NAEP Validity Studies White Paper: Revision of the NAEP Science Framework and Assessment

James W. Pellegrino
University of Illinois Chicago

October 2021
Commissioned by the NAEP Validity Studies (NVS) Panel

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U.S. Department of Education or the American Institutes for Research.

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

Panel Members:

Project Director:

Sami Kitmitto
American Institutes for Research

Project Officer:

Grady Wilburn
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS) Panel
American Institutes for Research
1333 Broadway Ave, Suite 300
Oakland, CA 94612
Email: skitmitto@air.org

CONTENTS

TABLES

FIGURES

OVERALL PURPOSE AND ORGANIZATION

The purpose of this white paper is to consider issues related to the scope and focus of a possible new framework for National Assessment of Educational Progress (NAEP) Science (hereafter, NAEP science), including its possible expansion to include aspects of what is represented in NAEP Technology and Engineering Literacy (TEL) (hereafter, NAEP TEL). The goal is to provide the NAEP Validity Studies (NVS) Panel and the NAEP program with input about possible directions for the future and the rationale for choosing among them. Five major sections comprise this paper.

Section I sets the stage for the sections that follow by providing brief background information about the history and projected future uses of the NAEP Science Framework and Assessment as well as the NAEP TEL Framework and Assessment. It also summarizes the National Center for Education Statistics (NCES) and the National Assessment Governing Board (NAGB) timeline for consideration of possible revisions to the NAEP science framework in anticipation of its use to guide the NAEP Science Assessment scheduled for 2028.

Section II contains information on analyses comparing the current NAEP science framework and the NAEP TEL framework to the overall science and technology framework and related set of standards that emerged in the United States in the early part of the last decade. The section begins with a brief synopsis of the content and focus of the NAEP Science and TEL frameworks followed by a brief synopsis of the National Research Council (NRC) *Framework for K–12 Science Education* (NRC, 2012) (hereafter, NRC framework) and the derivative *Next Generation Science Standards* (NGSS) (NRC, 2013). Following that, results are presented from an extensive study comparing the alignment between NAEP Science and NAEP TEL and NGSS (Neidorf et al., 2016). In doing so, the section also considers some of the implications regarding assessments aligned with each reference source.

Section III focuses on the status of science standards and assessments in individual states since the publication of the NRC framework and the NGSS. It reviews the current status regarding state adoptions of science standards that are either identical to NGSS or that are partially aligned with the NGSS (i.e., NRC framework and NGSS "alike"), as well as states with science standards that have no claimed alignment with either the NGSS or NRC framework. For those states with science standards that are NRC framework/NGSS alike, results are summarized from a study examining content alignment between those state standards and the NAEP science framework (Dickinson et al., 2021). The section also includes a summary of the status of the design and implementation of state science assessments relative to their currently adopted standards. This consideration is limited to states that have adopted the NGSS and those whose adopted standards are NRC framework/NGSS alike. The section includes a brief review of the status of the implementation of curricular and instructional practices in states relative to the NRC framework and NGSS. Results are based on the most recent (2018) National Survey of Science and Mathematics Education. The section concludes with a consideration of trends in NAEP science performance for the last 12 years and some possible implications for future NAEP science assessments.

Section IV provides a brief discussion of advances in technology as related to the assessment of science and engineering knowledge and skills. It considers how various developments in digital technologies should be considered in reviewing the existing NAEP Science framework and assessment and envisioning possibilities for their updating. Discussion focuses on the affordances of technology with respect to the constructs that could be included in a revised framework and the associated task design, data capture, and data analytic issues involved in an assessment aligned to an updated framework. The section concludes with a brief discussion of practical and equity concerns related to digitally based assessment of science and technology proficiency.

Section V contains a set of conclusions and recommendations as input to the NCES and NAGB process of reviewing the NAEP science framework and considering possible revision. Conclusions and recommendations are based on the major findings presented in the prior sections.

SECTION I: BACKGROUND, TIMELINE, AND INPUTS

Relevant History: NAEP Science and NAEP TEL

NAEP Science

NAEP science is based on a framework that was adopted in 2005 for the 2009 assessment (NCES, 2009, 2014). That framework was used for the 2015 and 2019 administration of science at grades 4, 8, and 12. It will be used once more for the 2024 (originally 2023) administration of science at eighth grade only. The 2028 (originally 2027) operational administration of the science assessment at grades 4 and 8 at the national, state, and large urban district levels is supposed to be based on an updated science framework.

NAEP TEL

The NAEP TEL assessment is based on a framework developed for grades 4, 8, and 12 in the 2011–2012 period for the 2014 assessment at grade 8. That framework was used for the 2018 TEL administration for grade 8. It will be used twice more for the 2024 (originally 2023) and 2028 (originally 2027) TEL administrations for grade 8. Both planned TEL administrations overlap with NAEP science administrations: 2024 overlaps with the current science framework and assessment, and 2028 overlaps with the new science framework and assessment.

NAEP Science and TEL—Possible Merger

Discussions have been held within NAGB about possibilities for combining NAEP science and TEL, especially because both are now digitally based assessments. Doing so may make logical sense given overlaps in conceptual coverage with contemporary U.S. science and technology frameworks. Another benefit could be cost savings realized by having a single assessment representing key aspects of knowledge and skill for science and technology. Such a merger clearly would be most beneficial for the planned 2028 administration of both science and TEL. NAGB therefore may wish to consider developing a single 2028 assessment based on a new integrated science and technology framework.

Status and Plans for Review, Update, and/or Revision of the NAEP Science Framework

NAGB has started the process needed to consider updating the science framework for application in the design of the 2028 grades 4 and 8 science assessment. Given the current timeline, it appears that a decision about the need for and the scope of a science framework revision will be completed during 2022. Work toward making such a decision includes:

- Detailed information available in an NCES report issued in 2016 titled *A Comparison Between the Next Generation Science Standards (NGSS) and the National Assessment of Educational Progress (NAEP) Frameworks in Science, Technology and Engineering Literacy, and Mathematics* (Neidorf et al., 2016). Information about the results of this study is presented in Section II.
- A recently completed study by HumRRO titled *Comparative Analysis of the NAEP Science Framework and State Science Standards* (Dickinson et al., 2021) in which content overlap was examined between the NAEP science framework and the science

standards of individual states. Classification of state standards was based on information from the National Science Teachers Association (NSTA) specifying which states have current standards that are identical to NGSS, partially NGSS, or non-NGSS. The focus for the analysis was on alignment between the NAEP science framework and the standards of the partial NGSS and non-NGSS states. Information about the results of this study is presented in Section III.

- Input from a group of five or more experts, each of whom would consider the information derived from the two studies mentioned above—the 2016 AIR comparison of NAEP to NGSS (Neidorf et al., 2016) and the more recent HumRRO analysis of state standards relative to NAEP (Dickinson et al., 2021)—as well as other factors given the expert's experience in the field of science education, to present their thoughts on whether the framework needs to be changed and why.
- NAGB recently issued a public call for input on the NAEP science framework regarding its revision. NAGB requested responses from interested parties by October 15, 2021.

NAGB is scheduled at its March 2022 meeting to consider whether to move ahead with a revision of the science framework for application in the design of the 2028 science assessment. The board also will consider the input received from the various sources mentioned above. The timing of these activities should NAGB choose to recommend a science framework revision would easily extend into 2023 if not beyond. Given existing statutes, NAGB will convene two panels based on their policy (NAGB, 2018a, p. 5):

- The **Framework Visioning Panel** shall formulate high-level guidance about the state of the field to inform the process, providing these in the form of guidelines. The major part of the Visioning Panel work will be at the beginning to provide initial guidance for developing a recommended framework. The Visioning Panel shall be composed of the stakeholders referenced in the introduction above. At least 20 percent of this panel shall have classroom teaching experience in the subject areas under consideration. This panel may include up to 30 members with additional members as needed.
- The **Framework Development Panel** shall develop drafts of the three project documents and engage in deliberations about how issues outlined in the Visioning Panel discussion should be reflected in a recommended framework. As a subset of the Visioning Panel, the Development Panel shall have a proportionally higher representation of content experts and educators, whose expertise collectively addresses all grade levels designated for the assessment under development. Educators shall be drawn from schools across the nation, who work with students from high-poverty and low-performing schools, as well as public and private schools. This panel may include up to 15 members, with additional members as needed.

The timeline for initiating and completing the work of the panels remains to be specified, and because the work of the development panel follows from the work of the visioning panel, its work would end sometime in 2023 or later, pending public review of a draft framework and commentary with subsequent revision and then final adoption by NAGB. A revised framework would be used to develop the design and tasks for the 2028 NAEP science assessment.

SECTION II. ANALYSIS OF THE NAEP SCIENCE FRAMEWORK RELATIVE TO OTHER CONTEMPORARY SCIENCE AND TECHNOLOGY FRAMEWORKS

This section examines how the NAEP science framework and assessment and NAEP TEL framework compare with the NRC *Framework for K–12 Science Education* (hereafter, NRC framework) and the derivative *Next Generation Science Standards* (NGSS). It begins with a brief description of key elements of each of the four reference sources and is followed by a summary of results from a detailed study of the correspondences between the two NAEP frameworks and the NGSS. Highlighted in the summary are important areas of similarity and dissimilarity and some of the implications relative to assessment.

Overview of the NAEP Science Framework and Assessment

As noted earlier, the current NAEP science assessment is based on a framework originally developed for the 2009 assessment administration at grades 4, 8, and 12. That framework also was used for the 2011 administration at grade 8 and the 2015 and 2019 administrations at grades 4, 8, and 12. The framework is scheduled to be used once more for the 2024 administration for eighth grade only. The scheduled 2028 operational administration of science for grades 4 and 8 is supposed to be based on an updated science framework.

The current NAEP science framework (NAGB, 2008, 2014) was developed approximately 4 years before the 2009 administration and incorporated ideas from contemporary theory and research on science learning and assessment including synthesis volumes from the NRC: *How People Learn: Brain, Mind, Experience and School* (Bransford et al., 2000); *Knowing What Students Know: The Science and Design of Educational Assessment* (Pellegrino et al., 2001); Systems of State Science Assessment (Wilson & Bertenthal, 2005) and *Taking Science to School* (National Research Council, 2007). The framework included important ideas about the learning and knowing of both science content and science practices with a particular emphasis on their integration as discussed below.

Science Content. The science content for NAEP is defined by a series of statements that describe key facts, concepts, principles, laws, and theories in three broad areas: physical sciences, life sciences, and Earth and space sciences. Table 1 shows the major topics and subtopics within each of the three major science domains. The nature of the specific content knowledge changes in both scope and sophistication across the three grade levels.

Table 1. NAEP science content areas and topics

Physical sciences	Life sciences	Earth and space sciences
Matter • Properties of matter • Changes in matter	Structures and functions of living systems • Organization and development • Matter and energy transformations • Interdependence	Earth in space and time • Objects in the universe • History of Earth
Energy • Forms of energy • Energy transfer and conservation		Earth structures • Properties of Earth materials • Tectonics

Physical sciences	Life sciences	Earth and space sciences
Motion • Motion at the macroscopic level • Forces affecting motion	Changes in living systems • Heredity and reproduction • Evolution and diversity	Earth systems • Energy in Earth systems • Climate and weather • Biogeochemical cycles

SOURCE: National Assessment Governing Board, 2014, Exhibit 4, p. 19. Reprinted with permission.

Science Practices. The second dimension of the framework is defined by four science practices: Identifying Science Principles, Using Science Principles, Using Scientific Inquiry and Using Technological Design. In the NAEP science framework, the first two practices (Identifying Science Principles and Using Science Principles) generally are considered as "knowing science," and the last two practices (Using Scientific Inquiry and Using Technological Design) are considered as the application of that knowledge to "doing science" and "using science to solve real-world problems."

Table 2 provides a high-level description of the nature of each specific practice in terms of the types of cognitive demands placed on students as they engage in a practice as applied to a topic from a specific science content area.

Table 2. NAEP science practices: General labels and specific applications

	Practice Label	Practice Applications			
↓—Communicate accurately and effectively—→	Identifying Science Principles	Describe, measure, or classify observations.	State or recognize correct science principles.	Demonstrate relationships among closely related science principles.	Demonstrate relationships among different representations of principles.
	Using Science Principles	Explain observations of phenomena.	Predict observations of phenomena.	Suggest examples of observations that illustrate a science principle.	Propose, analyze, and/or evaluate alternative explanations or predictions.
	Using Scientific Inquiry	Design or critique aspects of scientific investigations.	Conduct scientific investigations using appropriate tools and techniques.	Identify patterns in data and/or relate patterns in data to theoretical models.	Use empirical evidence to validate or criticize conclusions about explanations and predictions.
	Using Technological Design	Propose or critique solutions to problems given criteria and scientific constraints.	Identify scientific tradeoffs in design decisions and choose among alternative solutions.	Apply science principles or data to anticipate effects of technological design decisions.	

