observed to assess that construct.

4.4.2 Assessing new constructs

Technology is reshaping the cognitive, interpersonal, and intrapersonal demands of tomorrow's world (Jenkins, 2009; Levy & Murnane, 2005). If NAEP is to fully inform policymakers and the public about students' acquisition of the range of skills the modern workplace will require, assessment frameworks will need to be expanded, and assessments of new constructs will need to be developed. Conceptualizing and piloting these new assessments will be a key activity for the NAEP Innovations Laboratory. Technological affordances are already transforming school activities, as with reading using hypertext, for example. Consider also the "21st-century skills," including clusters defined by the NRC (2011a) as cognitive skills (e.g., non-routine problem-solving, systems thinking, critical thinking), interpersonal skills (e.g., clear communication, collaboration, leadership), and intrapersonal skills that aid one in problem-solving (e.g., adaptability, metacognition/self-regulation, emotional intelligence). While construct definition issues abound, the NRC report makes a strong case that these non-traditionally measured competencies in K-12 should become part of our education assessment focus. These constructs would not stand alone as separate scales, but would instead be integrated into expanded subject matter frameworks.

With regard to cognitive skills, some curriculum-embedded "performance assessments" may be of particular interest in suggesting potential assessment task formats. These are tasks involving technologies that allow monitoring not only of end products, but also intermediate products and in some instances, learning processes. An example is recording the sequence of changes to text done with word processing software. The challenges of analyzing such recordings are formidable, but they may eventually provide evidence of editing skills or may even inform a student model underlying the writing performance. As assessments move toward computer administration, collecting such records of intermediate performance and processes will become more feasible. With further R&D, interactive assessments may prove especially promising in making performance process-embedded assessments structured enough to enable efficient, reliable scoring. In mathematics, delivering assessments on new technologies permits direct access to student use of manipulatives and to their interactions with alternative KR, such as graph transformations, algebraic manipulations, and choices of numbers to substitute to check values for equations. Students' behaviors and sequences of actions may serve as the basis for new assessments capable of characterizing problem-solving processes. They might also improve models of student knowledge, providing better next-stage instructional guidance (Koedinger, McLaughlin & Stamper, 2012, in press).

With regard to interpersonal skills, what each individual can accomplish is expanded by the opportunity to work with others. Collaboration is thus a common ingredient to work in society and an integral component of a number of knowledge-building pedagogies. Furthermore, we know that much future work that today's NAEP examinees will encounter will be collaborative. Today's gaming and tomorrow's augmented reality technologies will bring a press for assessments designed to measure collaborative learning. One could imagine that NAEP might seek to assess the range of such skills in examinees in the future. As part of their work with the Organisation for Economic Co-operation and Development to develop the frameworks for the 2015 Programme for International Student Assessment (PISA) (used in 74 countries during the 2009 test), Pearson will be developing a new Collaborative Problem-Solving assessment in recognition of the ways young people will have to learn and work throughout their lives. Below we posit that currently available knowledge representation technologies may point the direction toward future NAEP assessments in this area as well.

With regard to intrapersonal skills, research on formative assessment has shown that when students are more aware of the goals for learning and take responsibility for peer and self-awareness of learning progress, they are more effective in learning. Formative assessment practices have also emphasized encouraging students to assume responsibility for learning. Research has shown that successful learners respond productively to challenge, with curiosity, engagement, and persistence. The panel is not recommending any expansion of NAEP to incorporate these constructs in the near-term, but does recommend continued monitoring of research and applications that may have future potential for expanding large-scale assessments in this direction, especially technology-supported expansions of formative curriculum-embedded assessments.

4.5 Technology and Education Data Infrastructure

4.5.1 Expanding field of assessment programs and interest in cross-program linking

The number of large-scale testing programs has increased dramatically since the inception of NAEP. These programs include state and district accountability tests as well as international efforts such as the Trends in International Mathematics and Science Study (TIMSS) and PISA. As testing programs proliferate, so do their results, and the opportunities for cross-test comparison of rankings, trends, and gaps increase exponentially. If the market for tests and test-based inferences continues to expand, the corresponding proliferation of possible cross-test comparisons represents a challenge for interpretation, especially to the degree that results are inconsistent across tests. An attractive solution to the problem is the formulation of a linking procedure that current and future testing programs can use to report their results, perhaps on a NAEP scale.

This proposal has a historical precedent. In the late 1990's, the discussion of the Voluntary National Test led to proposals to establish NAEP as a common scale for state tests. This motivated an important NRC report that addressed the question, "Is it feasible to establish an equivalency scale that would enable commercial and state tests to be linked to one another and to NAEP?" After that committee concluded that the answer to this question was, "no," a second NRC report addressed a more specific procedure that involved the embedding of NAEP questions in state tests in order to obtain individual scores locatable on a national scale. The second committee determined that this procedure was also infeasible.

There are at least three arguments that motivate reopening or reshaping the question of whether NAEP can serve as a common scale for other testing programs. The first argument assumes (a) that test scores are more visible and consequential than ever before, (b) that the general public deserves to have a resolution of cross-test discrepancies when they arise, and, most importantly, (c) that some person, or organization will try to resolve cross-test discrepancies with available information if an official linking is not provided, and that such efforts are unlikely to be as thoughtfully conducted or carefully contextualized as an official, albeit imperfect effort might be.

An example consistent with this first argument is the recent effort to map state "proficiency" scores onto the NAEP scale. With a public well aware that proficiency cut scores varied in stringency across states, a number of studies used readily available information to show that the reported proficiency percentages on state vs. NAEP tests had very weak relationships. An official and more analytically defensible response known as the Mapping State Proficiency Standards[11] lagged years behind these initial demonstrations. These mappings are not technically equivalent to the linking of entire score scales, but the argument supporting the linkage of one score to another is very similar to that supporting the linkage of one scale to another. This argument is precisely what the previous NRC reports determined to be insufficient.

Although this panel has reservations about the stability and interpretability of the NCES state mappings, there may be a lesson learned about the value of supplanting popular and seemingly inevitable cross-test linking endeavors with the best possible analysis, well caveated, with clear statements about and examples of indefensible extensions. To borrow a sports metaphor, there may

[11] See http://nces.ed.gov/nationsreportcard/studies/statemapping/

be situations where the best defense is a good offense.

A second argument helps to reshape of the question of whether NAEP scores can serve as a common scale. The NRC reports were justifiably cautious about a full linking of two score scales that support individual reporting. However, some tests, like NAEP, TIMSS, and PISA, do not support the reporting of individual scores, let alone reportable classroom and school scores. As the level of aggregation is raised to states and countries, some (but not all) of the concerns about the unreliability of scale linking are diminished. This has helped to motivate the ongoing NAEP-TIMSS linking that may serve as a model for future linkages between NAEP and other large-scale tests that focus exclusively on aggregate reporting. It should be noted that even as aggregate-level reporting reduces concerns about the unreliability of scale linking, it may increase other concerns. Comparability of score distributions would depend not only on the accuracy of assessment scale linking, but also on comparable population definitions and appropriate sampling. Also, subtle differences in construct definition that might be swamped by measurement error at the individual level could become more salient in aggregate-level comparisons where measurement error is averaged out over large groups of examinees.

Perhaps most significantly for this report, a third argument for reopening the linkage question is the opportunity the panel sees to modify the NAEP design so as to facilitate better linkages. Specifically, we argue elsewhere in this report for the flexibility to modestly expand NAEP exercise pools beyond coverage of the NAEP frameworks. This expansion is intended to enable inclusion of additional content to improve alignment between NAEP and other major assessments to which more accurate linking is desired. The NAEP Innovations Laboratory might also field special bridging studies to account for differences in school calendars and testing windows. In short, if better test linkage is a higher priority, and if resources are allocated to address the problem, then it should be possible to move beyond some (although not all) of the constraints that led to the earlier NRC committees' pessimistic conclusions.

Together, these arguments support a recommendation to investigate the feasibility of a standardized procedure for establishing cross-test NAEP links, evaluating their validity, and monitoring the validity of the linkage over time. For example, one part of this approach might take the form of a small number of representative item blocks in key subjects and grades that could be inserted into target tests. These would not be made public, but would be provided upon request, subject to strict guidelines for security and the linking design. A key requirement of the linking protocol would be a procedure for continued investigation of the link over time, as the drift in any test link over time represents strong evidence against its use—evidence that is not available at the time of the initial link. Linkages should be investigated for one-time use on a case-by-case basis and then reinvestigated at each major reporting interval.

4.5.2 Alignment of infrastructure with state data warehouses

The data warehouses that collect, store, and support analysis of state testing data have become increasingly sophisticated in recent years. The Statewide Longitudinal Data Systems (SLDS) grant program has provided funding to 41 states to support these systems and, in many cases, establish links to early childhood and postsecondary outcomes. The recent release of the Common Education Data Standards (CEDS) Version 2 and evolving standards for question and test interoperability (e.g., http://www.imsglobal.org/) will further facilitate assessment and data sharing. Even without longitudinal linkages, the kinds of data gathered by these warehouses offer at least three opportunities: (1) gaining greater perspective on the representativeness of the NAEP sample and the stability of sample characteristics over time; (2) supporting feasibility studies for links between state and NAEP score scales; and (3) addressing ongoing concerns about inconsistent and unrepresentative NAEP sampling of English language learners and students with disabilities.

NAEP already capitalizes on the results from state assessment programs to improve the efficiency of within-state samples. If schemes for adaptive testing are to be extended so that assessment items can be targeted according to students' prior instructional histories, current learning environments, or available technology supports,

then state data warehouses might also be tapped for these kinds of information.