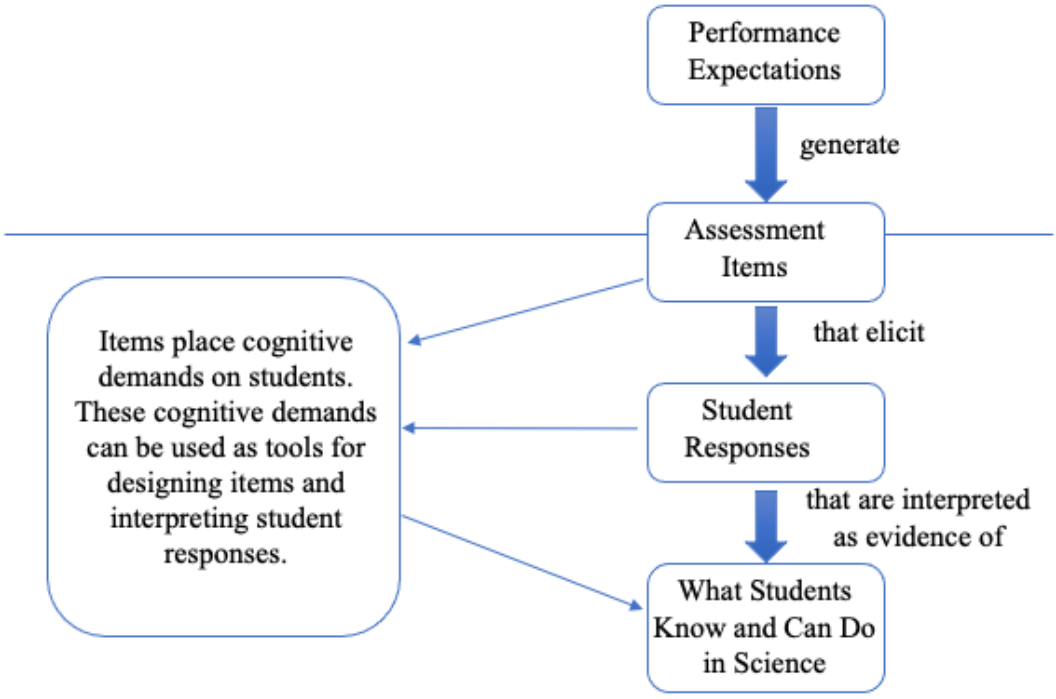
SOURCE: National Assessment Governing Board, 2014, Exhibit 13, p. 76.

Performance Expectations—Combining Content and Practices. The design of the NAEP science assessment is guided by the framework's descriptions of both the science content and science practices to be assessed but with the key assumption that the practices are to be combined with a science content statement to generate specific student performance

expectations that serve as the target for assessment. Assessment items are then developed based on the description of each specific performance expectation.

Using the logic of specific performance expectations as a guide for item development processes, items are then designed to vary the cognitive demands of tasks, a process that then influences the conclusions to be made about student performance. Such a process of item development can be represented schematically as shown in Figure 1.

Figure 1. NAEP assessment item development model



SOURCE: National Assessment Governing Board, 2014, Exhibit 2, p. 12.

In 2009, 2011, and 2015, NAEP science was administered as primarily a paper-and-pencil test. In 2019 a major shift occurred when NAEP science was administered for the first time as an entirely digitally based assessment. The Nation's Report Card (2019) provides a description of the new digital assessment:

> The NAEP digitally based science assessment consisted of standalone, discrete questions, and scenario-based tasks comprising a connected sequence of questions. Scenario-based tasks were designed to engage students in scientific inquiry through hands-on activities and computer simulations set in real-world contexts. The tasks provided students opportunities to demonstrate their knowledge and skills in each of three science content areas and four science practices. The science assessment included two types of scenario-based tasks:

- Interactive computer tasks (ICTs). ICTs use real-world simulations to engage students in scientific investigations that require the use of science inquiry skills and application of scientific knowledge to solve problems.
- Hybrid hands-on tasks (HHOTs). Students perform hands-on scientific investigations using materials in kits provided by NCES. The "hybrid" in HHOTs denotes that these tasks combine hands-on investigations with digital activities. Students use NCES-supplied tablets to view kit instructions, record results and data, and answer assessment questions.

Overview of the NAEP TEL Framework and Assessment

As noted earlier, a TEL framework was developed for the first TEL assessment in 2014 at grade 8 and was used again for the 2018 TEL at grade 8. It is scheduled to be used twice more for the 2024 and 2028 TEL administrations at grade 8.

The development of this framework and assessment was motivated by several factors. In the science education community, a call for preparing students with technology and engineering literacy has been long awaited. The *Science for All Americans* report (American Association for the Advancement of Science, 1990) explicitly suggested that science education should incorporate technology and engineering as a form of scientific inquiry. Bybee (2010) proposed an advance to STEM education by integrating technology and engineering with science and mathematics education. He argued that "there are very few other things that influence our everyday existence more [than technology] and about which citizens know less" (Bybee, 2010, p. 30). Bybee suggested extending traditional information communication technology education by integrating ICTs with other subjects. He further pointed out that involving students in engineering activities could promote their abilities for both problem solving and innovation. He also acknowledged that engineering as typically presented in schools was inconsistent with its careers and contributions to society, and thus authentic scenarios needed to be developed for both learning and assessment (Bybee et al., 2009).

The NRC report, *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*, identified information literacy and ICT literacy as two of the most frequently mentioned critical competencies for students to succeed in the 21st century (Pellegrino & Hilton, 2012). That report discussed various foundations for education, and STEM education in particular, including preparing future entrants to the labor market with the ability to adapt to technological changes in society rather than simply acquiring static bits of knowledge. Similarly, another 2012 NRC report, the *Framework for K–12 Science Education* (NRC, 2012), framed one of the overarching goals of science education as the development of students who "are careful consumers of scientific and technological information related to their everyday lives" (p. 1). The framework explicitly includes "Engineering, Technology, and Applications of Science" as one of four disciplinary core ideas and describes "defining problems, design solutions, and using computational thinking" as critical components of science and engineering practices. Further discussion of the NRC framework follows this section on TEL.

These and other trends related to technology and engineering literacy spurred the development of a TEL framework and inclusion of the TEL assessment as part of the NAEP program. The goal of TEL has been to obtain information about students'

understanding of technology and its effect on our society and environments, as well as students' ability to design solutions to solve real-world problems. The TEL framework describes TEL as the "capability to use, understand, and evaluate technology as well as to understand technological principles and strategies needed to develop solutions and achieve goals" (NAGB, 2013, p. xi). Specifically, the framework identified three interconnected areas to be assessed (NAGB, 2018b, p. xii) as follows:

- *Technology and Society* deals with the effects that technology has on society and the natural world and with the sorts of ethical questions that arise from those effects. Knowledge and capabilities in this area are crucial for understanding the issues surrounding the development and use of various technologies and for participating in decisions regarding their use.
- *Design and Systems* covers the nature of technology, the engineering design process by which technologies are developed, and basic principles of dealing with everyday technologies, including maintenance and troubleshooting. An understanding of the design process is particularly valuable in assessing technologies, and it can also be applied in areas outside technology, since design is a broadly applicable skill.
- *Information and Communication Technology* includes computers and software learning tools, networking systems and protocols, hand-held digital devices, and other technologies for accessing, creating, and communicating information and for facilitating creative expression. Although it is just one among several types of technologies, it has achieved a special prominence in technology and engineering literacy because familiarity and facility with it is essential in virtually every profession in modern society.

Students taking the TEL assessment are expected to succeed in the following three types of thinking and reasoning practices:

- *Understanding technological principles* focuses on students' knowledge and understanding of technology and their capability to think and reason with that knowledge;
- *Developing solutions and achieving goals* refers to students' systematic application of technological knowledge, tools, and skills to address problems and achieve goals presented in societal, design, curriculum, and realistic contexts; and
- *Communicating and collaborating* centers on students' capabilities to use contemporary technologies to communicate for a variety of purposes and in a variety of ways, working individually or in teams. (NAGB, 2018b, pp. 3-2–3-3)

The TEL assessment has developed scenario-based tasks designed to engage students in multimedia environments to gauge students' understanding of technological and engineering principles and their ability to apply such principles to determine design solutions. Most of TEL's assessment tasks are computer simulation problems involving technology and engineering scenarios.

Overview of the NRC Science Education Framework and Next Generation Science Standards

Based on multiple sources of evidence and discussions about the knowing and learning of science, the nature of science education as it had been practiced in the United States, and evidence of relatively poor student achievement in science across K–16+, agreement emerged during the early part of this century about the need for substantial change in science standards, instruction, and assessment, including what we expect students to know and be able to do in science, how science should be taught, and how it should be assessed.

> Recognition of this science education problem can be found in reports spanning elementary, secondary, and postsecondary education (K–16+). These reports present a consistent description of the nature of competence in science and include NRC reports on K–8 science education in formal and informal learning environments (NRC, 2007, 2009); curriculum and assessment frameworks for Advanced Placement (AP) science courses (e.g., College Board, 2011a, 2011b); and even revisions in the nature of the science knowledge required for entry to medical school and assessed on the Medical College Admissions Test (e.g., American Association of Medical Colleges, 2012). (Pellegrino, 2016, p. 5)

Reconceptualization of the nature of science competence emergent from these many and diverse sources was captured to some extent in the College Board's standards for success in high school science (College Board, 2009). Their most complete expression for all K–12 science education was presented in the 2012 NRC report, *A Framework for K–12 Science Education. Practices, Crosscutting Concepts and Core Ideas.* The NRC framework report contains many important key ideas, including articulation of three interconnected dimensions of science competence as denoted in the report's title. The three dimensions are Disciplinary Core Ideas (DCIs), Crosscutting Concepts (CCCs), and Science and Engineering Practices (SEPs). The NRC framework provides detailed descriptions of each dimension, the concepts that each dimension encompasses, and the rationale for their inclusion. Figure 2 provides a list of the dimensions and their associated high-level concepts.

DCIs are the big ideas associated with a discipline, like life science, and which are essential to explaining phenomena. CCCs are ideas like systems thinking that are important across many science disciplines and provide a unique lens to examine phenomena. SEPs are the multiple ways of knowing and doing science and engineering, like developing models and constructing explanations that scientists and engineers use to study the natural and designed world. The framework focuses on the need for the integration of these three dimensions in science and engineering education. The knowledge associated with each of the three dimensions must be integrated in the teaching, learning, and doing of science and engineering, and in assessing what students know and can do. The framework emphasizes research indicating that learning about science and engineering "involves integration of the knowledge of scientific explanations (i.e., content knowledge) and the practices needed to engage in scientific inquiry and engineering design" (NRC, 2012, p. 11). The disciplinary core ideas, crosscutting concepts, and science and engineering practices serve as thinking tools that work together to enable scientists, engineers, and learners to design solutions to problems, reason with evidence, and make sense of phenomena. When learners engage in science and engineering practices integrated with DCIs and CCCs to make sense of

compelling phenomena or design solutions to complex problems, they build new knowledge about all three dimensions and come to understand the nature of how scientific knowledge and engineering solutions develop.

Figure 2. The three dimensions of the NRC framework

Scientific and Engineering Practices
- Asking questions (for science) and defining problems for engineering
- Developing and using models
- Planning and carrying out investigations
- Analyzing and interpretating data
- Using mathematics and computational thinking
- Constructing explanations (for science) and designing solutions (for engineering)
- Engaging in argument from evidence
- Obtaining, evaluating, and communicating information

Crosscutting Concepts
- Patterns
- Cause and effect: Mechanism and explanation
- Scale, proportion, and quantity
- Systems and system models
- Energy and matter: Flows, cycles, and conservation
- Structure and function
- Stability and change

Disciplinary and Core Ideas

Physical Sciences
PS 1: Matter and its interactions
PS 2: Motion and stability: Forces and interactions
PS 3: Energy
PS 4: Waves and their applications in technologies for information transfer

Earth and Space Sciences
ESS 1: Earth's place in the universe
ESS 2: Earth's ecosystems
ESS 3: Earth and human activity

Life Sciences
LS 1: From molecules to organisms: Structures and processes
LS 2: Ecosystems: Interactions, energy, and dynamics
LS 3: Heredity: Inheritance and variation of traits
LS 4: Biological evolution: Unity and diversity

Engineering, Technology, and the Applications of Science
ETS 1: Engineering design
ETS 2: Links among engineering, technology, science, and society

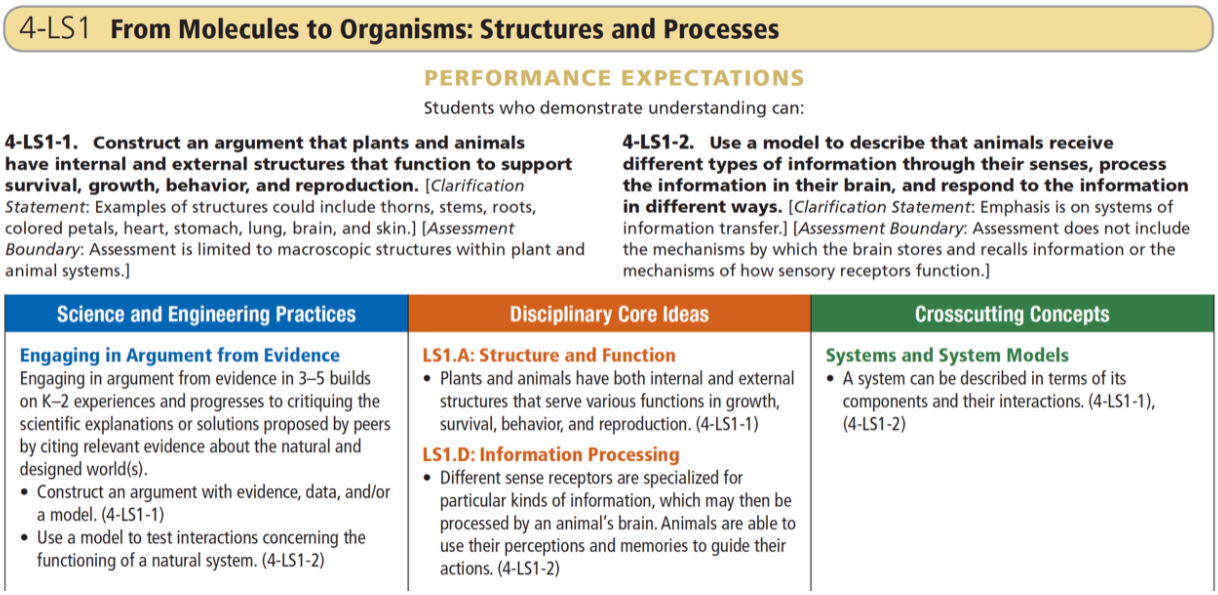
SOURCE: NRC 2012, Box S-1, p. 3.