There remain considerable design inconsistencies across state data warehouses, so near-term recommendations must be modest and begin with checks of sampling validity in particular states. As database linkage becomes the norm, however, longitudinal information can support additional opportunities, including: (1) exploration of a vertical scale bridging grades 4 and 8; (2) investigation of predictive relationships between early childhood and early grade outcomes and NAEP; and (3) investigation of the predictive relationships between NAEP and future outcomes including later grade tests and postsecondary data. The committee does not recommend pursuing these particular opportunities in the near-term, as they extend far beyond the current infrastructure and purview of NAEP. However, as a useful short-term goal, NAEP might investigate and describe the extent and feasibility of state database linkage, both cross-sectionally and longitudinally. This would both inform near-term, NAEP-related studies and help identify the opportunity space for the future.

Of particular significance is the emerging development of the Shared Learning Infrastructure (SLI), a non-proprietary interoperable educational data store and associated web services that aims to facilitate the implementation of the CCSS by helping states and districts provide teachers with the instructional data and tools they need through shared online services, integrating educational data created and managed by a variety of state education agency (SEA) and local education agency (LEA) source systems. Leveraging a broad variety of open standards, the SLI will provide an Application Programming Interface (API) layer and Software Developer Kit (SDK) to make those data (such as student attendance, transcript, assessment data, class schedule and customized data unique to a particular SEA, LEA or application) available for third-party application development by for- and non-profit vendors and content creators. Since the CCSSrecognize that there are multiple pathways for a learner through a 'learning map' of competencies, the personalized learning vision is a much more fine-grained and dynamic learning data record that can be developed for students, enabling easier discovery of instructional content adapted to their learning needs and informing more personalized support from their teachers. The Shared Learning Collaborative, LLC (SLC) is a temporary governing entity established for the design and development of SLI technology and the long-term SLC organizational model. The SLC Initiative aims to accelerate the progress of public schools toward personalized learning for all students, and is led by the vision of the Council of Chief State School Officers (CCSSO) and nine participating states that collectively serve 11 million K-12 students (Phase 1: Colorado, Illinois, Massachusetts, New York, North Carolina; Phase 2: Delaware, Georgia, Kentucky, Louisiana), with funding from the Bill & Melinda Gates Foundation and Carnegie Corporation of New York.

Such massive data warehouses as the SLC may make it possible for NAEP to incorporate achievement information from other sources in ways not yet envisioned. Nearer-term, existing databases may enable stronger linkages between NAEP and other large-scale assessment programs, including international assessments.

4.6 Implications for NAEP

Today's NAEP assessments require same-time and same-place provision for the people involved. Furthermore, they all have to be doing roughly the same thing, though with NAEP's matrix sampling, students in the same room may be responding to different booklets of test questions. Given that substantial learning of relevance to educational assessments is taking place outside school, in online environments, and in the physical world (NRC, 2009), and increasingly using smartphones, the time may soon come to consider assessing educational progress in ways that relax the same-time, same-place, same-thing provisions of NAEP today. To the extent that learning migrates away from formal school settings, current NAEP sampling frames may become inadequate.

Learner performance assessment in virtual worlds and immersive games has been a recurrent theme in federal and foundation grant programs in the past few years (e.g., Ainsworth et al., 2005; MacArthur Foundation's Digital Media and Learning Initiative, 2006-2011; NSF Cyberlearning Task Force, 2008; National Education Technology Plan, 2010; NRC, 2011b; PCAST, 2010). Shute, Ventura, Bauer, and Zapata-Rivera (2009) articulate a vision where

assessment becomes a more frequent, regular, embedded and unobtrusive part in everyday activities like gaming.[12] As discussed elsewhere in this report, formative classroom assessment can sometimes be embedded unobtrusively into classroom instructional activities. One might imagine that someday NAEP, too, might rely on assessments embedded in learning work itself, instead of as pull-out activities from the classroom. This possibility, if realized, would bring a sea change in assessment practice.

As we consider assessing new constructs, suggested by the earlier discussion on the dynamic knowledge representations used for learning and assessment involving science simulations, it seems likely that there will be significant stresses placed on today's measurement technologies. One way to frame the psychometric challenge here is in terms of standardization. So long as assessment tasks, testlets, blocks, or other modules are defined such that they can be replicated across examinees and used in different times and places, then it should in principle be possible to calibrate them using current psychometric methods and incorporate them into an item bank or exercise pool. However, when the constraints of standardization are relaxed, as with less highly specified, more open-ended activities or problems admitting of multiple possible solutions it will become more challenging to use students' responses to estimate their proficiencies on a common scale. As NAEP moves into its next decade and beyond, it will be important to continue fundamental research on psychometric models and methods if we are to have the tools at hand to capitalize on new sources of information about student achievement and the nation's educational progress.

[12] Similarly, as e-textbooks enable more dynamic embedded assessments and large-scale educational data tracking and mining, more personalized learning recommendations can be provided as patterns are identified of greater and lesser success in learning pathways using learning resources tagged for the core learning standards with which they are associated.

5. NAEP Reporting and Use

Any argument for the validity of NAEP-based interpretations must eventually address the metrics and mechanisms for score reporting and delivery. Inferences based on NAEP are mediated through NAEP reporting metrics, tables, and graphics.

NAEP has been successful over the years in conveying complex assessment results to a wide range of audiences. However, there are several ways in which the reporting metrics and mechanisms can be improved to make the information more accessible and useful, and to ensure that the inferences drawn from the results are valid.

In this section, the panel makes recommendations for the future of NAEP reporting metrics and their media for delivery. Specifically, we propose: shifting away from achievement levels as the primary means of reporting NAEP results and enhancing the interpretability of NAEP scales in other ways, improving the accuracy and consistency of expanded population estimates, investigating ways to respond more adequately to demands for small subgroup reporting, increasing the use of active formats that permit interactive tools for producing reports on NAEP results, and developing reporting metrics responsive to the likely demand for measurement of performance against the CCSS.

5.1 Background and History

Over its 40-year history, NAEP has taken several approaches to reporting results. These approaches have attempted to balance the desire for high precision on a global reporting scale with the desire to provide task specific, educationally informative and descriptive accounts of student educational progress. Over time, there has been a shift in emphasis away from detailed reporting focused on specific tasks, toward heavier reliance on numerical scales that summarize performance over broader domains. At the same time, there has been a trend from summarizing performance for large groups (population and subpopulation estimates for four broad geographic regions within the United States) toward reporting for smaller regions (states and large school districts).

Originally, NAEP reported results for individual items (then called exercises), indicating the percentage of test takers who answered each correctly. NAEP's founders believed that this approach would make more sense to NAEP's audience than an overall score on a collection of unrelated items (Linn and Dunbar, 1992). In addition, they believed that the reports would provide educationally valuable information, since the assessment at the time consisted largely of complex tasks, rather than discrete selected-response questions.

However, this approach did not satisfy policymakers, who were more interested in more global findings. So NAEP, which was then operated by the Education Commission of the States, added reports that indicated the average percent-correct scores on sets of exercises. This approach had limitations—for example, it could not show trends over time if the exercises differed—but it was the primary method of reporting for nearly ten years.

The reporting method shifted again in 1984, when Educational Testing Service (ETS) took over the NAEP operation. The new approach that ETS undertook relied on IRT to create a scale with a theoretical range from 0 to 500 that spanned ages and grade levels. The mean for the scale across all three age cohorts tested was 250.5, and the standard deviation was 50. ETS identified items associated with performance at selected anchor points—150, 200, 250, 300, and 350. These items were then used to develop descriptions of what students at each anchor point knew and were able to do (Beaton & Allen, 1992).

NAEP has used the scale scores and anchor points since 1984. But the Governing Board, which was created by 1988 legislation that reauthorized the assessment and called for reporting levels of achievement on NAEP, established "achievement levels," or performance standards, and reported the proportion of students at the Basic, Proficient, and Advanced achievement levels on each assessment, together with the remaining proportion Below Basic. Achievement levels were intended to enhance NAEP reports by adding an evaluative component; in addition to describing what students knew and

were able to do, the achievement levels were intended to indicate how students performed compared to what they should know and be able to do (Koretz & Deibert, 1993; Pellegrino, Jones, and Mitchell, 1999).

The achievement levels were first used to report results from the 1990 mathematics assessment, with achievement level results appearing in a separate publication from the main report. They were incorporated into regular NAEP reports beginning with the 1992 national assessment in reading. The establishment of achievement levels has been subject to considerable criticism and repeated negative evaluations, despite repeated modifications of standard setting procedures (Brown, 2000; Pellegrino, Jones, & Mitchell, 1999; Vinovskis, 1998, pp. 41-57). Despite these concerns, NAEP achievement levels remain the principal vehicle for reporting NAEP findings; are widely cited by policymakers; and, pursuant to the 2001 NCLB legislation, have also served as a model for state test reporting.

The goal of NAEP reporting, from printed and online reports to the NAEP Data Explorer, is to summarize NAEP data in a manner best suited to support various lines of inquiry in research and policy. The summarization of data, in metrics, tables, and graphics, cannot be a casual, passive process but instead requires careful consideration of a target audience and a target inference. The recommendations that follow are specific examples of this deliberative process. Reporting metrics well suited for some inferences will be misleading in support of others. Reporting metrics that are well understood for some audiences will require tutorials for others. The committee makes the general recommendation that NAEP reporting should represent an active and even instructional practice of providing particular audiences with the metrics, tables, and graphics best suited to support their desired inferences.

5.2 Shift Achievement Level Reporting to the Background

The legislative history of the three NAEP achievement levels—*Basic*, *Proficient*, and *Advanced*—begins with their establishment as goal statements (1988) and, later, indications of appropriate student performance (1994). At the time of NCLB (2001), achievement level reporting became a mandated reporting metric for all states, as reflected by the goal of 100 percent student proficiency by 2014. NAEP achievement level reporting has never officially emerged from its "trial" status,[13] and a 1998 report by the NRC, echoing the language of an earlier evaluation by the National Academy of Education, characterized the standard setting procedures as "fundamentally flawed" (Pellegrino, Jones, & Mitchell, 1999). Nonetheless, achievement level reporting, particularly with respect to "proficiency," remains not only the most widespread reporting metric but a central reference point of the rhetoric of school reform.