The rationale for the choice of the specific DCIs is important to note here relative to other previous standards and frameworks. One criticism of U.S. K–12 science curricula relative to those of other countries was that they were "a mile wide and an inch deep" (Schmidt et al., 1997, p. 62). The same concerns about breadth versus depth were made in an NRC Report on advanced study of science in U.S. high schools (NRC, 2002). In reaction, the framework focused on core ideas in each of the four content domains with the directive that students should continue to be exposed to these core ideas with increased levels of complexity and

explanatory power relative to a range of phenomena and problem contexts throughout their schooling.

While each of the three dimensions matters, a central argument of the framework is that proficiency is demonstrated through *performances* that require the integration of all three dimensions. Such demonstrations are labeled *Performance Expectations (PEs)* because they specify what students at various levels of educational experience should know and be able to do. The Next Generation Science Standards (NRC, 2013) are an expression of the integrated knowledge vision contained in the framework, and provide a set of standards expressed as performances expectations for students from Kindergarten to 12th grade. The NGSS appear as clusters of performance expectations related to particular aspects of a core disciplinary idea (see Figure 3 for an example at grade 4). Each performance expectation requires students to draw upon knowledge of a specific practice and a crosscutting concept in the context of specific elements of disciplinary core knowledge. Across the set of performance expectations at a given grade level or grade band, each practice and crosscutting concept appears in multiple standards. A student demonstrates grade-level proficiency by completing performances that demonstrate that they can make use of their knowledge. To truly know and understand science is to be able to use the three dimensions of scientific knowledge together to explain compelling phenomena and/or provide solutions to complex problems.

Figure 3. NGSS Performance Expectations for Grade 4 Life Science 1: From molecules to organisms: Structures and processes



4-LS1 **From Molecules to Organisms: Structures and Processes**

PERFORMANCE EXPECTATIONS

Students who demonstrate understanding can:

4-LS1-1. Construct an argument that plants and animals have internal and external structures that function to support survival, growth, behavior, and reproduction. [*Clarification Statement*: Examples of structures could include thorns, stems, roots, colored petals, heart, stomach, lung, brain, and skin.] [*Assessment Boundary*: Assessment is limited to macroscopic structures within plant and animal systems.]

4-LS1-2. Use a model to describe that animals receive different types of information through their senses, process the information in their brain, and respond to the information in different ways. [*Clarification Statement*: Emphasis is on systems of information transfer.] [*Assessment Boundary*: Assessment does not include the mechanisms by which the brain stores and recalls information or the mechanisms of how sensory receptors function.]

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
Engaging in Argument from Evidence Engaging in argument from evidence in 3–5 builds on K–2 experiences and progresses to critiquing the scientific explanations or solutions proposed by peers by citing relevant evidence about the natural and designed world(s). • Construct an argument with evidence, data, and/or a model. (4-LS1-1) • Use a model to test interactions concerning the functioning of a natural system. (4-LS1-2)	**LS1.A: Structure and Function** • Plants and animals have both internal and external structures that serve various functions in growth, survival, behavior, and reproduction. (4-LS1-1) **LS1.D: Information Processing** • Different sense receptors are specialized for particular kinds of information, which may then be processed by an animal's brain. Animals are able to use their perceptions and memories to guide their actions. (4-LS1-2)	**Systems and System Models** • A system can be described in terms of its components and their interactions. (4-LS1-1), (4-LS1-2)

SOURCE: NRC, 2013, p. 38. Reprinted with permission.

An important issue relative to the present paper's discussion of NAEP Science and NAEP TEL is the NRC framework's emphasis on the connections among science, engineering, and technology. While these connections are somewhat separate across NAEP Science and TEL, key practices and ideas from engineering are included in the NRC framework because of important interconnections between science and engineering and because evidence shows that engaging in engineering design can help leverage student motivation and increase

learning in science. One goal of including ideas related to engineering, technology, and the applications of science in the framework for science education is to help students understand the similarities and differences between science (the natural world) and engineering (the designed world) by making the connections between the two fields explicit and by providing all students with an introduction to engineering.

The NGSS expanded upon the framework's adoption of the logic of learning progressions to describe students' developing proficiency in the three intertwined domains across grades K–12, noting that "If mastery of a core idea in a science discipline is the ultimate educational destination, then well-designed learning progressions provide a map of the routes that can be taken to reach that destination" (NRC, 2012, p. 26). The stress on learning progressions is supported by research on science knowing and learning described in the 2005 NRC report *Systems of State Science Assessment,* the 2007 NRC report *Taking Science to School* and in other documents describing research on the progression of student learning and understanding in science (e.g., Alonzo & Gotwals, 2012; Corcoran et al., 2009). The framework built in the idea of a developmental progression of student understanding across the grades by specifying grade band end point targets at grades 2, 5, 8, and 12 for each component of each disciplinary core idea. For the practices and crosscutting concepts, the framework also provided sketches of possible progressions for learning each practice or concept but did not indicate the expectations at any particular grade level. The NGSS built on these suggestions and developed tables that define what each practice might encompass at each grade level. The NGSS also defined the expected uses of each crosscutting concept for students at each grade level.

The NRC framework and NGSS stand in sharp contrast to prior generations of U.S. science standards (e.g., American Association for the Advancement of Science, 1992; NRC, 1996, 2000) that treated content and inquiry as separate strands of science learning. Unfortunately, both instruction and assessment followed suit. The form the standards took contributed to this separation: Content standards stated what students should know, largely in the form of declarative knowledge, and inquiry standards stated what they should be able to do, largely in the form of procedural knowledge. Consequently, instruction often separated content learning from inquiry and vice versa. Science education often was often criticized as "lots of hands on but not much minds on." In a similar fashion, assessments separately measured content knowledge in the absence of application or inquiry practice components in the absence of content concerns. Thus, the NGSS idea of an integrated, *multidimensional science performance* represents a different way of thinking about science proficiency. Disciplinary core ideas and crosscutting concepts serve as thinking tools that work together with scientific and engineering practices to enable learners to solve problems, reason with evidence, and make sense of phenomena. Such a view of competence signifies that measuring proficiency solely as the acquisition of core content knowledge or as the ability to engage in general inquiry processes is neither appropriate nor sufficient.

In the context of assessment, the importance of this integrated perspective of what it means to know science is that one should be attempting to assess where a student can be placed along a sequence of progressively more "scientific" understandings of a given core idea and successively more sophisticated applications of practices and crosscutting concepts. This idea is relatively unfamiliar in the realm of science assessments, which more often have been viewed as simply measuring whether students know or do not know particular grade-level

content (Pellegrino, 2013). To support an integrated and developmental approach to science learning, the framework explains that assessment tasks "must be designed to gather evidence of students' ability to apply the practices and their understanding of the crosscutting concepts in the contexts of specific applications in multiple disciplinary areas" (NRC, 2012, p. 218). Assessments must strive to be sensitive both to grade-level-appropriate understanding and to those understandings that may be appropriate at somewhat lower or higher grades. This is particularly important for assessment materials and resources to support ongoing classroom instruction. The challenges of designing such multidimensional assessments for classroom and large-scale assessment use are substantial. Potential approaches and solutions were discussed in detail in another NRC report, *Developing Assessments for the Next Generation Science Standards* (Pellegrino et al., 2014).

Comparing the NAEP Science and TEL Frameworks and NGSS

Given the brief descriptions provided above, it should be clear that there are multiple similarities and overlaps as well as differences between the NAEP science framework and the NGSS and between NAEP TEL and NGSS. Even though the NAEP science framework predates the 2012 NRC framework and the derivative 2013 NGSS, overlapping content exists, each has a description of science practices, and both make use of the idea of performance expectations that involve the intersection of content and practice. The NAEP TEL framework was developed about the same time as the NRC framework and overlaps with the latter's highlighting of engineering practices alongside science practices, and its inclusion of Engineering, Technology, and the Application of Science as one of the four disciplinary areas.

Although some of the ideas that are part of the NRC framework and NGSS have found their way over time into the NAEP Science assessment and NAEP TEL assessment, including the design of scenario-based tasks in both NAEP assessments and enacted through technology, neither NAEP framework is reflective of the more dramatic shifts found in the NRC framework and NGSS. NAEP TEL focuses on various aspects of technology and engineering literacy and shares certain things in common with the NRC framework and NGSS. In addition, when it was developed and implemented as a technology-based assessment, TEL included more innovative scenario-based item types than the paper-and-pencil NAEP science assessment. The 2019 digitally based NAEP science assessment has moved in a similar direction. Interestingly, when the NRC framework and NGSS were published, NCES leadership often used TEL items as illustrations of performance tasks in NAEP of the type implied by the NGSS, in part because the paper-and-pencil NAEP science assessment did not include such items at the time.

The most significant difference between NAEP science and NAEP TEL and the NRC framework and NGSS is the singular focus of the latter two on the idea of *knowledge in use*—that competence is demonstrated by being able to use DCI and CCC conceptual knowledge in the context of one or more SEPs to solve problems, explain phenomena, and/or design solutions to challenging problems (Harris et al., 2019). Thus, a major concern regarding the future of the NAEP science and TEL assessments is the nature and degree of the alignment between current NAEP frameworks and the NGSS, especially if most states have adopted NGSS or NRC framework/NGSS alike standards and have implemented state assessments aligned with those standards. A related question is whether states, districts, and schools have accordingly modified curricular choices and instructional practices in ways consistent with

their own standards (NRC framework or NGSS) and assessments. If a serious misalignment between NAEP science and the science and technology instruction and assessment practiced in schools exists, the validity and value of the NAEP science assessment results for the 2024 or 2028 administrations could be seriously questioned.

The remainder of this section includes the results from a detailed examination of the alignment between each of NAEP science and TEL frameworks with NGSS.[1] These data are critical in thinking about whether changes are needed in NAEP to better align with contemporary U.S. frameworks and standards as well as the extent to which a single assessment framework more like the NGSS would suffice to create a NAEP science and technology assessment rather than two NAEP science and technology assessments as is currently the case. Section III examines the situation with respect to (a) state science standards relative to the NGSS, (b) state science assessments relative to their current standards, and (c) implementation of new science standards in terms of curricular choices and instructional practices in the field.

Comparative Study of the NAEP Science and TEL Frameworks and NGSS

The main purpose of *A Comparison Between the Next Generation Science Standards (NGSS) and the National Assessment of Educational Progress (NAEP) Frameworks in Science, Technology and Engineering Literacy, and Mathematics* (Neidorf et al., 2016) was "to determine the extent to which the NGSS performance expectations are aligned with the content objectives and definitions of practices in the NAEP science and TEL frameworks. An additional purpose was to determine the extent to which the NGSS performance expectations involving mathematics-related practices are aligned with the content objectives in the NAEP mathematics framework." (Neidorf et al., 2016, p. 2).[2]

A comparison of the NGSS with the NAEP STEM frameworks can yield multiple important outcomes with potential implications for a revision of NAEP science and a possible merger of NAEP science and TEL. Neidorf et al. (2016) listed the following (p. 2):

- For the science comparisons, similarities suggest areas where NAEP may provide useful science assessment examples and national achievement data on the student understandings in the natural sciences described in the NGSS. Differences suggest areas where NAEP and NGSS-based science assessments may each provide unique contributions.
- The TEL comparisons augment these findings by identifying additional areas of overlap with the engineering and technology content and practices in the NGSS. Together, these comparisons explore how completely the full range of content and practices in the NGSS are covered by the NAEP science and TEL frameworks as well as the unique aspects of each.

[1] The NAEP Science framework and assessment also can be compared to international large-scale science assessment programs in terms of content focus, assessment practices, and future directions. Doing so is beyond the scope of this paper, but for those interested in the PISA and TIMSS science assessment programs, such information is available in a forthcoming chapter on large-scale science assessment (Zhai & Pellegrino, in press).

[2] The Neidorf et al. (2016) study was conducted prior to the adoption of the 2019 math framework for administration in 2026.

- The mathematics comparisons, while more limited, explore the degree of alignment between the mathematics-related performance expectations in the NGSS and the NAEP mathematics framework. The NGSS are not intended to guide mathematics assessments, and the performance expectations in science and engineering do not specify explicit mathematics requirements. However, the mathematics students may need to use in responding to items developed to assess these performance expectations can be inferred and compared to the mathematics included in NAEP across grades. Thus, such comparisons can provide information on how assessments based on the NGSS might compare with NAEP in terms of the level of mathematics and quantitative skills that would be required of students.

Three research questions guided this comparison study (Neidorf et al., 2016, p. 3):

1. *Related to the NAEP science framework*: How similar (or different) are the NGSS performance expectations in physical sciences, life sciences, and Earth and space sciences to the content and practices in the NAEP science framework at the corresponding grade levels?
2. *Related to the NAEP TEL framework*: How similar (or different) are the NGSS performance expectations in engineering, technology, and applications of science to the content and practices in the NAEP technology and engineering literacy framework at the corresponding grade levels?
3. *Related to the NAEP mathematics framework*: To what extent are the mathematics-related NGSS performance expectations and practices aligned with the content and skills specified in the NAEP mathematics framework, and at which grade(s)?

Major Findings

The report discusses multiple ways in which the NAEP science and TEL frameworks and the NGSS were compared and contrasted, including different directions and forms of comparison. A plethora of findings are reported and what follows is excerpted from a summary of the major results of those comparisons. It is taken directly from the AIR report.

> There was a moderate to substantial degree of ***content overlap*** between the NGSS and the NAEP science and TEL frameworks. About half of the NGSS performance expectations in the upper elementary grade band (grades 3–5) covered content that overlaps with NAEP science or TEL at grade 4. In contrast, there was much less content in NAEP science that overlapped with the NGSS at grade 4 (and in TEL that overlapped at any grade).
>
> Ninety percent or more of the NGSS performance expectations at the middle school and high school levels covered content that overlaps with NAEP science or TEL at grades 8 and 12, respectively. A somewhat lower, but still substantial, percentage of content in NAEP science at grades 8, and 12 (from 74 to 88 percent) overlapped with the NGSS.
>
> Because of differences in the depth, breadth, detail, or focus of the overlapping content, *content alignment* was lower than *content overlap* when the NGSS was compared to the NAEP science and TEL frameworks together. Moreover, when relevant performance expectations in the natural sciences (physical sciences, life

sciences, and Earth and space sciences) and in engineering, technology, and applications of science (ETS) were compared to the NAEP science and TEL frameworks individually, content alignment differed by grade and by content domain.

Across frameworks, content alignment of the NGSS with the NAEP science and TEL frameworks was moderate. Roughly half of the NGSS performance expectations aligned to NAEP (science or TEL) at each grade level. At grades 3–5, 38 percent of performance expectations were aligned with the science framework and 13 percent with the TEL framework, with 2 percent in the sciences aligned with both NAEP and TEL. At the middle school level, 44 percent of performance expectations were aligned with the science framework and 13 percent with the TEL framework, with 3 percent in the sciences aligned with both. At the high school level, 44 percent of performance expectations were aligned with the science framework and 13 percent with the TEL framework (with no performance expectations aligned with both).

When looking only at the performance expectations in science, the content alignment of the NGSS with the NAEP science framework was low at grade 4 (36 percent) and moderate at the middle school and high school levels (about 50 percent at each grade level). Comparing NAEP science to the NGSS, alignment at grades 4 and 8 was similarly low (23 percent) and moderate (56 percent), respectively; at grade 12, the alignment of NAEP to the NGSS was substantial (71 percent).

Across grades, the greatest degree of alignment between the NGSS and the NAEP science framework was in life sciences and the lowest was in physical sciences, based on the content similarity ratings at both the objective level and at the content area level as a whole. From 48 to 54 percent of NGSS performance expectations in life sciences were aligned with NAEP objectives compared to from 29 to 42 percent of NGSS performance expectations in physical sciences. Looking at the content areas as a whole, life sciences was the only content area rated as similar at two grades (grades 8 and 12) whereas physical sciences was rated as similar only at grade 12, and Earth and space sciences only at grade 8. None of the content areas as a whole were rated as similar at grade 4.

When looking only at the performance expectations in engineering, technology, and applications of science (ETS), content alignment to the NAEP TEL framework was strong for NGSS performance expectations in engineering design (at least 75 percent at each grade level), but weaker for those in the sciences with connections to ETS, especially at the upper grades (as low as 38 percent). The alignment of NAEP TEL with the NGSS, in contrast, was weak at all grade levels, because many more assessment targets are in NAEP TEL as well as assessment areas or subareas that do not have corresponding disciplinary core or component ideas in the NGSS. In addition to engineering design at all three grade levels, both the NGSS and NAEP TEL include the effects of technology on society and the natural world at the middle and high school levels.

The NGSS and NAEP science framework emphasize some content at different grades. That is, some content that was not similar at the corresponding grade level was aligned at a higher or lower grade level in the other framework. In general, the percentage of objectives aligned at a different grade was low—representing no more than one fifth of the objectives. The one exception was for NAEP science at grade 4, where 59 percent of content statements were aligned at a lower of higher grade in the NGSS. The percentage aligned at a different grade decreased over the grade levels for both the NGSS and the NAEP science framework.

Notably, the NGSS and NAEP objectives at middle school/grade 8 that were aligned to other grades were only aligned at the higher grade level in the other framework (high school/grade 12)—i.e., none of the middle school performance expectations were aligned with NAEP grade 4 content statements in science, and none of the NAEP grade 8 content statements in science were aligned with NGSS performance expectations in grades K–5. In addition, some objectives at high school/grade 12 in both the NGSS and NAEP were aligned at the middle school/grade 8 level in the other framework. Thus, the difference between the NGSS and NAEP science framework at grade 8 was more in terms of what content is emphasized in middle school versus high school.

Both the NGSS and the NAEP science and TEL frameworks include objectives at each grade level that cover *unique content*. This reflects nongrouped objectives covering content that is in one framework but not in its counterpart at any grade. (Examples are given in exhibits 10–12 for science and exhibit 13 for TEL). The unique content, together with content that overlapped but was not aligned at any grade in the counterpart framework, represented between 43 and 48 percent of NGSS performance expectations in science and between 18 and 28 percent of NAEP science content statements. Unique content also represented between 14 and 55 percent of NGSS performance expectations in ETS and between 72 and 87 percent of NAEP TEL assessment targets. Unique content reflects areas where each program can contribute different information about student outcomes.

Practices alignment was uniformly strong, but the emphasis of NGSS performance expectations across the NAEP science and TEL practices differed from the emphases specified in the NAEP frameworks.

Ninety-nine percent of NGSS performance expectations in science were aligned with NAEP science practices and 81 percent of performance expectations in ETS were aligned with NAEP TEL practices.

The NGSS performance expectations in science were more strongly concentrated in the NAEP science practice of *using science principles* (60 percent across grades) than was specified in the NAEP science framework (30 to 40 percent across grades). In contrast, very few of the NGSS performance expectations aligned with *identifying science principles* (4 percent across grades) compared to the 20 to 30 percent specified for NAEP across grades. The emphasis on *using scientific inquiry* (22

> percent) and *using technological design* (13 percent) was more comparable to NAEP science (30 and 10 percent, respectively, across grades).
>
> The NGSS performance expectations in ETS were strongly concentrated in the NAEP TEL practice of *developing solutions and achieving goals* (62 percent across grades), which was greater than what is specified in the NAEP TEL frameworks (40 percent across grades). Only small percentages of NGSS performance expectations aligned with NAEP's *understanding technological principles* (12 percent) and *communicating and collaborating* (7 percent) (compared to 30 percent in each practice across grades in NAEP TEL).
>
> **However, despite some strong indications of alignment between the NGSS and NAEP content and practices dimensions separately, when both content and practices were considered together, the NGSS and NAEP science framework were found to be not aligned at the *overall framework level*. That is, at each grade level, the two frameworks were rated as not similar. This was generally because panelists thought that the individual NGSS performance expectations often went beyond what would be expected based on the descriptions of the practices in the NAEP framework when they are applied to specific content statements, even if the science content covered was similar to that in the NGSS. (Neidorf et al., 2016, pp. 94–97, emphasis added)**

Major Conclusions and Implications

The AIR report (Neidorf et al, 2016) also included a set of major conclusions about the relationships among the NAEP science and TEL frameworks and NGSS based on all the various comparisons executed in the study and the judgments made by experts. It focused on implications regarding possible similarities and differences in the demands of assessments aligned to each of the three reference sources. The following is taken directly from the AIR report.

> Together, the results from the various components of the comparison study suggest that NGSS-based assessments and NAEP science and TEL assessments would be aligned to some degree, but each would also have unique content and different emphases in terms of science and TEL practices. This is because some of the grouped NGSS and NAEP objectives with overlapping content—those that were aligned—would likely lead to similar assessment items, but some were different enough that they would likely lead to assessment items with a different content focus. Additionally, those objectives that were not grouped (and either aligned at a lower or higher grade or not aligned at all) would represent unique content at the given grade.
>
> For example, content alignment of an NGSS-based assessment with the NAEP science assessment would likely be low at grade 4—moderate if the entire upper elementary grade band was considered—and moderate at the middle and high school levels. The lower alignment at grade 4 relates to the greater breadth of content in NAEP (evidenced by the greater number of nongrouped objectives)

and the fact that some of the content in NAEP at grade 4 may be covered at a different grade in the NGSS's upper elementary grade band.

An NGSS-based assessment also would likely have a much greater emphasis—over half the assessment—on *using science principles* and a much lesser emphasis on *identifying science principles* than a NAEP science assessment—only 4 percent. This is not surprising given that NAEP explicitly includes declarative knowledge in this latter practice, where the NGSS emphasize the application of science knowledge.

Another implication looking across the study is that the content and practices embodied in NGSS performance expectations that involve engineering design are not fully covered by either the NAEP science or NAEP TEL framework, despite strong alignment with the engineering design assessment targets in NAEP TEL. This includes both performance expectations in engineering design and those in the sciences that involve design applications. Thus, assessment tasks involving engineering design could look quite different in the two programs despite these areas of overlap.

The NAEP science framework—which specifies the practice of *using technological design* (with which many of the NGSS performance expectations in science that involve design applications aligned)—is restricted to the consideration of scientific criteria, constraints, and trade-offs in making design decisions. This is in contrast to the NGSS (and NAEP TEL), which more fully reflect the engineering design process and include a broader range of considerations such as social and economic factors (excluded in NAEP science). Additionally, the NAEP TEL framework and assessments do not expect prior science content knowledge, in contrast to the NGSS, which require the application of science concepts. NAEP TEL, rather, provides the background on the science concepts needed to be successful on the items and tasks measuring the engineering design process.

A final implication is that the tasks that could be developed to assess the NGSS performance expectations in science and engineering would likely require students to use some mathematics that is beyond the corresponding grade level in the NAEP mathematics framework; in contrast, the NAEP science and TEL assessments require mathematics at or below the corresponding grade. In other words, some of the mathematics that could be required in an NGSS-based assessment would be at a higher level than what is required in NAEP science and TEL assessments. (Neidorf et al., 2016, pp. 98–99)

SECTION III. ANALYSIS OF THE NAEP SCIENCE FRAMEWORK AND ASSESSMENT RELATIVE TO STATE SCIENCE POLICY AND PRACTICES: STANDARDS, ASSESSMENTS, AND CLASSROOM INSTRUCTION

This section examines how the NAEP science framework aligns with science standards and assessments that have been adopted and implemented in the states. Three main questions are of interest: (1) Since publication of the NRC framework and the NGSS, how many states have adopted the NGSS or standards that are similar in nature? (2) How do the standards of those states that have not completely adopted the NGSS align with NAEP? and (3) For those states that have adopted the NGSS or similar standards, what is the status of the design and implementation of their state assessments relative to their standards? The section then seeks to establish what the states are doing in the way of instruction as related to the NRC framework and NGSS. It closes with an examination of trends in NAEP science assessment performance between 2009 and 2019 and what those results might imply about the current state-of-science education. Overall, the information provided in this section has substantial implications for considering where states are likely to be in science instruction and assessment by the time the current NAEP science assessment is administered in grade 8 in 2024 and when the updated science assessment is administered in grades 4 and 8 in 2028.

NAEP, NGSS, and State Science Standards Comparisons

Since the publication of the NRC framework and NGSS states, 21 states have explicitly adopted the NGSS as their state science standards and 24 other states have adopted standards that NSTA has designated as partial NGSS in that they are multidimensional standards like the NGSS. In such cases they have based their standards development on the NRC framework and have typically adhered to the central idea of integrated performance expectations based on two or more dimensions as in the NGSS.

In February 2021, HumRRO published a report for NAGB entitled *Comparative Analysis of the NAEP Science Framework and State Science Standards* (Dickinson et al., 2021).

> The method used to conduct this comparative study relied heavily on obtaining experts' judgments regarding the overlap of subject matter between the NAEP science framework and states' science standards…. The comparative analysis included only the standards from states that did not fully adopt the NGSS (i.e., 6 states) and those that partially adopted the NGSS (i.e., 24 states, including the Department of Defense schools). The science standards from the partial NGSS adopting states, which are based on the NRC framework, were included in the study. However, NGSS performance expectations were excluded from the analysis, given the previous study comparing NAEP and NGSS. (Dickinson et al., 2021, p. 1.)

Table 3 below shows which state's standards were included in the analysis.

To execute this analysis. the HumRRO team started by pulling out all content statements, objectives, and performance expectations outside NGSS. The focus was on the content overlap and not the practice overlap. They did some preliminary distillation by matching state and NAEP content statements to look at state and NAEP content side by side to rate the overlap. Also, they identified content-related practices in state statements. They then

developed a consensus statement to give the overall impression of where states are doing things differently. They tried to include only statements in the science domains and cut out technology and engineering statements if easy to do so. They did not look explicitly at the TEL framework. An important point to note is that in conducting this work, the comparison of NAEP to state standards is based on an aggregation of all the states' standards rather than a state-by-state individual comparison. Thus, the comparison paints a very broad picture of overlap between the NAEP framework and the partial NGSS and non-NGSS states as a whole. Further details about the methodology and specific sets of outcomes can be found in the complete report.