There are three flaws with achievement level reporting that together support our recommendation to move achievement level reporting to the background, to the level of appendices and footnotes. The default selection of average scale scores in the NAEP Data Explorer is the correct choice, and the panel applauds the shift away from achievement level reporting for gaps and gap trends in NAEP reports. However, the panel recommends further de-emphasis, particularly for trends, accompanied by an explanation of the limitations of achievement level reporting, to further discourage its potential selection as a reporting metric by secondary data analysts. If the flaws were few or inconsequential, the panel would recommend passive continuation of the metric in the interest of providing users with a familiar frame of reference. The panel's recommendation to shift achievement level reporting to the background arises from the wealth of literature on the limitations of achievement level reporting, both as a statistical metric and as a policy tool.

The first flaw with achievement level reporting is that the cut scores defining successive levels are

[13] A typical, recent NAEP report includes the following wording: "As provided by law, the National Center for Education Statistics (NCES), upon review of congressionally mandated evaluations of NAEP, has determined that achievement levels are to be used on a trial basis and should be interpreted with caution. The NAEP achievement levels have been widely used by national and state officials." (National Center for Education Statistics, 2011, p. 6)

determined judgmentally and are ultimately arbitrary. This was the primary criticism of the 1998 NRC Panel. This alone is not a sufficient reason to shift achievement levels to the background, as even an arbitrarily determined standard can gain meaning and relevance over time. However, if "percent-above-cut" (PAC) reporting is maintained, then a persuasive argument might be made for having less arbitrary methods of establishing cut scores. Beaton, Linn, and Bohrnstedt (2012) surveyed several alternatives that have been proposed—determining cut scores empirically to be maximally predictive of some future, valued outcome or linked to another familiar scale; benchmarking to international standards; and benchmarking to norms established in some baseline year. Hybrid methods are also considered. Retaining a PAC metric would mean that reports would still provide the estimated percentages of various student populations who were at or above one or more defined levels, but instead of at-or-above proficient, for example, reports might show percentages at-or-above the median achievement averaged across five high-performing countries in 2012.

The second flaw with achievement level reporting relates to defects shared by any PAC reporting scale. That flaw is the distortion the metric imparts to trends, gaps, and gap trends (Holland, 2002). Any trend, gap, gap trend, or relative ranking that uses, as its basis, the percentage of students above an achievement level standard will be confounded with the strictness of the standard and lead to inaccurate conclusions about status, progress, and gaps (Ho, 2008). These distortions are systematic—in general, using the percentage proficient metric, jurisdictions with overall proficiency percentages near 50% will see magnified trends, gaps, and gap trends when compared to jurisdictions with more extreme percentages. However, the distortions are also frustratingly unpredictable. An apt characterization of this reporting metric is that it is "short-sighted" with respect to the most commonly desired test-based inferences. As trends, gaps, gap trends, and cross-state comparisons are essential inferences that NAEP supports, a metric with these distortions is unacceptable for primary reporting, and any secondary reporting should include strong caveats.

Third and perhaps most subtly, the panel observes that achievement level reporting has coincided with a limiting framework for educational policies, wherein 100 percent proficiency is the only rhetorically acceptable goal. These policies restrict incentives to a particular region of the score distribution—those just below the "proficient" level—and can decrease incentives to teach both students far below and far above the cut score. These policies also politicize the selection of a cut score, where aspirational standards become diluted for short-term inflation of proficiency percentages. The problem is not the selection of a "proficiency" standard but the narrow use of a metric that can only detect whether students are above or below this standard. The panel believes that NAEP should lead a return to average-based reporting along with increased attention to average subgroup performance and percentile-based reporting. Rothstein, Jacobsen, and Wilder (2006) demonstrate convincingly that statistically sound metrics like percentiles or averages can also be standards-based. The average or percentile is simply compared to the standard. Standards-based reporting need not and should not be synonymous with achievement level reporting.

Together, these flaws motivate a relegation of achievement level reporting to the background in published guides, reports, and data tools and an active discouragement of achievement level reporting to reporters and secondary data analysts. Although rhetorically compelling, headlines framed by percentages of proficient students rely on proficiency as a weasel word, one that fosters the perception of a shared definition when no such shared understanding exists.

5.3 Alternatives to Achievement Level Reporting

NAEP scales might someday come to seem as familiar and intuitive to the general public as those used on the Scholastic Aptitude Test (SAT), American College Testing (ACT), intelligence quotient (IQ) tests, and Advanced Placement (AP) tests, but absent any guideposts, the current 0-500 NAEP scales have almost no meaning for most people. Labeling specific points along the scales as "Basic," "Proficient," and "Advanced" was intended to enhance interpretability, but critics have

contended that these markers are arbitrary and misleading. Nonetheless, reporting in terms of achievement level percentages is likely to remain an attractive option so long as the NAEP scales remain poorly understood.

NAEP has long provided item maps that illustrate the knowledge and skills demonstrated by students performing at different scale scores (e.g., National Center for Education Statistics, 2011, p. 29). These item maps show specific items tied to locations along the score scale. They are best able to communicate scale information when the complete item is viewable, not simply the content standard that it measures. For this reason, the NAEP Questions Tool becomes a powerful resource for any user who wishes to explore specific tasks on which students at a given scale score can succeed.[14] Printed NAEP reports already provide complete sample items anchored at the average scale score or at selected percentiles (e.g., National Center for Education Statistics, 2011, pp. 30-36). This kind of reporting could be expanded dramatically in online reports with links to the NAEP Questions Tool.

Market-basket reporting[15] is a kind of average-based reporting that relies on the idea of a representative sample of items comparable to a shopping cart at a supermarket. This model could be used for NAEP reporting in any of several ways, so long as the "market basket" collection was representative of some well understood collection of items. Scale scores, or averages in the scale-score metric, would be transformed to predicted performance for the market basket item collection, yielding a percent correct (across dichotomous items) or percent of total possible score (for dichotomous and polytomous items).

In one application of this idea, the market basket might be a changing but transparent collection of representative items, similar to the CPI model discussed previously. This kind of market basket could be used to maintain a meaningful score scale even as the mix of items changed gradually over time. A weakness of this approach is that it cannot support inferences about improvement in proficiency for a stable construct over time. This "dynamic assessment framework" deliberately confounds trends for fixed subdomains with changes to the domain itself. Clear communication about the contents of the market basket and changes thereto would be essential to ensure appropriate trend interpretation under dynamic frameworks.

Alternatively, the market basket might be a "NAEP lite" collection of released items designed to represent the entire exercise pool or some meaningful subdomain (e.g., the CCSS), modest enough in size that an interested user could review the entire collection and thereby gain some insight into the meaning of a given percent correct score. Even further, with a market basket of released items, an average scale score could be expressed in terms of the chances of answering each market basket item correctly.

The comfort and familiarity of the percentage metric is an advantage of this reporting process. This is simultaneously the weakness of this metric, as users are likely to map percentages onto intuitive percentage correct metrics such as, 90% and above is an A, 80% and above is a B or, worse, to the percentage proficient metric.[16] Users who do not have full appreciation of the contents of the market basket will not appreciate the scale.

Percentile-based reporting uses one or more of the 10th, 25th, 50th (i.e., the median), 75th, and 90th percentiles (e.g., National Center for Education Statistics, 2011, p. 9). Percentiles are reported on the NAEP score scale and are particularly useful for describing trends for students at lower or higher points in the distribution. Additionally, graphs of different percentiles over time can communicate trends in the variability in distributions over time. Inferences about trends at different achievement levels are sometimes inadvisably addressed by

[14] Conversely, the *NAEP Questions Tool* might be configured to show the probability of a correct response to any item by students at any given scale score.

[15] See http://www.nap.edu/openbook.php?record_id=10049&page=50

[16] In principle, of course, one might construct a market basket of items in such a way that 90% correct corresponded to "proficient," but as originally formulated, the intent was to have the market basket faithfully reflect the NAEP exercise pool. Even if one were able to map, say 90% correct to "proficient," it would be technically challenging to map simultaneously "Basic" and "Advanced" to additional points chosen in the market basket percent correct metric.

comparing trends in the percentages above Basic, Proficient, and Advanced achievement levels. This comparison suffers from the same distortion as all percentage-cut-metrics: trends for percentages closer to 50% will be magnified simply because more students happen to be close to the cut score. Percentiles are the correct alternative (Holland, 2002).

As Rothstein, Jacobsen, and Wilder (2006) argue, percentiles can be referenced to judgmental standards just as averages can, and policy goals can be expressed in terms like, "the average, median, or 25th percentile should surpass the cut score by a certain target horizon." The trends, gaps, and gap trends that arise from percentiles and averages are not subject to the same insidious distortions as the more familiar PAC metrics are. However, a superior approach to encouraging standards-based interpretations is one that does not rely on judgmental cut scores and instead fosters intuition about the score scale through item maps as noted above. A working example of percentile-based reporting is included in this section.

5.4 NAEP Inclusion Policies and Reporting of Full/Expanded Population Estimates

Current NAEP protocols allow for accommodations for students with disabilities and English language learners. Students who cannot participate meaningfully in the assessment, even with accommodation, are excluded from the assessment. Exclusions and, more specifically, varying exclusion rates over time, between groups, and between states, represent a serious threat to the most popular uses of NAEP scores, including the reporting of trends, gaps, gap trends, and state comparisons.