Table 3. Non-NGSS, partial NGSS, and full NGSS adopting states

Non-NGSS Adopting States	Partial NGSS Adopting States	Full NGSS Adopting States
Florida	Alaska	Arkansas
North Carolina	Alabama	California
Ohio	Arizona	Connecticut
Pennsylvania	Colorado	Delaware
Texas	Department of Defense Education	District of Columbia
Virginia	Activity	Hawaii
West Virginia	Georgia	Illinois
	Idaho	Iowa
	Indiana	Kansas
	Louisiana	Kentucky
	Massachusetts	Maine
	Minnesota	Maryland
	Missouri	Michigan
	Mississippi	Nevada
	Montana	New Hampshire
	North Dakota	New Jersey
	Nebraska	New Mexico
	New York	Oregon
	Oklahoma	Rhode Island
	South Carolina	Vermont
	South Dakota	Washington
	Tennessee	
	Utah	
	Wisconsin	
	Wyoming	

SOURCE: Dickinson et al., 2021, p.12.

The following conclusions, based on the analyses completed by both the HumRRO staff and the outside experts, were offered in the report. They are reprinted here verbatim from that document (Dickinson et al., 2021, pp. 6–7).

1. When examining the content covered by the full set of states' science standards (with any NGSS performance expectations removed), there are many state statements that do not overlap in content with any NAEP statement.
   - At grade 4, 31 percent of all state content statements reviewed by HumRRO experts and external science experts were rated as not overlapping a NAEP content statement.

- At grade 8, 32 percent of all state content statements reviewed by HumRRO experts and external science experts were rated as not overlapping a NAEP content statement.
- At grade 12, 55 percent of all state content statements reviewed by HumRRO experts and external science experts were rated as not overlapping a NAEP content statement.

2. Considering only the state content statements that the experts reviewed, all NAEP statements at least partially overlap in content with at least one state statement. In most cases, NAEP statements overlap in content with multiple state statements. Finally, in some cases, NAEP content statements are fully reflected in a combination of multiple state content statements.
   - For each NAEP content statement HumRRO identified multiple state content statements with overlapping content. Review by external experts verified content overlap with at least one of these pairings for each NAEP content statement.
   - Experts noted that there were instances where a combination of state content statements would fully cover the content in a NAEP content statement.
3. Experts rated the least amount of content overlap between NAEP and states' standards at grade 12.
   - Overall, at grade 12, 19 percent of state content statements reviewed by expert panelists were rated as having no content overlap with a NAEP content statement.
4. As with the NAEP-to-NGSS comparison, experts rated the least amount of overlap in content between NAEP and states' standards for the Physical Science domain, especially at grades 8 and 12.
   - At grade 8, 9 percent of state Physical Science content statements reviewed by expert panelists were rated as not overlapping a NAEP content statement.
   - At grade 12, 25 percent of state Physical Science content statements reviewed by expert panelists were rated as not overlapping a NAEP content statement.
5. Science experts identified the grades 4 and 8 state content statements to most frequently reflect NAEP's Identifying Science Practices and the grade 12 state content statements to most frequently reflect NAEP's Using Science Practices. The experts least frequently identified the states' content statements to reflect NAEP's Using Technological Design.
   - At grades 4 and 8, 54 percent of all state content statements reviewed by expert panelists were rated as reflecting NAEP's *Identifying Science Practices*.
   - At grade 12, 51 percent of all state content statements reviewed by expert panelists were rated as reflecting NAEP's *Using Science Practices*.
   - Across the grade levels, between 1 percent and 5 percent of all state content statements reviewed by expert panelists were rated as reflecting NAEP's *Using Technological Design*.
6. Science experts noted that states whose standards are based on the NRC K–12 Science Framework have more in common with NAEP than states whose standards are not based on that framework.
   - Consensus statements developed by both the grade 8 and grade 12 expert panels included assertions that they observed more content overlap between NAEP and the science standards of states who based their standards on the NRC K–12 Science Framework.

State Science Policy and Practices: Standards, Assessments, and Classroom Instruction

Thus far we have established three important findings that bear on a judgment about the validity of results from the NAEP science assessment at the time of its next implementation in 2024 and subsequently in 2028 if substantial revision is not made to both the framework and the derivative assessment before the 2028 administration. First, as described in Section II, major differences exist between the NAEP framework and the NRC *Framework for K–12 Science Education* and the derivative *Next Generation Science Standards* in science content, science and engineering practices, and in their juxtaposition in the form of performance expectations. Second, currently, 45 states (including Department of Defense Education Activity) have either fully adopted the NGSS as their state standards (21) or adopted NGSS-like state science standards (24). Third, when the latter states' standards and those of non-NGSS adopting states (6) are compared with NAEP content, several substantive differences arise. Thus, it seems reasonable to conclude that the current NAEP science framework may be substantially at variance with and lagging a contemporary view of what we want students to know and be able to do in science at grades 4, 8, and 12 and how we would expect them to show proficiency. That view of proficiency has become policy for the preponderance of states and is realized via their state science standards.

How far out of synch the NAEP framework and assessment may be with what instruction and science assessment look like in most states in 2024 and 2028 and with what students know and can do in science depends very much on the following timelines: (a) state adoption of new standards following publication of the NRC framework and NGSS, (b) implementation of new state assessments aligned with those standards, (c) availability of curricular and instructional resources reflecting the new vision of science learning and instruction, and (d) implementation of teacher professional learning programs relative to each of a–c. We provide information relevant to these concerns in the following material.

Time Course for Adoption of New State Standards and Assessments

An article that includes information about adoption of new science standards by Smith (2020) discusses results from the two most recent National Survey of Science and Mathematics Education (NSSME) completed in 2012 and 2018 (see also Banilower et al, 2018). Table 4 shows the pattern of adoption of the NGSS or NGSS-like standards by the states as of 2018. The 16 early adopters did so between 2013 and 2015 while the 24 late adopters did so between 2015 and 2017, and non-adopters had not adopted by spring 2018 when NSSME collected data. Note that there are some differences between Table 4 and the Table 3 shown earlier regarding NGSS adoptions. For example, Florida, North Carolina, Ohio, Pennsylvania, Virginia, and Texas remain nonadopters as of 2021 and they have been joined by West Virginia, which was previously designated as a late adopter. In contrast, Arizona, Alaska, Maine, and Minnesota have moved from the nonadopter group into the late adopter group.

Table 4. Adoption of NGSS or NGSS-like standards – August 2018

Early Adopters	Late Adopters	Non-Adopters
California*	Alabama	Alaska
Delaware*	Arkansas*	Arizona*
District of Columbia	Colorado	Florida
Illinois*	Connecticut	Maine
Kansas*	Georgia*	Minnesota*
Kentucky*	Hawaii	North Carolina*
Maryland*	Idaho	North Dakota
Nevada	Indiana	Ohio*
New Hampshire	Iowa*	Pennsylvania
New Jersey*	Louisiana	Texas
Oklahoma	Massachusetts*	Virginia
Oregon*	Michigan*	
Rhode Island*	Missouri	
South Carolina	Mississippi	
Vermont*	Montana*	
Washington*	Nebraska	
	New Mexico	
	New York*	
	South Dakota*	
	Tennessee*	
	Utah	
	West Virginia*	
	Wisconsin	
	Wyoming	

* Lead state

SOURCE: Data are from Smith, 2020.

One of the many factors driving instructional practice relative to the vision of science teaching, learning, and assessment contained in the NRC framework and state science standards aligned with that vision is the status of each state's large-scale science assessment relative to its adopted standards. Consistent with federal requirements, states that have adopted new science standards are obligated to implement new assessments aligned with those standards having the minimum requirement for at least one assessment in each of the elementary school grade bands (grades 3–5), the middle school grade band (grades 6–8), and the high school grade band (grades 9–12). An analysis for this paper by AIR staff of the 21 states that have fully adopted the NGSS (14 of which are shown as lead adopters in the table above) reveals that all but one of those 21 states, Arkansas, has already developed and in most cases implemented a large-scale science assessment that they claim is aligned with the NGSS. The timeline of assessment implementation varies from 2014 to 2019, with some implementations planned for 2020 but delayed until 2021, given suspension of all large-scale assessments in spring 2020 due to the COVID-19 pandemic. The timelines for implementation of new science assessments for the states classified as partial NGSS are less clear although for the majority of those states their websites indicate that their standards and assessments require integration of the disciplinary core content and practices described in the NRC Framework and many include mention of the third dimension of crosscutting concepts. Some have adopted many if not all the performance expectations from the NGSS. For some states, the timeline for full implementation of new assessments extends to 2025.

Survey Information on Science Instructional Practices: 2018 vs. 2012

NSSME has provided periodic snapshots of K–12 science instruction in the United States for more than 40 years. Study topics include teacher backgrounds and beliefs, professional learning opportunities, course offerings, instructional objectives and activities, resources for instruction, and policies affecting instruction. The two most recent studies were conducted in 2012 and 2018. The 2012 study provides baseline data on multiple indicators prior to publication of the NGSS. From 2013 to 2018, 39 states and the District of Columbia adopted the NGSS or NGSS-like standards. By the time the 2018 survey was conducted, NGSS states accounted for more than two thirds of the nation's K–12 students. The 2018 study provides a snapshot of the state-of-science instruction in 2018 relative to the vision of the NRC framework and the NGSS, including the opportunity to observe any impact on instructional beliefs and practices relative to 2012 in light of the publication of the NRC framework in 2012 and the NGSS in 2013.

Smith's 2020 analysis and discussion of results from the 2018 NSSME (Banilower et al., 2018) shows that states have been slow in the full implementation of their new science standards in terms of making a difference in instructional practice. As discussed by Smith, one reason for the slowness is the lack of good curriculum materials aligned with the new standards. Another reason for the slowness is the need for substantial teacher professional development related to understanding the science and engineering practices as well as the meaning and manifestation of integration of the multiple dimensions expressed by the performance expectations. Related to the latter, valid, high-quality assessments reflecting the kinds of performances expected from students also have been lacking. In general, during the period in question there was a paucity of such examples for classroom use as well as at the large-scale state assessment level given the timeline for implementation of new NGSS-aligned assessments as described above from the analysis of state websites by AIR staff.

Regarding professional development, Smith (2020) reports that roughly four of five secondary science teachers (i.e., middle school and high school) participated in science-focused professional development in the preceding 3 years, in contrast to three of five elementary science teachers. Only about half of schools or districts offered any science-focused professional development in the preceding 3 years, and participation data were largely unchanged since 2012. About a third of secondary teachers participated in more than 35 hours of professional development in the 3 years preceding 2018, and more than 4 in 10 elementary teachers had none. As Smith notes, even 35 hours, spread over 3 years, is not much considering prominent instructional practices and the shifts that the framework and NGSS entail.

Among the other results summarized by Smith were results regarding data on instructional practices and emphases in elementary, middle school, and high school classrooms (see Smith 2020, Table 1). Most importantly, in 2018 the most frequent "heavy emphasis" instructional objective reported by Science teachers was "understanding science concepts," particularly in middle and high schools (47 percent of Science teachers in elementary schools, 77 percent in middle schools, and 76 percent in High schools). In contrast, the second most frequent objective with a heavy emphasis reported by teachers was "learning how to do science" but only in 26 percent of Science classes in elementary schools, 46 percent in middle schools, and 41 percent in High schools. Smith concluded that:

> Despite widespread adoption of the NGSS and NGSS-like standards, data from the NSSME+ point to few differences in science instruction compared to 2012. Further, the data from teachers in adopting states vary little from those in non-adopting states. Among the few differences, we do see encouraging signs. Among them, classes in adopting states were more likely to emphasize learning how to do engineering, and they were less likely to emphasize learning vocabulary and facts. In terms of instructional activities, classes in early-adopting states were less likely to rely on lecture and more likely to have students do hands-on activities. However, the data overall suggest that much work lies ahead to achieve the vision laid out in the framework and the standards themselves (Smith, 2020 p. 608).

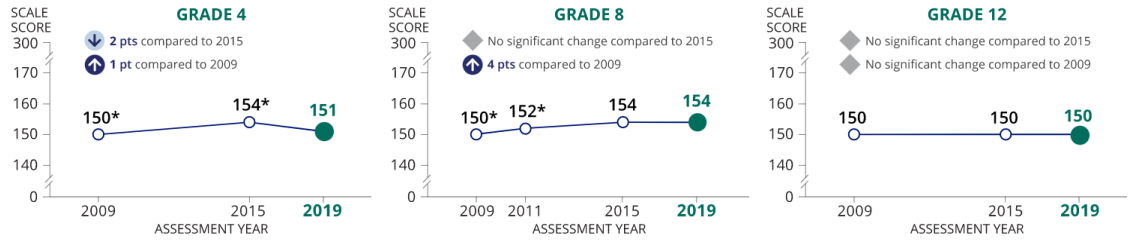
Perhaps not surprising is that substantial changes in science instructional practices were not observed in the 2018 NSSME survey relative to 2012 and that aspects of the vision for science teaching and learning embodied in the NRC framework and NGSS were less well represented in teacher beliefs and instructional practices. As noted by Smith (2020), 5 years may not be enough time. Many of the critical factors needed to spur change are only now becoming more prominent with further changes on the horizon during the next 2 years when NAEP science is set to be administered again for grade 8 only. Among the drivers of change are new state science assessments reflecting the NGSS or similar science standards. In addition, growth in both commercially available and open education resources (OER) aligned with the NGSS has been significant. One of the largest of the OER curricular initiatives is the foundation-funded OpenSciEd project (https://www.openscied.org/about/), which has generated instructional units covering all the middle school NGSS performance expectations and is working on similar materials for other grade levels. At the classroom level, assessment resources have been developed to support formative and summative assessment practices in ways aligned with the multidimensional assessment vision described in the 2014 NRC report, *Developing Assessments for the Next Generation Science Standards* (Pellegrino et al., 2014). See for example the materials available from the *Next Generation Science Assessment Project* (http://nextgenscienceassessment.org) and from the Stanford *NGSS Assessment Project* (https://scienceeducation.stanford.edu/assessment).