A 2010 policy statement from NAGB sets a target of 95% overall inclusion and a goal of 85% inclusion for both English language learners and students with disabilities. Any sample that falls short of this goal will have its scores flagged prominently in reports. This policy will tamp down cross-state differences in inclusion policies as well as future differences over time, but the shift in policy will have a particular influence on trends over the phase-in period. The so-called full population estimates or expanded population estimates are a useful tool for investigating this influence, as they use missing data procedures to impute excluded scores. The panel is reassured by the small magnitudes of difference between these adjusted trends and the observed trends.

However, the panel recommends strongly that the reporting of expanded population estimates continue and in a manner more accessible to secondary researchers. Although the inclusion goals are admirable, they continue to leave room for variability across states, across time, and, more worryingly, across groups within a state, where there may be disproportionate inclusion of some subgroups. Particular subgroups can drop below inclusion targets without violating policy guidelines, which will threaten estimates of gaps and gap trends. A near-term recommendation is to extend the research on exclusion bias to commonly reported gaps and gap trends. A longer term recommendation is to make the expanded estimates available for as many years and subgroups as possible and incorporate them as an option for reporting in the NAEP Data Explorer. These recommendations arise from the understanding that NAEP data support a wide variety of analyses that cannot be fully anticipated, and nonzero and variable exclusion rates may threaten some analyses more than others.

5.5 Small Subgroup Reporting

Along the same lines, NAEP is being increasingly asked to report on small subgroups. Examples include both regional/jurisdictional groups for which adequate samples are difficult to construct, and groups in parts of the proficiency distribution that are difficult to measure. Sometimes these two demands intersect, as when a particular region or jurisdiction is associated with either extreme of the proficiency distribution.

The panel recommends further investigation of methods for producing small-group reports within the confines of NAEP's design and legal restrictions. Some of these issues can be addressed with modifications of survey and assessment design. For example, difficult to sample subgroups can be better represented with oversampling methods. And extremes of the proficiency distribution can be better measured by targeting blocks to those

regions, as with KaSA and block adaptive testing. When these options are not available, post-hoc adjustments such as small area estimation methods may be possible. Knowing in advance that such methods might be used should also influence design of the survey and assessment components of NAEP.

While some demands for small group reporting can be met with the above or similar methodologies, without undermining core NAEP roles and responsibilities, others cannot. The goal of further investigations in this area would be to carefully delineate the conditions under which new small group reporting could be done, and those under which it could not.

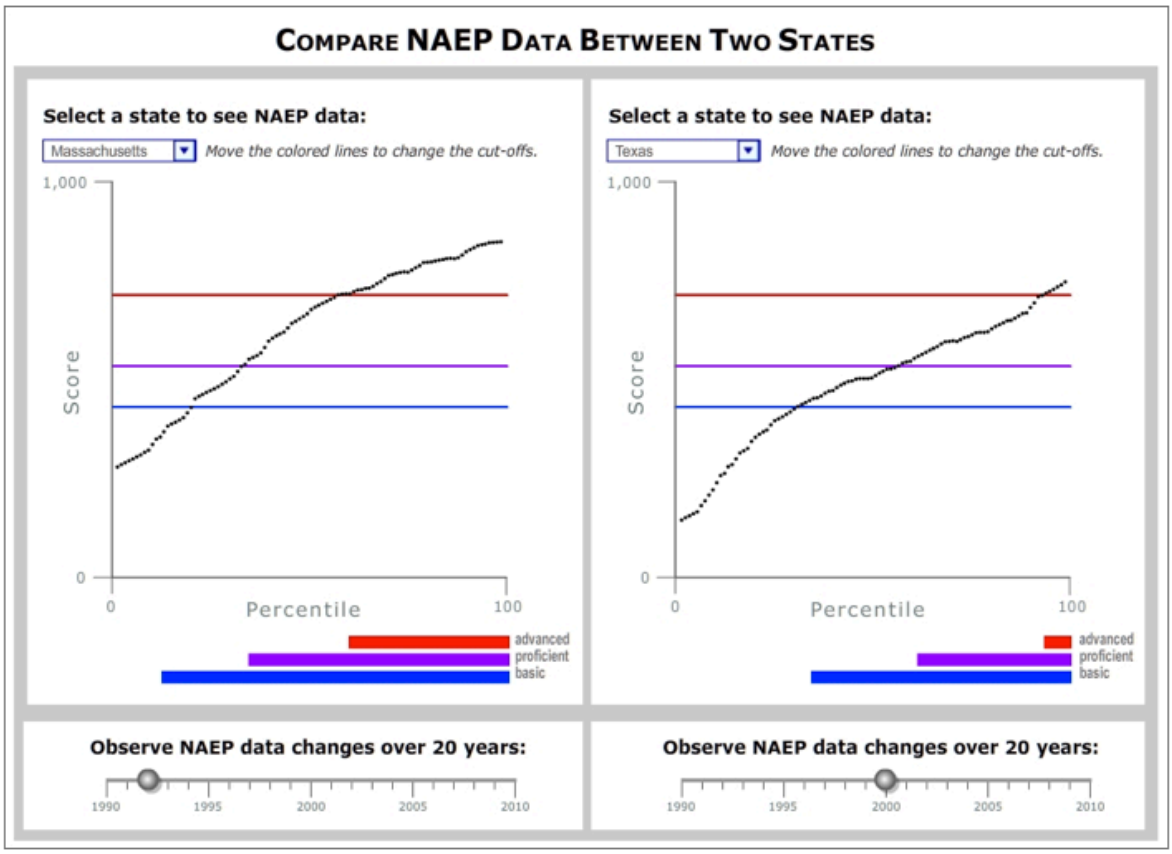
5.6 "Active" Reporting

Just as we assume that in the future all NAEP assessments will be delivered electronically, we also expect that the reports that NAEP produces will increasingly be viewed on a computer. And just as the change from paper to screen will make possible innovative assessment items, the evolution of dissemination technology is likely to induce major changes in how NAEP reports are compiled, used, and interpreted.

An online version of a report should not simply be a digital copy of the paper version, any more than an online test should be constrained to a static question-and-answer format. Even today the NAEP Data Explorer provides powerful data-mining tools that stakeholders can use to examine trends, gaps, and gap trends. We suggest that in the future such tools will be embedded in the NAEP reports themselves in the form of interactive graphs, contextualized and customized for particular purposes and to illustrate specific points. An obvious and natural example is the use of animation to represent changes over time. Where a static report might employ a graph to represent the trend of a set of percentile averages, an interactive graph might substitute time for position along the x-axis, freeing up that dimension to represent additional information.

Earlier we pointed out the problem with using percent proficiency above cut score measures to make comparisons between cohort performances. The reason, we suggest, that such measures are so popular is that they simplify a set of complicated data into a single, seemingly straightforward and relevant number. The fact that the measure is a woefully incomplete characterization of the data, easily leading to misguided policy choices, has not been sufficient to discourage its widespread reporting and use (Ho, 2008). The fundamental problem is that assessment data are complex by nature and do not lend themselves to easy compression. In mathematical terms, student scores form a distribution and distributions are only imperfectly represented by statistics.

We suggest that interactive, manipulable graphs embedded in NAEP reports may help readers understand the data at a deeper level than is commonly achieved by static representation. We provide a simple illustration of this general point on the next page.



The graphs represent NAEP data[17] from two different states, which can be selected from a drop-down menu. In both cases the horizontal axis is the percentiles and the vertical axis is the average score for each percentile. Thus, for example, the height of the left-most dot is the average score for the first 1% of students in each state. The three colored horizontal lines (red, purple, and blue) are at the positions of the cutoff scores for the Advanced, Proficient, and Basic achievement levels, respectively. Below the horizontal axis of each graph are three line segments, coded to the same color as the horizontal cutoff lines. These lines extend from the point at which the students' scores cross the respective cutoff scores all the way to the 100th percentile; they therefore represent the percentage of students in each of the three categories. Note that in the interactive version of the graph the cutoff scores can be changed by moving the horizontal lines up and down. The default position for each line, however, is the score adopted by NAEP for that category.

Below each graph is a slider representing time.[18] The reader can use this slider to examine data from different years or alternatively can animate the graph and have time go by automatically at a pre-determined rate. As this happens, the dots move smoothly[19] up and down, the percentage of dots above the cutoff scores also varies, and the line segments at the bottom of the graph change their lengths.

[17] The data and score scale used are fictitious. NAEP scores are reported on a scale from either 0 to 300 or 0 to 500 depending on the assessment.

[18] For some purposes we may want to provide a single slider that controls time for both graphs, or enable the user to link the two sliders so that the graphs always show data from the same year.

[19] We interpolate between scores in order to make the motion of the dots smooth, even though the data is typically collected only every two years. While this is in some sense a "cheat," in that the interpolated data is not "real," it is no more of a cheat than the common practice of drawing lines between discrete data points on a static graph.

These line segments are, of course, nothing more than a bar graph, albeit somewhat re-arranged, of the percentage of students scoring at or above the three achievement level categories. In order to drive home this point, we might consider animating a process by which the line segments fly up, turn 90 degrees, and land next to each other on the horizontal axis of a separate graph. Animating this graph will illustrate the time behavior of the proficiency scores. Interstate differences could be illustrated similarly. By altering the cutoff scores, users could see for themselves why the "percentage proficient" measure is unreliable and potentially misleading.

5.7 NAEP Reporting and the Common Core State Standards

The CCSS in ELA and mathematics have been adopted by 46 states and the District of Columbia. A similar national effort toward the adoption of common standards for science is well underway. In addition, two multistate consortia, PARCC and SBAC, are developing comprehensive assessment systems aligned with the CCSS, which are to be available by 2015. These new initiatives have raised questions as to how, or even whether, NAEP will continue to be relevant or necessary.

In response to these challenges, it is important to distinguish NAEP's purposes and design features from those of the consortium assessments, and to distinguish NAEP's assessment frameworks from the CCSS. NAEP and the CCSS-aligned consortium assessments are complementary to one another, meeting different information needs within an evolving ecology of educational assessments.

Most fundamentally, the CCSS is designed as a framework for curriculum and instruction. It represents a particular set of choices as to scope and sequence for K-12 education. NAEP assessment frameworks are emphatically not intended to direct curriculum and instruction. Instead, they serve as blueprints for the creation of assessments that broadly sample their respective subject matter domains, encompassing multiple potential curricular choices. This breadth of coverage equips NAEP to reveal changing profiles of achievement within a content area as curriculum changes. As one example, the long-term trend component of NAEP offers trend lines in mathematics from which the main NAEP mathematics trend lines have diverged over time. These distinct trends demonstrate that mathematics achievement is multidimensional. Curriculum and instruction have evolved, and students today are learning more of some new things, but less of some old things. It has only been possible for NAEP to document this effect because the long-term trend and main NAEP item pools together represent a range of earlier and more recent learning objectives. As Lee J. Cronbach, one of the original architects of NAEP, long ago observed, "If you wish only to know how well a curriculum is achieving its objectives, you fit the test to the curriculum; but if you wish to know how well the curriculum is serving the national interest, you measure all outcomes that might be worth striving for" (Cronbach, 1963, p. 680).

Now, however, conceptualizing NAEP's role vis-à-vis the consortia is a highly speculative endeavor at best. A critical question, as yet unanswered, is which assessment(s) will be the most innovative and will best represent the higher level thinking and application skills called for in the CCSS. When the U.S. Department of Education first set aside stimulus monies to fund the development of the next generation of assessments, the original hope was that an intensive research and development program would lead to major breakthroughs in assessment design. Given the large number of states involved in each consortium, however, and the very short timelines for fielding operational assessments, it now appears likely that apart from computer-based administration, the SBAC and PARCC assessments will bring incremental, not radical changes and improvements.

If the consortium assessments fulfill their promise and provide more ambitious representations of the intended achievement domains than the current NAEP, then NAEP will be playing catch up. If they fall short, then greater investments will be called for to make NAEP the locus for next-generation assessment development—a role it played decades ago. In either case, the development of new NAEP frameworks in each content area should explicitly attend to the overlap and unique components among several competing frameworks, including the frameworks implied by each consortium, the original intentions of the CCSS, international frameworks, and the existing NAEP frameworks. This is not to say that new frameworks should be an

incoherent blend of all of these. Rather, assuming that the CCSS will be successful as the basis of many state assessments going forward, exercise pools should be designed explicitly to retain those elements that are needed to track progress toward old goals as well as new goals and to report on important differences between these two views of progress.

Maintaining NAEP assessment frameworks that are broader than the CCSS is also critical to the central role we envision for NAEP within a network of linkages among large-scale assessments. Content alignment with multiple large-scale assessments, both within the United States and around the world, will require content sampling beyond the particular scope and sequence chosen for the CCSS.

Along with the fact that NAEP assessment frameworks are broader than the CCSS and designed for different purposes, the design for NAEP administrations enables broader sampling. The SBAC and PARCC summative assessments will be either fixed forms, designed to be administered in their entirety to each examinee at a given grade level, or adaptive tests that present different items to different examinees for the primary purpose of improving measurement precision. Under either of these approaches, the goal of maximizing comparability among individual students' scores is at odds with the goal of broad content sampling. In contrast, NAEP exercise pools are much larger for each subject and grade assessed. A matrix sampling design is employed, whereby each student responds to just a small portion of the available items. Responses across students are then used to construct a much fuller portrait of achievement distributions for groups of examinees.

Beyond these points, we note the obvious fact that the consortium assessments and NAEP assessments cover distinct, only partially overlapping subject-by-grade configurations. NAEP covers many more subject areas than the consortia will be able to assess by 2015, testing in-depth at just three grade levels (4, 8, and 12), whereas the SBAC and PARCC assessments will cover all grade levels from 3 to 8, as well as one grade in high school.

Historically, NAEP has served a critical "audit" function, offering an extremely helpful reference point in the interpretation of score trends on "high-stakes" tests used for school accountability. The main NAEP scales have served this function well even though high-stakes state assessments were not always closely aligned with the corresponding NAEP assessments. The frequently observed pattern of greater score gains over time on high-stakes tests versus NAEP may be attributable in large part to the reallocation of instructional time and resources toward just those content elements appearing on the high-stakes tests (Koretz, 2008). This "audit" function for NAEP, and the corresponding pattern of typical findings, further reinforces the importance of independent national assessment all-encompassing frameworks that reach beyond the CCSS.

For the reasons just presented, we quickly rejected the notion that the CCSS might replace NAEP assessment frameworks in ELA and mathematics. A more difficult question was whether NAEP should incorporate a new reporting scale designed to match the CCSS at each grade assessed. Such scales initially appeared attractive in the light of anticipated interest in direct comparisons between NAEP and the SBAC and PARCC summative assessments. Upon further reflection, however, we concluded that such scales would probably not be helpful. The CCSS is not adequately specified to support a single, definitive interpretation. The SBAC and PARCC assessments, each aligned to the CCSS, will nonetheless differ from one another in important respects, and the degree to which they will be directly comparable is not yet known. If NAEP fielded a third, competing interpretation of the CCSS, it would probably not agree with either of the other two. Rather than proposing new CCSS scales within NAEP, we propose the development of mechanisms for flexible linking of NAEP to other scales. This would include reweighting of content within NAEP if necessary, so as to maximize alignment with any of a range of large-scale assessment programs, including the SBAC and PARCC summative assessments as well as PISA, the Progress in International Reading Literacy Study (PIRLS), TIMSS, and others.

5.8 A General Approach to Reporting and Design

Historically, the development of assessments for NAEP has been primarily driven by content and the frameworks. Statistical and psychometric analysis has entered into the development process primarily to discard extremely poor items and to assemble blocks and booklets with adequate measurement properties, subject content, and framework criteria. This has worked in the past in part because item writers are well-trained in writing items that tend to scale with IRT models and in part because the reporting demands for NAEP did not depend on precise measurement across a broad range of proficiencies or subpopulations.

However, in recent years, increasing focus on reporting in a variety of proficiency ranges and subpopulations has taxed the current development methodology. For example, it is difficult to estimate proficiency distributions well in certain low-performing subgroups, with currently assembled assessments, so it is difficult to measure improvement in those groups. In addition, measuring complex multidimensional proficiencies that are analytically defined in terms of component knowledge and skills is inherently more demanding than measuring numerical position on a one-dimensional scale. On the other hand, concern has been expressed that too many resources are spent on measuring at the high end of the scale—as a rule, fewer than 5% of students score at the Advanced level.

With these ideas in mind, the panel believes that the time has come to augment content guidelines with statistical and psychometric guidelines to help ensure measurement precision sufficient for the reporting demands now being made on NAEP. Content and quantitative guidelines would work together to guide the development of items, blocks, booklets and assessments.

A fruitful plan for developing these guidelines is to approach them with the end in mind. Begin by carefully identifying and clarifying reporting goals (e.g., a list of main and subgroup reporting variables together with desired standard errors for each), and types of assessment activities (test items, more complex tasks for individual or collaborative work, etc.). Then let these decisions drive macro design issues such as assembly of blocks into booklets, booklet length, distribution of booklets to students using a PBIB or more adaptive design, etc. Practical requirements at this macro level (e.g., testing time for each student) would help determine micro design decisions such as standard error of measurement targets for assembling blocks and booklets, item selection strategies, etc.

This approach to designing NAEP assessments will be crucial to integrating the many suggestions made in this report, as well as other demands being made of NAEP, into a coherent national survey of educational progress in the future. Studying ways to make an approach like this feasible for NAEP is critical to its design, reporting, and relevance in the new assessment ecosystem.

6. Summary and Conclusions

The panel strongly affirms the continuing importance of the National Assessment of Educational Progress. Indeed, in this time of transition, with the adoption of the CCSS, the new science content framework, and the development of promising new accountability assessments by the PARCC and SBAC consortia, as well as the DLM, NCSC, and WIDA consortia, the need for NAEP is greater than ever before. NAEP is something to hold steady to, an essential resource uniquely capable of showing whether current and planned innovations in curriculum, in teaching and learning activities, and in accountability testing actually bring hoped-for improvements in learning outcomes. Long experience has shown that high-stakes tests themselves cannot be relied upon as the sole indicators of educational progress, making NAEP's low-stakes "audit" function indispensable. Beyond its value as a low-stakes assessment program maintaining stable trend lines over decades of time, NAEP is also more important than ever precisely because it is not fully aligned with the CCSS. As discussed in this report, reading, mathematics, and other broad subject matter proficiencies can be defined in multiple ways, and assessments based on alternative construct definitions will often show different trends over time. Even the PARCC and SBAC summative assessments may be expected to diverge over time, despite their common reliance on the CCSS.[20]

Several of our recommendations, especially in the area of new technology, are driven by the observation that schooling has become more complex and more heterogeneous than it was in the days when the foundations of the present NAEP program were established. The NAEP design from that era drew questions posed to all 9-year-olds, for example, from a common exercise pool. Curricular variations might equip some better than others to respond to those questions, but apparently little attention was paid to the possibility that different children's learning environments might be so diverse that different sets of exercises would be needed for valid assessment of those in one sort of classroom versus another. Since that time, there have been some forays into NAEP testing of special populations, but NAEP has generally not focused on important learning trends represented in significant numbers of classrooms though not yet universal. Today and increasingly for the future, the panel sees a need for NAEP to target some assessments in a more focused way. In part, this can be done with conventional adaptive testing approaches, extending the range of accurate measurement into the lower and upper tails of the achievement distribution. But where some classroom environments provide qualitatively different affordances for learning, more sophisticated targeting may be required for assessment tasks aligned with those new affordances.