NAEP Science Performance Changes Over Time

One final source of information about possible changes in science education in the United States over time might be gleaned from an examination of performance on the NAEP science assessment for the period from 2009 when the new science framework and assessment were first implemented to 2019 when NAEP science was delivered as a digitally based assessment, in contrast to prior years. These data track student performance both before and after the NRC framework and NGSS.

The 2019 NAEP science scale score results are shown in Figure 4 for each of the grade levels in comparison to prior administrations back to 2009. As can be seen in Figure 4, the average science score for the nation at grade 4 was lower by 2 points compared to 2015, whereas average scale scores at grades 8 and 12 did not significantly differ from 2015. At grades 4 and 8, average scale scores were higher when compared to 2009, while the average scale score at grade 12 was not significantly different across years.

Figure 4. Average scores in NAEP science, by grade: 2009–2019

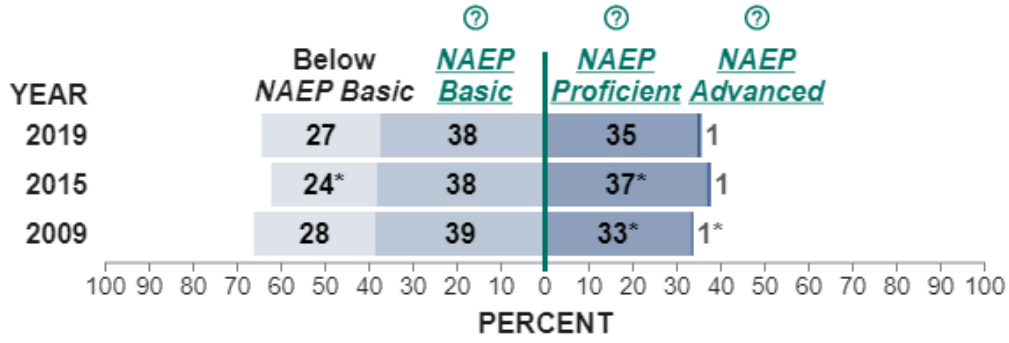


*Significantly different ($p < .05$) from 2019.

SOURCE: The Nation's Report Card, 2019. Reprinted with permission.

Although the absolute levels of the scale scores and the trends in those scores are important indicators of student performance, of particular significance is the reporting of results in terms of achievement levels. As shown below in Figures 5, 6, and 7, the rates by which students were classified into the achievement levels varied across the grades with the highest rate of *Proficient* classifications occurring in grade 4, slightly lower levels of proficiency at grade 8 and substantially lower student proficiency classifications at grade 12. Note that at all three grade levels, there is a very low level of classification of student performance at the *Advanced* level. This finding holds across years.

Figure 5. NAEP achievement-level results in NAEP science for fourth-grade students: 2009, 2015, and 2019

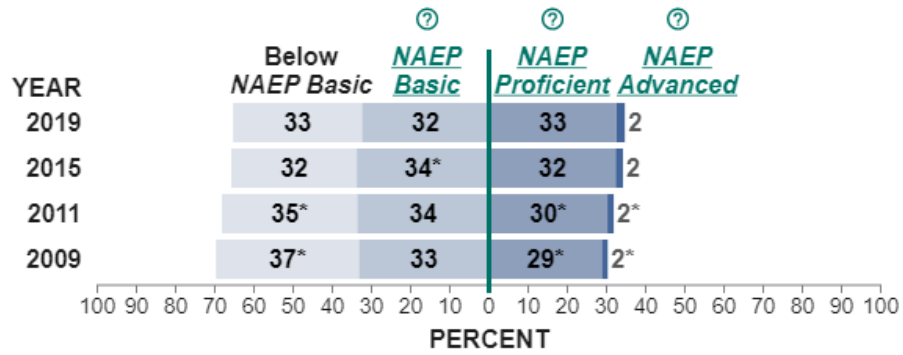


* Significantly different ($p < .05$) from 2019

Note: NAEP achievement levels are to be used on a trial basis and should be interpreted and used with caution.

SOURCE: The Nation's Report Card, 2019. Reprinted with permission.

Figure 6. NAEP achievement-level results in NAEP science for eighth-grade students: Various years, 2009–2019

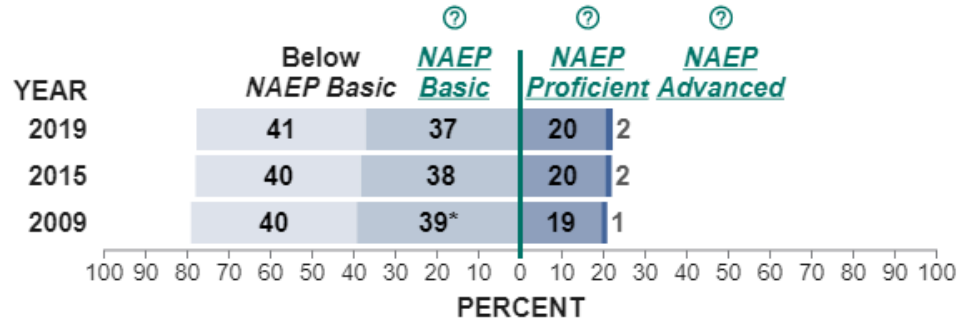


* Significantly different (p < .05) from 2019.

Note: NAEP achievement levels are to be used on a trial basis and should be interpreted and used with caution.

SOURCE: The Nation's Report Card, 2019. Reprinted with permission.

Figure 7. NAEP achievement-level results in NAEP science for twelfth-grade students: 2009, 2015, and 2019



* Significantly different (p < .05) from 2019.

Note: NAEP achievement levels are to be used on a trial basis and should be interpreted and used with caution.

SOURCE: The Nation's Report Card, 2019. Reprinted with permission.

Perhaps there are two major takeaways from this examination of the NAEP science assessment results. First, not much has changed over time implying that science instruction also has not changed substantially despite the existence and adoption of new standards with higher expectations about what students are supposed to know and be able to do. Despite their differences in content and format of science assessment, the most recent trend results from the PISA science assessment and the TIMSS science assessment largely corroborate the lack of change in U.S. science performance during the last decade. Second, those new standards are much needed because science performance across the grade bands is relatively poor and only declines across grades. The vast majority of students are below *Proficient* as defined by the NAEP achievement levels.

The real concerns then are threefold: (1) whether instruction aligned with the new standards will take hold in ways envisioned by the NRC framework and NGSS and change

performance, (2) whether the NAEP science assessment can track the impact of those changes given the differences between the NAEP framework, the NGSS and the majority of state science standards, and (3) whether NAEP science and/or TEL will have sufficient instructional sensitivity to reveal what has and has not happened over time when next administered in 2024 or 2028.

SECTION IV. TECHNOLOGY IMPLICATIONS FOR NAEP SCIENCE

This section briefly considers how various developments in digital technologies need to be considered in reviewing the existing NAEP science framework and assessment and envisioning possibilities for their updating. The discussion that follows focuses on the affordances of technology regarding the constructs that could be included in a revised framework and the associated task design, data capture, and data analytic issues involved in an assessment aligned to an updated framework. The section concludes with a brief discussion of practical and equity concerns related to digitally based assessment of science and technology proficiency.

Technology and NAEP Assessment

During the last two decades, much has been written and speculation made about the power of technology to both improve and transform assessment across a range of assessment contexts and purposes (e.g., Behrens et al., 2019; Bennett, 2008; Drasgow, 2016; Gane et al., 2018, Pellegrino & Quellmalz, 2010; Pellegrino et al., 2001). Although technology's potential for improving and transforming assessment has yet to be fully realized, the vast majority of national-, international-, and state-level assessments of science and technology have moved almost entirely to digital presentations of materials accompanied by technology-based data capture for purposes of scoring, analysis, and reporting. Within the past decade, PISA (2015, 2018), eTIMSS (2019), NAEP Science (2019), and NAEP TEL (2014, 2018) have been delivered via technology using various types of devices including laptops, tablets, and desktops.

Not only has technology changed assessment delivery, response capture, and scoring, it also has had a significant effect on assessment design. This includes the types of tasks and situations that can be presented to students with the goal of tapping into various forms of scientific thinking and reasoning aligned with the practices of science and engineering as found in the NAEP science and TEL frameworks and NGSS. For the NAEP program, some of the newer task types that take advantage of some of technology's affordances were briefly described in Section II, including the scenario-based tasks added to the NAEP science assessment in 2019. The latter were modeled to a great extent after the digitally based tasks were first introduced in NAEP TEL in 2014. The literature on NAEP has considered a number of the affordances of technology for the assessment program, including implementation and analysis of the types of scenario-based tasks in science piloted by NAEP in 2015 and included as part of NAEP 2019, including analyses of student response data (e.g., Bennett, 2008; Bergner & von Davier, 2019; Duran et al., 2020; Lee at al., 2019; Mullis, 2019). The purpose of the discussion that follows is to briefly highlight some of the possibilities for the future of NAEP science as related to both the framework and the assessment.

Opportunities and Possibilities for NAEP Science

As discussed in prior sections of this paper, conceptions of scientific and technological competence have evolved during the last 10–15 years, some of which align with the current NAEP framework and assessment while others go beyond both. Thus, in considering possible changes for the design of the 2028 administration of the science assessment, it will be important to consider how some of the affordances of technology discussed below may

influence the nature of the competencies included in the framework, the design of the assessment tasks needed to provide evidence of those competencies, and the associated measurement and interpretive challenges, especially in light of goals for reporting the results. The assessment as evidentiary reasoning argument presented in the NRC report *Knowing What Students Know: The Science and Design of Educational Assessment* (Pellegrino et al., 2001) frames the discussion. In Chapter 7 of that report many of the affordances of technology for advancement of assessment design and practice are discussed in terms of the three interconnected components of the assessment triangle: *Cognition*, *Observation,* and *Interpretation*. As argued in that report:

> The role of any given technology advance or tool can often be differentiated by its primary locus of effect within the assessment triangle. For linking *cognition* and *observation*, technology makes it possible to design tasks with more principled connections to cognitive theories of task demands and solution processes. Technology also makes it possible to design and present tasks that tap complex forms of knowledge and reasoning. These aspects of cognition would be difficult if not impossible to engage and assess through traditional methods. Related to the link between *observation* and *interpretation*, technology makes it possible to score and interpret multiple aspects of student performance on a wide range of tasks carefully chosen for their cognitive features, and to compare the resulting performance data against profiles that have interpretive value. (Pellegrino et al., 2001, p. 252)

The discussion that follows elaborates on these general ideas regarding NAEP science. It focuses is on the constructs that could be represented in an updated framework, the ways in which those constructs could be realized in the assessment environment, and some of the interpretive challenges and solutions associated with doing so for purposes of measurement and reporting.

The *Cognition* vertex of the assessment triangle. What matters in assessment is what we are trying to reason about – the contemporary conception of student *Cognition* in a domain like science that matters to scientists, educators, and society. A contemporary view of multidimensional proficiency in science includes the expectation that learners should be able to use their disciplinary core knowledge to engage in a variety of science practices in the service of explaining phenomena and designing solutions while answering challenging questions (NRC, 2012). As the conception of student cognition changes and expands in terms of what students are supposed to know and be able to do, as has been the case for science, technology affords opportunities for substantially changing and extending the *Observation* and *Interpretation* components of the assessment triangle in order to more adequately represent and provide evidence about the constructs of interest. Doing so enhances the entire evidentiary reasoning process and the validity of the NAEP science assessment given its intended interpretive use as an index of trends in U.S. science achievement.

The *Observation* vertex of the assessment triangle. Technology provides opportunities for presentation of dynamic stimuli (e.g., videos, graphics, 2- and 3-D simulations) that can be interacted with in the service of eliciting relevant sets of responses from students. Simultaneously, technology enables the generation and capture of a variety of response products, including situations in which students generate responses using multiple modalities

(e.g., drawing and writing). In general, *technology-enhanced assessments* are defined by their capacity to provide novel stimuli and/or responses that would not be possible with traditional, paper-and-pencil assessment formats. Technology-enhanced assessments such as those included in NAEP science 2019 and NAEP TEL enable engagement with a variety of science and engineering practices (e.g., generating models, planning and carrying out investigations, engaging in computational thinking) by opening the door to interactive stimulus environments and response formats that better match the intended reasoning and response processes that form the basis for desired claims about student proficiency (Gorin & Mislevy, 2013).

Students' interactions with these technology-enhanced assessments can be logged to provide data on how they engage in particular processes. In certain applications such as engineering or experimental design, the process by which one completes the activity can be as important a piece of information about knowledge and skill as the final product. In these cases, understanding the operations that students performed in the process of creating the final product may be critical to evaluating students' proficiency. Log data offer the opportunity to reveal these actions, including where and how students spend their time, and what choices they make in situations like using a simulation. Such applications offer the potential to provide large volumes of "click-stream" and other forms of response process data that might be useful for inferences about student thinking as discussed by Ercikan and Pellegrino (2017). Such data can be complex, however, and must be segmented and analyzed in construct-relevant ways if they are to be reliable and valid for a given interpretive use. An ongoing challenge is identifying how to take massive volumes of log data and distill it into actionable information to make judgements about students' knowledge, skills, and abilities (e.g., Bergner & von Davier, 2019).

The *Interpretation* vertex of the assessment triangle. Technology offers significant opportunities for enhancement of the reasoning-from-evidence process given the types of observations described above. Collecting the types of data just mentioned in the discussion of observations makes little sense unless there are ways to reliably and meaningfully interpret them. This can evolve through mechanisms such as automated scoring of responses and application of complex parsing, statistical and inferential models for response process data. Much has been written recently about the opportunities of student-response-process data for capturing what students are doing when they solve problems and answer questions related to science and technology (see Ercikan & Pellegrino, 2017). Such data include the time taken to perform various actions, the actual activities chosen, and their sequence and organization. The potential exists for examining the global and local strategies students use while solving assessment problems and the implications, including how such strategies relate to the accuracy or appropriateness of final responses. Although capturing such data in a digital environment is "easy," making sense of the data is far more complicated. The same can be said for capturing data to constructed response questions where students may be expressing in written and/or graphical form an argument or explanation about some scientific problem or phenomenon, describing the design of a scientific investigation, or representing a model of some structure or process.

The data capture contexts described above are challenging regarding scoring and interpretation. It is here that AI and machine learning may play a significant role in future science assessments. Machine learning mimics human scoring processes by first "learning"

from scoring by human experts to develop algorithmic models and then applying those models to automated scoring of new student responses (Zhai, Yin, et al., 2020). Advances have been made in the automated scoring of short, written, constructed responses for various topics and content in science and other subjects (see Beggrow et al., 2014; Nehm et al., 2012; Williamson et al., 2012). However, automated scoring of other types of constructed response products, such as the features that might be included in drawings and other forms of graphical representation associated with a practice like modeling, has not yet been explored in-depth (see Gerard et al. [2016] for one promising attempt). For both written and graphical responses, well-designed task models that define the features of responses that matter for scoring are needed. This likely will have a considerable impact on the development of automated scoring systems that are both reliable and practical for implementation across a variety of assessment contexts.

Developments in machine learning also may allow researchers to analyze complex response process data of the type described above (Zhai, 2021). Traditional statistical methods are often difficult or inappropriate to apply to such data. Machine learning, however, might assist in analyzing these types of data to reveal patterns that provide important insights into students' cognitive processes in problem solving (Zhai, Haudek, et al., 2020; Zhai, Yin, et al., 2020). Such data may prove to be especially informative about student thinking and reasoning and thus add greatly to the knowledge gained about student competence from large-scale assessments like NAEP that go beyond the performance accuracy data they now provide. An interesting example was provided in a recent study by Pohl et al. (2021). The authors showed that differences in student response processes, of the type described above, when combined with scoring methods, can significantly change the interpretation of a country's performance on a large-scale assessment such as PISA. Their study findings showed that current reporting practices in PISA confound differences in test-taking behavior with differences in competencies and can do so in a different way for different examinees, threatening the validity and fairness of comparisons. Thus, their argument is that test-taking behavior is not a confounding factor introducing construct-irrelevant variance, but that it is something that provides important information on how examinees approach tasks, which can be meaningful outside the testing situation. Disentangling and reporting all these factors as part of a performance portfolio could result in fairer comparisons across groups and enables a better understanding of student competencies and important possible causes of variations in performance. Explorations of the analysis and interpretation of response process data have been initiated for some of the NAEP science tasks (Bergner & von Davier, 2019; Lee at al., 2019) and the results suggest that this is a fertile area for future exploration, albeit taking into consideration some of the cautions mentioned below.

Areas of Concern for NAEP Science

Assessments that can tap into and measure multidimensional knowledge take the form of *knowledge-in-use* tasks (Harris et al., 2019). Technology can make practical the design, administration, and scoring of such tasks. An area of concern is that technology by itself is not enough: Technology cannot fix assessments that are poorly designed or misaligned with the desired learning targets. Instead, technology considerations need to be integrated with assessments through a transparent and principled design process. As the targets of assessment become more conceptually complicated, with demands such as jointly measuring science practices and conceptual knowledge, a principled design process is essential for

developing relevant and valid assessment tasks (Gorin & Mislevy, 2013; Pellegrino et al., 2014). A principled design process like *Evidence Centered Design* (Mislevy, 2018; Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006) that identifies task and response features that matter can also move the scoring process from a black box statistical approach to one that is more transparent and defensible. Explicit task and response models with defined response features can lead to improved human scoring as well. A caveat, in a general sense, for NAEP science is that if NAEP wants to capture more complex forms of scientific thinking and reasoning using digital environments, this cannot be done by simply applying technology to the sense-making process "after the fact," which seldom is well done or efficient. Thus, a very deliberate design process needs to be used for task design and data capture that takes into consideration the relevant forms of evidence and the means for interpretation of that evidence throughout the task design, task refinement, and task validation processes.

Although technology can enhance many aspects of large-scale assessment, concerns have arisen about the equity and fairness of digitally based assessment. An area of concern is comparability of results and validity of inferences derived from performance obtained across different modes of assessment, especially for varying groups of students (see Berman et al., 2020). As NAEP science has moved from paper-and-pencil assessment to digitally based assessment, the general focus has been on mode comparability and concerns about student familiarity and differential access to the hardware and software used (see Way & Strain-Seymour, 2021). As the digital assessment world advances, a significant issue for future large-scale science and technology assessments is determining how student background characteristics including language, culture, and educational experience influence performance on different types of tasks and innovative assessment designs that leverage the power of technology. As the tasks become more innovative, equity and fairness concerns may become even more important than general mode comparability effects.

Another area of concern relates to cost, efficiency, and feasibility. Complex, scenario-based tasks such as those found in NAEP science and TEL are challenging to design well and costly to create relative to more conventional tasks. They typically also take significant amounts of time for students to complete. Given the nature of the scenarios, they also tend to be memorable because they depict interesting, engaging, and often realistic problem-solving situations. They exemplify and perhaps magnify many of the challenges that have long been noted about the inclusion of performance tasks in large-scale testing programs such as NAEP. Davey et al. (2015) provided an excellent discussion of the many challenges associated with development and deployment of performance assessments for constructs represented in science standards such as the NGSS. Their report included a discussion of many of the measurement and statistical challenges associated with the interpretation and reporting of performance data. Thus, NAEP science will have to consider tradeoffs associated with inclusion of technology-based assessment tasks relative to adequate representation and sampling of the constructs of interest. The fact that NAEP science uses a matrix-sampled block design for selection and administration of tasks may mitigate some of the many concerns noted by Davey et al. (2015). NAEP can offer leadership to the large-scale science assessment field in providing a vision and examples of how science and technology competence can and should be assessed and reported.

SECTION V. CONCLUSIONS AND RECOMMENDATIONS

The purpose of this white paper is to consider the need for a revised NAEP science framework and its possible scope and focus including expansion to aspects of what is represented in NAEP TEL. The goal is to provide the NAEP NVS Panel and the NAEP program input about possible futures for NAEP science. As such, the paper can also serve as input to NAGB's deliberations in 2022 about the need and possible directions for a revision of the science framework that would in turn serve as the basis for development of the NAEP science assessment scheduled for 2028.

Topics Covered Across Sections I–IV

- A brief history of the current NAEP science and TEL frameworks and assessments and their projected use over the next seven years through 2028
- Brief descriptions of the content and focus of the NAEP science and NAEP TEL frameworks and assessments as well as the National Research Council's *Framework for K–12 Science Education* (NRC, 2012) and the derivative *Next Generation Science Standards* (NRC, 2013)
- Results from an extensive comparison of the content and focus of both NAEP frameworks with the NGSS
- Information on the timeline and status of state adoptions of the NGSS or similar science standards derived from the NRC framework
- Results from a study comparing the content of state science standards with the NAEP science framework for states with science standards similar but not identical to the NGSS together with states with standards unrelated to the NGSS or NRC framework
- Information about the status of development and implementation of standards-aligned, large-scale state science assessments for those states that have either adopted the NGSS or similar standards
- Information about the conditions of science instruction based on the 2012 and 2018 National Survey of Science and Mathematics Education
- Trends in NAEP science assessment performance for 2009–2019 for students at grades 4, 8, and 12
- A discussion of the affordances of technology for consideration in refinements and revisions to the NAEP science framework and assessment

Conclusions and Implications

Alignment of NAEP Science and NAEP TEL With Other Frameworks and Standards

The frameworks for NAEP science and NAEP TEL were developed before the NRC framework and NGSS and all within a window of approximately 6–7 years. All four drew upon bodies of theory, research, and practice regarding the knowing, learning, and teaching

of science and technology available at the time of their development. Given time lags among them, it should come as no surprise that there are both significant similarities between the two NAEP frameworks and the NGSS and substantial differences as determined by a 2016 AIR comparison study (Neidorf et al., 2016).

Conclusion 1. Overlap exists between NAEP science and NGSS in terms of the focal science content areas—physical science, life science, and Earth and space science—and subtopic areas within each domain, but substantial differences exist in specific content. The differences are magnified in the movement from grade 4 to grade 8 to grade 12. One reason for the pattern of differences across grade levels is that the NGSS is based on a set of four disciplinary core ideas (DCIs) in each domain of science, and each DCI is elaborated across grades in terms of knowledge expectations. This was a deliberate design decision in the NRC framework that is replicated in the NGSS. In contrast, the NAEP framework changes content emphasis and focus across grades 4, 8, and 12 with an increasing emphasis on physical science content at grades 8 and 12, especially at grade 12.

Conclusion 2. Overlap exists between the NAEP framework and NGSS regarding the concept of science practices that describe ways of thinking about and reasoning with science content. The NAEP science practices and the NGSS science practices are different in at least two ways, however. Two of the four NAEP practices are considered to be more focused on "knowing science" in contrast to the other two that are more focused on "doing science." In contrast, the NGSS includes eight specific science and engineering practices, each of which fall under the category of science inquiry ("doing science") and/or engineering design. In general, the NGSS science and engineering practices are more demanding than at least two of the NAEP practices, and this is especially apparent when the practices are combined with content to form performance expectations as noted below.

Conclusion 3. Although both NAEP and NGSS express the targeted knowledge and skills for students in the form of performance expectations, the NGSS performance expectations are considered to demand much more in the way of application of disciplinary content knowledge to answer a question involving a science practice to demonstrate proficiency. Regarding the latter point, the 2016 AIR comparison study concluded: "… despite some strong indications of alignment between the NGSS and NAEP content and practice dimensions separately, when both content and practices were considered together, the NGSS and NAEP science framework were found to be not aligned at the *overall framework level*. That is, at each grade level, the two frameworks were rated as not similar. This was generally because panelists thought that the individual NGSS performance expectations often went beyond what would be expected based on the descriptions of the practices in the NAEP framework when they are applied to specific content statements, even if the science content covered was similar to that in the NGSS" (Neidorf et al., 2016, p. 97).

Conclusion 4. The NGSS includes a fourth dimension in its content framework—engineering, technology, and the applications of science as well as two engineering practices—defining problems and designing solutions. The AIR comparison study (Neidorf et al., 2016) showed that the NGSS has overlap with both NAEP science and NAEP TEL with respect to certain aspects of engineering, technology, and design. The overlap is highly variable, however, depending on grade level and direction of comparison. A significant difference between NGSS and TEL is that NGSS performance expectations related to

technology and design require science content knowledge, which is not true of the TEL assessment that provides relevant science content in the task situation.

Conclusion 5. Given differences between NAEP science, NAEP TEL, and the NGSS in terms of content, practices, and performance expectations, the AIR study (Neidorf et al., 2016) concluded that an assessment aligned to the NGSS could look substantially different from assessments aligned with either NAEP science or NAEP TEL. Much of this difference is associated with the demands of the NGSS performance expectations for science DCIs, as noted above. The same concern applies to performance expectations for the DCI designated as engineering, technology, and applications of science as well as performance expectations involving the engineering practices when combined with science disciplinary content. For the most part, the NGSS performance expectations likely would lead to more challenging assessment tasks than those found in either NAEP science or NAEP TEL.

Status of State Science Standards, Assessments, and Instruction

Given substantial differences between the NAEP science and NAEP TEL frameworks and the NGSS, an obvious question is the degree to which states have adopted the NGSS or similar standards and the status of implementation of policies and practices associated with those standards. Included among the latter is implementation of state large-scale assessments aligned to their current standards. A related concern is penetration of the NRC framework's vision for science learning, teaching, and assessment at the level of classroom practice. Such information has implications for the validity of results from the NAEP science assessment when it is re-administered in grade 8 in 2024 and when an updated science assessment is administered in grades 4 and 8 in 2028.