[20] Suppose that after the PARCC and SBAC assessments had been in use for a few years, samples of students from PARCC states and from SBAC states were each administered both assessments. It would be expected that each group of students would do relatively better on whichever assessment had been high-stakes for them or their schools over the past several years (see, e.g., Koretz, 2003; Linn, 2000). But that implies that PARCC-SBAC linking functions derived using these two student subpopulations would differ. This is sufficient to show that the two assessments would by that time be measuring distinct constructs.

6.1 Recommendations

6.1.1 Need for care and caution in redesigning NAEP

In pondering NAEP's future, the panel affirmed the values of both continuity and innovation. Over the past 40 years, the machinery of the assessment has become extremely complex as NAEP has adapted to serve multiple purposes. This complexity may not be a problem per se, but it does imply that future changes and improvements must be thoughtfully considered and carefully implemented. Changes in one area will entail changes elsewhere. There are virtually no nontrivial changes to the core NAEP functions of item development, test assembly, sampling, administration, scoring, analysis, and reporting that we could confidently recommend without some preliminary study. Accordingly, many of our recommendations take the form of R&D proposals to be prioritized by the Innovations Laboratory. NAEP needs to be more nimble, but pilot studies and bridging studies will still be essential. Inevitably, the annual rhythms of the school calendar will sometimes limit the speed with which new innovations can be implemented.

The August 2011 NAEP Summit, the January 2012 NAEP Summit for state education agency representatives, further advice we have received from various sources, and of course our own deliberations have generated a long menu of potentially valuable NAEP initiatives and innovations. We have sifted these, setting some aside and discussing others. However, even the more limited set of new activities alluded to in this report would be far too numerous to undertake all at once. As presented in Section 5.8, we propose that a principled approach to determining priorities may be to begin with the most urgent reporting needs, and work backwards, focusing on those design changes and research initiatives required to support intended inferences from NAEP scale scores. One possible set of near-term priorities is suggested in Section 6.2, concerning topics for the Innovations Laboratory.

6.1.2 Infrastructure recommendations

The considerations set forth in Section 6.1.1 do suggest one clear, immediate recommendation, on which virtually everything else depends. That is to expand NAEP R&D. We recommend the establishment of a NAEP "Innovations Laboratory" (IL) to strengthen and systematize NAEP R&D, maintain a coherent overview of both in-house and third party studies, and provide an access hub for vetting new ideas. The IL we propose will require significant start-up funding and a commitment to stable future funding so that longer-term as well as short-term studies can be planned and executed in an orderly and efficient way. We envision the IL housing surveys of new instructional practices (e.g., appropriation of new technologies by schools into their classrooms) both in the United States and abroad; monitoring of the quality, comprehensiveness, and comparability of information available from states' education data warehouses; theoretical work on psychometric models to better accommodate multipart items with branching structures or adaptive scaffolding; cognitive labs and field tests to study new item types (e.g., including the use of interactive simulations and models); continued research on automated scoring of constructed response items that allow for highly diverse responses; development of new, dynamic reporting tools; and studies with user groups to better design NAEP reports that communicate clearly and minimize misinterpretation. These and many more activities will vary in their timelines, budgets, staffing requirements, and in their probabilities of success. Apart from all these studies themselves, funding will be required to staff and manage the IL, set priorities, monitor research activities, disseminate findings, and assure that dollars are well spent.

Our second recommendation concerning NAEP infrastructure is to investigate a method for maintaining LTT without a separate assessment component in NAEP administration. The verb "investigate" is important here. Elsewhere, we describe several changes that together might make this possible, but the short-term proposal is to study these changes, not to implement them immediately. If supported by IL studies, the NAEP exercise pools might be modestly expanded beyond coverage of material in the assessment frameworks, making room for LTT content in the main NAEP

data collections. Suitable bridging studies, also overseen by the IL, would be required to account for differences in sampling frames and administration conditions between main NAEP and LTT. If realized, this change could significantly increase the precision of contrasts between main NAEP and LTT trend lines, because they would be estimated using common student samples. This change would also simplify comparisons of main NAEP versus LTT and make the LTT component more useful for secondary analysis (cf. Beaton & Chromy, 2010).

6.1.3 Assessment framework recommendations

Our first recommendation concerning NAEP assessment frameworks is the creation of standing committees of content experts. Historically, ad hoc framework committees have been convened to develop frameworks and then disbanded, in effect handing off their framework to another group that would then devise test specifications that, together with the framework, would then inform item development. Following item tryout and revision, exercise blocks and test forms would then be assembled. Through this process, there can be some loss of fidelity to the framework committee's original vision. Under our proposal, standing committees would review field test data, for example, and be aware when "after-the-fact" distortions of the intended domain occur because more ambitious item types fail to meet statistical criteria. Standing committees could also update assessment frameworks incrementally, at the same time assuring that the constructs underlying NAEP reporting scales did not drift to the point where new trend lines were indicated. In particular, assessment frameworks would be updated to accommodate changing learning environments. Inquiries with dynamic knowledge representations and simulations in science would be one example.

We also recommend that content guidelines should be augmented routinely with statistical and psychometric guidelines, to help ensure adequate measurement for the reporting demands now being made on NAEP. Content and quantitative guidelines should work together to guide the development of items, blocks, booklets and assessments. Content and psychometric guidelines might best be integrated if the standing committees of content experts also included one or more members with psychometric expertise. It is likely that in addition to assessment frameworks, additional test specifications would still be required, but we believe validity would be enhanced if the processes of developing content guidelines and psychometric guidelines were developed together.

A third, related recommendation does not address assessment frameworks per se, but does underlie several of the recommendations presented elsewhere in this report. We recommend a modest expansion of NAEP exercise pools, beyond what is specified in the assessment frameworks. The main NAEP, state NAEP, and TUDA reporting scales would continue to reflect as faithfully as possible the content described in the frameworks. Assessment frameworks would continue to define the achievement constructs NAEP was intended to measure. However, making more room for the administration of additional items would further several related goals. First, additional content could assure adequate representation of material now included only in the LTT component. Second, additional content in the reading and mathematics assessments could help assure full coverage of the CCSS. The goal would be to include at least some items resembling those found in both the PARCC and the SBAC summative assessments. To the extent possible, any elements of the CCSS omitted from the consortium assessments would also be addressed. Third, additional content could be included to improve alignment with other large-scale assessment programs for which linkages to NAEP were deemed worthwhile. This is, of course, not an entirely new idea. Additional exercise blocks have been included in NAEP assessments by way of tryouts and probe studies. We are recommending that this practice be expanded to accomplish a broader set of goals.

6.1.4 Technology recommendations

As stated earlier, virtually every aspect of NAEP is affected, directly or indirectly, by changing technology. Here, we focus primarily on the challenges NAEP faces in measuring and reporting the effects of new technologies on what, where, when, how, and why children learn. One orienting question is: What should NAEP lead with that is not being done universally in American schools today? It

is a particular challenge, of course, for NAEP to monitor and report on emergent schooling practices, precisely because they are not being done universally in American schools.[21] Doing so will require targeted sampling of schools where significant new technologies are being used. We suspect that NAEP-produced evidence of technology trends that appear to have promise can both accelerate their growth in practice and spur the policy and research communities to deepen their efforts to get rigorous and more conclusive evidence of effectiveness.

Today, for example, we assume universal use of paper texts in reading and mathematics instruction; we do not assume the same universality for e-texts or science simulation inquiry environments. NAEP studies reporting on the achievement of students in classrooms where these affordances are available might use some technology-mediated exercises that would be inappropriate for students who did not have the same learning opportunities. Conversely, it might be impossible to document some aspects of these students' achievement without using exercises based on the learning tools with which they were familiar. Note that this is not a recommendation for curriculum evaluation or for studies of the effectiveness of particular interventions. Clear criteria would be required for deciding when a new technology had reached a point where such study was warranted.

We cannot predict with certainty which of the areas we have outlined are likely to be among the first taken up in schools (or out-of-school contexts) and shown to be productive in fostering higher levels of learning. Subject of course to priorities established by NAEP's governance, these would be the most likely near-term targets for NAEP R&D. Based on marketplace trends, however, we consider user interface modalities of gesture/touch and voice to provide near-term priorities for R&D, with visual recognition/augmented reality and sketching a longer-term horizon.

A second key area of application for new digital media and new communication modalities is technology-enhanced accommodations for specific student subpopulations. NAEP must strive to ensure fair assessments for learners marginalized in many educational settings: students with disabilities, students from low-income communities and minorities, English language learners, students who are gifted and talented, students from diverse cultures and linguistic backgrounds, and students in rural areas (e.g., see the National Education Technology Plan, 2010, section on Universal Design for Learning).

Further technology-related recommendations are included in Section 6.2, concerning topics for the Innovations Laboratory.

[21] Students learning via new interactive media may best display their understanding using familiar technology tools. Items best suited to these students may be inappropriate for students in more conventional classrooms. Just as advanced problem-solving items may fail on psychometric criteria if very few children can answer them correctly, so technology-based items may appear defective if they are used inappropriately with unprepared students. For the most part, the current NAEP design relies on items suitable for near-universal administration at a given grade level.

6.1.5 Reporting recommendations

Reporting demands should drive NAEP design. NAEP reports are the most visible products of the entire NAEP enterprise. Recognizing the centrality of NAEP reporting, the panel recommends that NAEP R&D priorities be determined by starting with desired reporting scales and reporting subpopulation definitions and then working backwards to design the items, sampling plans, and other aspects of data collection and analysis required to meet reporting requirements (see Section 5.8). In particular, small-group reporting demands should be taken into account in the NAEP design. As discussed in Section 5.5, when researchers or policymakers turn to NAEP for information about small demographic groups, samples are often insufficient to provide precise information. Anticipating and satisfying such specialized reporting needs is a continuing challenge. When future policy or research questions can be anticipated, small groups of potential interest may be oversampled. Small area estimation methods may prove useful. Adaptive testing may improve precision when small groups are located largely in a tail of the achievement distribution.

NAEP reporting scales should be better tied to assessment frameworks. Historically, assessment frameworks have served as the blueprints for NAEP assessments; in effect defining the intended constructs underlying NAEP reporting scales. In practice, however, the exercises ultimately included in NAEP data collections have sometimes fallen short of frameworks' specifications (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007). Several of the panel's recommendations elsewhere in this report are intended to strengthen the connection between NAEP assessment frameworks and NAEP reporting scales. First, as described in Section 3.2.2, standing subject-matter panels should be involved not only in the initial creation or updating of assessment frameworks, but also in reviewing items and the content balance of the entire set of exercises on which reporting scales are based. Second, as described in Section 5.8, statistical and psychometric guidelines should be developed in tandem with content guidelines.

Much less emphasis should be placed on achievement levels. For the past two decades, achievement levels—Basic, Proficient, and Advanced—have played a central role in NAEP reporting. Achievement level reporting was developed by the National Assessment Governing Board in response to their interpretation of a statutory mandate. Reporting of percentages of students at or above score levels labeled Basic, Proficient, and Advanced appears to answer questions about the proportion of students meeting expectations or doing well enough. Achievement level reporting is well intentioned and entirely understandable, but the panel concurs with a long series of reviews and evaluations that find it seriously problematical. In response, we offer three recommendations.

First, beginning immediately, NAEP reports should more clearly explain the limitations of achievement level reporting. Caveats concerning the interpretation of gaps, trends, and (most problematically) gap trends should be included, and these interpretations should be strongly discouraged, because achievement-level statistics depend in complex ways on the setting of the achievement levels. Second, current alternatives to achievement level reporting should be featured more prominently in NAEP reports. These include the use of scale score means and standard deviations of gaps, trends, and gap trends; reporting and comparison of scale scores representing chosen percentiles of score distributions; scale anchoring; item maps; and reports of performance on selected released items. The NAEP Questions Tool could be a useful resource in helping NAEP stakeholders understand what students in a given region of the score scale in fact know and are able to do. Third, through the Innovations Laboratory, studies should be undertaken to develop alternative reporting methods, including both new methods and improvements to the existing methods just mentioned. Among new methods, one promising possibility is market-basket reporting, as discussed in Section 5.3 of this report. Over time, if NAEP reporting scales become more broadly familiar, the attention paid to achievement levels may diminish.

Active reporting, including dynamic data visualization tools, should be incorporated into non-print versions of NAEP reports. As the dissemination of information moves from print media to computers and tablets, there are increasing opportunities to enhance NAEP reports with embedded tools for customizable, dynamic displays

of information. The NAEP Data Explorer provides some excellent tools already, but only a small fraction of the audience for NAEP reports is likely to access these tools. If dynamic displays of information were directly available in online versions of NAEP reports, they would be more widely used and appreciated. As discussed in Section 5.6, "active" reporting should be a priority for the Innovations Laboratory.

Full or expanded population estimates should be readily available. As discussed in Section 5.4, NAEP should continue to strive to include as large a proportion of eligible students as possible. The panel is hopeful that common data definitions, state comparisons within each of the multistate consortia, and increased data sharing will encourage moves toward greater uniformity in definitions, rules, and procedures for exclusions and accommodations, including definitions of students with disabilities and of English language learners. NAEP governance and NAEP contractors should do whatever they can to expedite these trends. Nonetheless, the goal of 100 percent participation is unlikely ever to be attained. Thus, work is still needed on imputation methods enabling more accurate full- or expanded-population estimates. We recommend considering the possibility of making full or expanded population estimates available routinely as an option in tables generated using the NAEP Data Explorer.

Developing a capability for linking NAEP to other assessment programs is critical. As discussed in Section 4.5.1, the panel places high priority on Innovations Laboratory studies of NAEP design changes to facilitate linkages between NAEP and other large-scale assessment programs, including the summative assessments developed by the PARCC and SBAC consortia at grades 4, 8, and possibly 12. The recommendation to broaden NAEP item pools should enable fuller representation of assessment frameworks underlying other assessments. Bridging studies to adjust for differences in administration conditions, including differences in testing windows, will be required as well.

6.2 Topics for the NAEP Innovations Laboratory

Below is a partial list of illustrative topics for the proposed NAEP Innovations Laboratory to explore. Each has its own unique challenges and timelines, but would better equip NAEP to maintain relevance in the changing assessment and education ecosystem.

Investigating and assuring the validity of intended inferences from NAEP

- Develop a detailed model for "dynamic assessment frameworks," including protocols for routine tracking of drift in construct definition as incremental changes accumulate (e.g., linking back to earlier construct definitions)
- Conduct an in-depth comparison of NAEP assessment frameworks in mathematics and ELA to the CCSS
- Explore ways to enhance student motivation/buy-in (especially at grade 12)
- Use technology to better accommodate learners with special needs
- Incorporate "universal design" principles in assessments delivered using new technology platforms
- Conduct studies of new psychometric models to better accommodate pattern scoring as necessary to assess constructs of interest
- Conduct studies of the clarity and interpretability of NAEP reports, including alternative reporting metrics keyed to specific purposes
- Continue R&D on full/expanded population estimates

Improving NAEP processes to balance among reducing respondent burden, shortening reporting time, increasing precision, and reducing costs

- Develop procedures for more rapid vetting of potential items/item formats adapted from curriculum-embedded assessments
- Implement booklet-level two-stage adaptive testing (e.g., MCBS, KaSA)
- Conduct alignment and bridging studies to investigate the feasibility of folding the long-term trend NAEP into the main NAEP data collection
- Facilitate linkages to state data systems for: (1) efficient sampling; (2) aligning NAEP scales to state assessment scales; and (3) maintenance of integrated longitudinal data structures
- Conduct studies on innovative item types/new test administration platforms

Expanding the range of achievement constructs NAEP can validly assess

- Survey/monitor current school-based applications of new knowledge representations and input modalities through which learners can create or interact with knowledge representations
- Use trend-tracking and field scanning concerning new practices and tools used in other nations' K-12 educational progress measurement
- Explore links between NAEP and the Gates Foundation-funded "Shared Learning Infrastructure"
- Consider adaptive test administration sensitive to instruction context as well as student proficiency

Enabling NAEP to serve new purposes

- Design main NAEP data collection with expanded slots for: (1) linking items; and (2) experimental item types
- Conduct alignment studies comparing NAEP mathematics and ELA frameworks to the CCSS, as well as SBAC and PARCC test specifications (when available) so as to position NAEP to inform progress relative to the CCSS
- Develop linking protocols and procedures that current and future testing programs can use to relate their results to NAEP results, perhaps with reporting on a NAEP scale
- Conduct bridging and linking studies on schedules of other major assessments, to support cross-assessment linkages
- Conduct feasibility studies of NAEP assessments reaching down to lower grade levels (even as low as pre-kindergarten)
- Assess students in two- or four-year colleges
- Assess home-schooled children
- Conduct feasibility studies for linkages to State Data Warehouses, or the NCES Schools and Staffing Survey (SASS)
- Conduct feasibility of changing NAEP sampling to facilitate longitudinal tracking, as opposed to one-time achievement snapshots

References

- Ainsworth, S., Honey, M., Johnson, W.L., Koedinger, K., Muramatsu, B., Pea, R., Recker, M., Weimar, S. (2005). Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda. Computing Research Association. Retrieved February 14, 2012 from http://cra.org/uploads/documents/resources/rissues/cyberinfrastructure.pdf.

- Bandeira de Mello, V., Blanksenship, C., McLaughlin, D. (2009). *Mapping state proficiency standards onto NAEP scales: 2005-2007* (NCES Report No. 2010-456). Washington, DC: Institute of Education Sciences. Retrieved March 26, 2012, from http://nces.ed.gov/nationsreportcard/pdf/studies/2010456.pdf.

- Beaton, A.E., Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 191-204.

- Beaton, A.E., & Chromy, J. R. (December 2010). NAEP trends: Main NAEP vs. Long-Term Trend (Commissioned by the NAEP Validity Studies [NVS] Panel). Retrieved March 18, 2012, from http://www.air.org/files/NAEP_Trends_12-6-10.pdf.

- Beaton, A.E., Linn, R.L., Bohrnstedt, G.W. (2012). Alternative approaches to setting performance standards for the National Assessment of Educational Progress (NAEP) (Commissioned by the NAEP Validity Studies [NVS] Panel). Retrieved March 23, 2012, from http://www.air.org/files/NVS_Achievement_Levels_Paper_Final.pdf.

- Beckwith, R., Theocharous, G., Avrahami, D., Philipose, M. (2010). Tabletop ESP: Everyday sensing and perception in the classroom. *Intel Technology Journal*, *14*(1), 18-33.

- Bennett, R.E., Persky, H., Weiss, A., Jenkins, F., et al. (2007). Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project, research and development series (NCES 2007-466). Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved March 21, 2012, from http://nces.ed.gov/nationsreportcard/pdf/studies/2007466.pdf.

- Bennett, R.E., Persky, H., Weiss, A., Jenkins, F. (2010). Measuring Problem Solving with Technology: A Demonstration Study for NAEP. *Journal of Technology, Learning, and Assessment*, *8*(8). Retrieved March 21, 2012, from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1627.

- Brown, W. (2000). Reporting NAEP by achievement levels: An analysis of policy and external reviews. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmations and improvements* (pp. 13-39). Washington, DC: National Assessment Governing Board.

- Chi, M.T.H. (2005). Common sense conceptions of emergent processes: Why some misconceptions are robust. *J. Learning Sciences*, *14*, 161-199.

- Council of Chief State School Officers (2008). Attributes of Effective Formative Assessment. Retrieved April 16, 2012, from http://www.ccsso.org/documents/2008/attributes_of_effective_2008.pdf.

- Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, *64*, 672-683.

- Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007, September). Validity Study of the NAEP Mathematics Assessment: Grades 4 and 8. (Conducted by the NAEP Validity Studies [NVS] Panel). Washington, DC: American Institutes for Research. Retrieved April 16, 2012, from http://www.air.org/files/Daro_NAEP_Math_Validity_Study.pdf.

- diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *Journal of the Learning Sciences*, *13*(1), 77-103.
- Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W., Hemphill, F.C. (1999.) Uncommon Measures: Equivalence and Linkage Among Educational Tests. Retrieved April 16, 2012, from http://www.nap.edu/openbook.php?record_id=6332.
- Forbus, K., Usher, J., Lovett, A., Lockwood, K., Wetzel, J. (2011). CogSketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, 3(4), 648-666.
- Glaser, R., Linn, R., & Bohrnstedt, G. (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. National Academy of Education.
- Glaser, R., Linn, R., & Bohrnstedt, G. (1992). *Assessing Student Achievement in the States: Background Studies.* Studies for the Evaluation of the NAEP Trial State Assessment Commissioned for the National Academy of Education Panel Report on the 1990 Trial. National Academy of Education.
- Glaser, R., Linn, R., & Bohrnstedt, G. (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. National Academy of Education.
- Glaser, R., Linn, R., & Bohrnstedt,G. (1994). *The Trial State Assessment: Prospects and Realities: Background Studies.* Studies of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1992 Trial State Assessment. National Academy of Education.
- Glaser, R., Linn, R., & Bohrnstedt, G. (1994). *Quality and Utility. The 1994 Trial State Assessment in Reading, Background Studies.* Studies of the National Academy of Education Panel on the Evaluation of the 1994 National Assessment of Educational Progress Trial State Assessment in Reading. National Academy of Education.
- Glaser, R., Linn, R., & Bohrnstedt, G. (1996). *Quality and Utility: The 1994 Trial State Assessment in Reading.* The Fourth Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1994 Trial State Assessment in Reading. National Academy of Education.
- Glaser, R., Linn, R., & Bohrnstedt, G. (1997). *Assessment in Transition: Monitoring the Nation's Educational Progress*. National Academy of Education.
- Glaser, R., Linn, R., & Bohrnstedt, G. (1997). *Assessment in Transition: Monitoring the Nation's Educational Progress.* Background Studies of the Final Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment. National Academy of Education.
- Heritage, M. (Feb 2008). *Learning Progressions: Supporting Instruction and Formative Assessment*. Council of Chief State School Officers: Washington, DC.
- Ho, A.D. (2008). The Problem with "Proficiency": Limitations of Statistics and Policy under No Child Left Behind. *Educational Researcher*, 37, 351-360.
- Ho, A.D., Haertel, E.H. (2007). Apples to Apples? The Underlying Assumptions of State-NAEP Comparisons. Paper commissioned by the Council of Chief State School Officers.
- Ho, A.D., Haertel, E.H. (2007). (Over)-interpreting Mappings of State Performance Standards onto the NAEP Scale. Paper commissioned by the Council of Chief State School Officers.

- Holland, P. (2002). Two Measures of Change in the Gaps between the CDFs of Test Score Distributions. *Journal of Educational and Behavioral Statistics*, 27, 3-17.

- Jenkins, H. (2009). *Confronting the challenges of participatory culture: Media education for the 21st century*. Cambridge, MA: MIT Press.

- Koedinger, K.R., McLaughlin, E.A., & Stamper, (2012, in press). Automated student model improvement. To appear in *Proceedings of the Fifth International Conference on Educational Data Mining*.

- Kitmitto, S. (Sept 2011). Measuring the Status and Change of NAEP State Inclusion Rates for Students with Disabilities: Results 2007-2009. Retrieved April 16, 2012, from http://nces.ed.gov/nationsreportcard/pubs/studies/2011457.asp.

- Klein, S.P., Hamilton, L.S., McCaffrey, D.F., Stecher, B.M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND.

- Koenig, J.A., Rapporteur, National Research Council (2011). Assessing 21st Century Skills: Summary of a Workshop. Washington, DC: National Academy Press.

- Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice*, *22*(3), 18-26.

- Koretz, D. (2008). *Measuring up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press.

- Koretz, D., Barron, S. I. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS). MR-1014-EDU. Santa Monica, CA: RAND.

- Koretz, D., Bertenthal, M.W., Green, B.F. (1999). Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests. Retrieved April 16, 2012, from http://www.nap.edu/openbook.php?record_id=9683.

- Koretz, D., & Deibert, E. (1993). Interpretations of National Assessment of Educational Progress (NAEP) Anchor Points and Achievement Levels by the Print Media in 1991 (Prepared for the National Center for Education Statistics). Santa Monica, CA: RAND.

- Levy, F., & Murnane, R.J. (2005). The New Division of Labor: How Computers Are Creating the Next Job Market. Princeton, NJ: Princeton University Press.

- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4-16.

- Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*(2), 177-194.

- Mislevy, R.J., Behrens, J.T., Bennett, R.E., Demark, S.F., Frezzo, D.C., Levy, R., Robinson, D.H., Rutstein, D.W., Shute, V.J., Stanley, K., Winters, F.I. (2010). On the Roles of External Knowledge Representations in Assessment Design. *Journal of Technology, Learning, and Assessment, 8*(2). Retrieved February 17, 2012, fromhttp://ejournals.bc.edu/ojs/index.php/jtla/issue/view/160.

- National Assessment Governing Board (March 6, 2010). NAEP Testing and Reporting on Students with Disabilities and English Language Learners: Policy Statement. Retrieved April 16, 2012, from http://nagb.org/policies/PoliciesPDFs/Reporting%20and%20Dissemination/naep_testandreport_studentswithdisabilities.pdf.

- National Assessment Governing Board (September 2010). Reading framework for the 2011 National Assessment of Educational Progress. Washington, DC: National Assessment Governing Board.
- National Education Technology Plan (2010). Transforming American Education: Learning Powered by Technology. Washington, DC: US Department of Education, Office of Educational Technology.
- National Institute of Statistical Sciences (July 2009). NISS/NESSI Task force on Full Population Estimates for NAEP: Final Report. Retrieved April 16, 2012, from http://www.niss.org/sites/default/files/tr172.pdf.
- National Research Council (2000). How People Learn: Mind, Brain, Experience and School. Washington, DC: National Academies Press.
- National Research Council (2009). Learning science in informal environments: People, places and pursuits. Washington, DC: National Academies Press.
- National Research Council (2011a). Learning science through computer games and simulations. Committee on Science Learning: Computer Games, Simulations, and Education; National Research Council. Washington, DC: National Academy Press.
- National Research Council (2011b). Report of a Workshop of Pedagogical Aspects of Computational Thinking. Committee for the Workshops on Computational Thinking, National Research Council. Washington, DC: National Academy Press.
- National Research Council (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Committee on Conceptual Framework for the New K-12 Science Education Standards, Board on Science Education, National Research Council. Washington, DC: National Academy Press.
- National Science Foundation Taskforce on Cyberlearning (June 2008). Fostering learning in the networked world: the cyberlearning opportunity and challenge. Arlington, VA: National Science Foundation.
- National Center for Education Statistics, Institute of Education Sciences. (2009). Mapping State Proficiency Standards. Retrieved April 16, 2012, from http://nces.ed.gov/nationsreportcard/studies/statemapping/.
- National Center for Education Statistics, Institute of Education Sciences. 2011 NAEP-TIMSS Linking Study. Retrieved April 16, 2012, from http://nces.ed.gov/timss/naeplink.asp.
- National Center for Education Statistics, Institute of Education Sciences. Statewide Longitudinal Data Systems Grant Program: Grantee States. Retrieved April 16, 2012, from http://nces.ed.gov/programs/slds/stateinfo.asp.
- National Center for Education Statistics, Institute of Education Sciences. (2012). Inclusion of Special-Needs Students. Retrieved April 16, 2012, from http://nces.ed.gov/nationsreportcard/about/inclusion.asp.
- National Center for Education Statistics, Institute of Education Sciences. (2011). The Nation's report card: Reading 2011 (NCES 2012-457). Retrieved March 18, 2012, from http://nces.ed.gov/nationsreportcard/pdf/main2011/2012457.pdf.
- Pellegrino, J.W., Jones, L.R., Mitchell, K.J. (Eds.). (1999). Committee on the Evaluation of National and State Assessments of Educational Progress. *Grading the Nation's Report Card*. Washington, DC: National Academy Press.

- President's Council of Advisors on Science and Technology (PCAST). (September 2010). Report to the President—Prepare and Inspire: K-12 Education in Science, Technology, Engineering, and Mathematics (STEM) for America's Future. Washington, DC: President's Council of Advisors on Science and Technology, Executive Office of the President.

- Rothstein, R., Jacobsen, R., Wilder, T. (2006). "Proficiency for all" – An oxymoron. In *Examining America's commitment to closing achievement gaps: NCLB and its alternatives*. Symposium conducted at the meeting of the Campaign for Educational Equity, New York, NY.

- Shute, V.J., Ventura, M., Bauer, M., Zapata-Rivera, D. (2009). Melding the Power of Serious Games and Embedded Assessment to Monitor and Foster Learning: Flow and Grow. (In U. Ritterfield, M. J. Cody, P. Vorderer, Eds.: *The Social Sciences of Serious Games: Theories and Applications*.)

- Vinovskis, M.A. (1998). Overseeing the nation's report card: the creation and evolution of the National Assessment Governing Board (NAGB). Retrieved March 18, 2012, from http://www.nagb.org/publications/95222.pdf.