Conclusion 6. Currently, 45 states (including the Department of Defense Education Activity) have either fully adopted the NGSS as their state standards (21) or adopted NGSS-like state science standards (24; Dickinson et al., 2021). These states represent a substantial proportion of the total U.S. student population across grades K–12. When the standards of states that have adopted NGSS-like standards (24) and those of non-NGSS-adopting states (6) are compared to the NAEP framework based solely on content, several differences arise. Such differences are not surprising given that standards based on the NRC framework are likely to show results that are highly similar to those obtained directly from comparison of content from the NAEP science framework with the NGSS. As mentioned above, the NRC framework and NGSS include a specific set of disciplinary core ideas that remain constant across grade levels while growing in depth and sophistication. State standards based on the NRC framework are likely to show the same pattern of content similarities and dissimilarities with NAEP within and across grades that were revealed in the AIR study (Neidorf et al., 2016) comparing NAEP and NGSS. Results reported in the HumRRO 2021 study of state content standards vis-à-vis NAEP are very similar in that regard (Dickinson et al., 2021). The implication is that at least at the policy level, significant differences exist between NAEP's view of science proficiency and its assessment and the view that has become policy for the preponderance of states and realized via their officially adopted state science standards. Given state science standards adoptions, the current NAEP science framework and assessment may be substantially at variance with a relatively pervasive national perspective on what is desired for students to know and be able to do in science at grades 4, 8, and 12 and how they could be expected to show proficiency via large-scale assessment.

Conclusion 7. The pace at which standards reflecting the NGSS or the NRC framework affects classroom teaching, learning, and assessment has been slow, perhaps not unexpectedly. Evidence shows that adoption of the new standards has been staggered across time since 2013, as has been the design and implementation of state large-scale assessments aligned to those new standards. The latter invariably lag two or more years behind adoption of new state standards. The most recent national survey of science education conducted suggests that little has changed between 2012 and 2018 in science instructional practice (Smith, 2020). Results from the NAEP science assessment from 2009 to 2019 also show little in the way of change in student performance across time (The Nation's Report Card, 2019). One major factor in the slow penetration at the classroom level appears to be limited availability and implementation of professional learning programs for teachers. Although state implementation of large-scale assessments aligned with the NGSS or NRC framework has progressed, and classroom instructional and assessment resources aligned with the NRC framework's vision of teaching, learning, and assessment have become more readily available, the current and future state of classroom practice remains to be determined. Regarding the latter, the National Academies of Science, Engineering, and Medicine (NASEM) is convening a two-day summit in October 2021 at which time the status of implementation of science standards with a focus on areas where additional work may be needed will be discussed. In summary, how far out of alignment the NAEP science framework and assessment may be with science instruction and assessment in most states in 2024 when the current assessment is to be used remains to be seen. It seems reasonable to conclude, however, that significant differences likely will exist in 2028 if the NAEP science framework and assessment are not updated and revised.

Technology and NAEP Science

Conclusion 8. Technology already has had a substantial impact on the NAEP program—and particularly on NAEP science. Both NAEP science and NAEP TEL currently are delivered as digitally based assessments and include new types of tasks that take advantage of some of the affordances of technology for task design, presentation and interaction, data capture, scoring, and analysis. Possibilities exist for capitalizing on the multiple affordances of technology in updating and revising the NAEP science framework and assessment. These include consideration of additional science and technology proficiencies that should be included in the framework, the capacity for their realization in the assessment in the form of tasks and situations that require particular forms of scientific and engineering reasoning, and opportunities for analysis and reporting of those proficiencies in ways that go well beyond overall accuracy. In general, innovative uses of technology offer NAEP science the possibility of leadership in the large-scale science assessment field by providing a vision and examples of how science and technology competence can and should be assessed and reported. Further movement in this direction must take into consideration design and analytic challenges together with equity, cost, and feasibility concerns.

Recommendations

Given the findings described, serious concerns exist about the capacity of the NAEP science assessment to fulfill its mission to provide valid and reliable information about the status of science achievement in the United States in 2028 and beyond unless a detailed review and revision of the NAEP science framework is recommended by NAGB in 2022 and then

pursued by an appropriate framework visioning panel followed by a framework development panel.

The major threat to the validity of NAEP science involves adoption by a preponderance of states of science and technology education standards that differ substantially from the NAEP science framework. Assuming continued implementation of assessments, curriculum materials, instructional practices, and professional learning opportunities aligned with those standards, whether the NAEP science assessment can track the impact of those changes on science achievement and whether NAEP science and/or NAEP TEL will have sufficient instructional sensitivity to reveal what has and has not happened over time when administered in 2028, and even quite possibly beforehand in 2024, is questionable.

Two broad recommendations consistent with these concerns and the related findings contained in this paper follow. For each recommendation, additional commentary is provided regarding matters that should be considered in acting upon each recommendation.

Recommendation 1

The NAEP Validity Studies Panel recommends that the NAEP science framework should be reviewed and revised to reflect contemporary changes in science standards, instruction, and assessment.

In reviewing and suggesting revisions to the science framework:

A. The panels should consider the distribution and focus of the content included in the framework regarding two factors. The first factor involves consideration about whether there should be continuity in the content foci within each domain of science across the grades, in ways similar but not necessarily identical to the disciplinary core ideas in life science, physical science, and Earth and space science described in the NRC framework. The second factor is related to the first and involves the specific set of topics included in each domain and across grades. A shift to this organization of content may allow the NAEP science assessment to provide important trend information across grades in the development of core knowledge in prioritized areas of each of the three major science disciplines.

B. The panels should consider NAEP's current science practices relative to a set of science and engineering practices that may be most important for students to understand and use. Such practices should be articulated in the framework as well as their implications for assessment at each grade level and across grades. Such a consideration includes the extent to which they emphasize active engagement with science and engineering practices, as articulated in the NRC framework, that is, the doing of science and engineering, when applied to science content rather than just knowing about those practices but not necessarily being able to use them.

C. The panels should consider the meaning of science proficiency and how that is expressed via performance expectations that integrate content and practice knowledge consistent with the separate but related considerations of science and engineering content and practices discussed above. Particular attention needs to be given to the

demands of those performance expectations and how they could be represented in assessments that make use of the affordances of technology.

D. The panels should consider the inclusion of technology and engineering content and practices, similar to their inclusion in the NRC framework and NAEP TEL. Further comments on technology and engineering in the NAEP science framework are included below under Recommendation 2.

E. The panels should gather the most recent information on the status of implementation and impact of current state science standards and projections for the remainder of this decade. The panels should seek information on these matters from the Board on Science Education from NASEM, the National Science Teacher Association, the Council of State Science Supervisors, the Science SCASS of the Council of Chief State School Officers, and the American Association for the Advancement of Science.

Recommendation 2

The NAEP Validity Studies Panel recommends that in reviewing and revising the NAEP science framework, consideration should be given to the possible merger of aspects of the TEL framework with the science framework to create an integrated science and technology framework and assessment for administration in 2028.

The NAEP TEL framework and assessment have served useful purposes since their development and initial implementation in 2014. As noted earlier, NAEP TEL is due to be administered twice more at grade 8—in 2024 and again in 2028. Given the representation and integration of technology and engineering with science content domains in contemporary science frameworks and standards, as well as the partial overlap of the latter with the NAEP science and TEL frameworks and assessments, worth considering is whether the most important aspects of the NAEP TEL framework could be included in a revised NAEP science framework.

While the NAEP TEL Framework covers grades 4, 8, and 12, the TEL assessment has been developed only for grade 8. In addition to the limitation of the assessment to a single grade, the TEL construct representation and focus on technology literacy may have lost some of its currency and value in the intervening decade. A review of the complete grades 4–12 framework and the grade 8 assessment seems warranted especially considering existing state standards that include integrated content and practice knowledge focused on technology, engineering, and applications of science across grades 4–12.

A. In reviewing and suggesting revisions to the science framework, the panels should consider NAEP TEL's current content, practices, and forms of assessment for possible inclusion in an updated NAEP science framework and assessment.

B. In considering inclusion of NAEP TEL content and practices in an integrated science and technology framework and assessment, the panels should simultaneously consider what important aspects of the NAEP TEL framework and assessment would be lost if the assessment was discontinued after 2024 and whether continuation of NAEP TEL through 2028 is advisable even if a combined science and technology framework is developed for the 2028 NAEP science assessment.

Considerations of Trend

One hallmark of the NAEP program is its focus on monitoring progress over time and the analysis and reporting of trends in performance. The NAEP science trend extends back to 2009 and NAEP TEL to 2014. Assuming implementation of both current assessments in 2024, there will be 15 years of trend data for science and 10 years for TEL. Given the likely scope of a revision to the NAEP science framework and the implications for the 2028 assessment, as well as the possibility of incorporating aspects of TEL in the new framework and assessment, it seems highly likely that preserving the science or TEL trend through 2028 will not be feasible or advisable. Whether breaking trend in either case in 2028 is both warranted and necessary demands careful attention in deliberations that ensue in NAGB's decisions about revisions to both NAEP science and TEL and their futures. In such deliberations, priority should go to insuring the validity of the revised science framework and assessment for 2028 and beyond. Doing so should not be compromised in a possibly misguided effort to preserve trend at all costs.

REFERENCES

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progression in science: Current challenges and future directions.* Sense.

American Association for the Advancement of Science. (1990). Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology. *Bulletin of Science, Technology & Society*, 10(2), 93–101.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy.* Oxford University Press.

American Association of Medical Colleges. (2012). *MR5 fifth comprehensive review of the Medical Colleges Admission Test (MCAT): Final MCAT recommendations.* https://www.aamc.org/download/273766/data/finalmr5recommendations.pdf

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). *Report of the 2018 NSSME+*. Horizon Research, Inc

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? *Journal of Science Education and Technology, 23*, 160–182.

Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of complex performances in digital environments. *The Annals of the American Academy of Political and Social Science*, *683*(1), 217-232.

Bennett, R. E. (2008). *Technology for large-scale assessment.* ETS Report No. RM-08-10. Educational Testing Service.

Bergner, Y., & von Davier, A. (2019). Process data in NAEP: Past, present and future. *Journal of Educational and Behavioral Statistics, 44(6),* 706–732.

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (Eds.) (2020). *Comparability of large-scale educational assessments: Issues and recommendations.* National Academy of Education.

Bransford, J., Brown, A., Cocking, R., Donovan, S., & Pellegrino, J. W. (2000). *How people learn: Brain, mind, experience and school (expanded edition).* National Academy Press.

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865–883.

Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. *Technology and Engineering Teacher, 70*(1), 30.

College Board. (2009). *Science: College Board standards for success.* https://secure-media.collegeboard.org/apc/cbscs-science-standards-2009.pdf

College Board. (2011a). *AP biology curriculum framework 2012–2013*. https://secure-media.collegeboard.org/digitalServices/pdf/ap/10b_2727_AP_Biology_CF_WEB_110128.pdf

College Board. (2011b). *AP chemistry curriculum framework 2013–2014*. https://secure-media.collegeboard.org/digitalServices/pdf/ap/IN120085263_ChemistryCED_Effective_Fall_2013_lkd.pdf

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform.* Columbia University, Teachers College, Consortium for Policy Research in Education, Center on Continuous Instructional Improvement.

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment.* http://www.ets.org/Media/Research/pdf/psychometric_considerations_white_paper.pdf

Dickinson, E. R., Gribben, M., Schultz, S. R., Spratto, E., & Woods, A. (2021). *Comparative Analysis of the NAEP Science Framework and State Science Standards: Technical report.* Unpublished manuscript. Retrieved from the National Assessment Governing Board

Drasgow, F. (Ed.) (2016). *Technology and testing: Improving educational and psychological measurement.* Routledge.

Duran, R., Zang, T., Sanosa, D., & Stancavage, F. (2020). *Effects of visual representations and associated interactive features on student performance on National Assessment of Educational Progress (NAEP) pilot science scenario-based tasks.* NAEP Validity Studies Panel.

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017) *Validation of score meaning for the next generation of assessments.* Taylor & Francis.

Gane, B. D., Zaidi, S. Z., & Pellegrino, J. W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *European Journal of Education, 53*(2), 176–187.

Gerard, L. F., Ryoo, K., McElhaney, K. W., Liu, O. L., Rafferty, A. N., & Linn, M. C. (2016). Automated guidance for student inquiry. *Journal of Educational Psychology, 108*(1), 60–81.

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment* [Paper presentation]. ITS Invitational Research Symposium on Science Assessment, Washington, DC, United States.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, 38*(2), 53–67.

Lee, Y.-H., Hao, J., Man, K., & Ou, L. (2019). How do test takers interact with simulation-based tasks? A response-time perspective. *Frontiers in Psychology, 24*.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement.* Routledge.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practices, 25*(4), 6–20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Erlbaum.

Mullis, I. V. (2019). *White paper on 50 Years of NAEP use: Where NAEP has been and where it should go next.* NAEP Validity Studies Panel. https://www.air.org/sites/default/files/2021-06/50-Years-of-NAEP-Use-June-2019.pdf

National Assessment Governing Board. (2008). *Science framework for the 2009 National Assessment of Educational Progress.* https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/science/2009-science-framework.pdf

National Assessment Governing Board. (2013). *Technology and engineering literacy framework for the 2014 National Assessment of Educational Progress.* https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/technology/2014-technology-framework.pdf

National Assessment Governing Board. (2014). *Science framework for the 2015 National Assessment of Educational Progress.* https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/science/2015-science-framework.pdf

National Assessment Governing Board. (2018a). *Revision of framework development policy for NAEP assessments.* https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/quarterly-board-meeting-materials/2017-11/12-framework-policy-revision.pdf

National Assessment Governing Board. (2018b). *Technology and engineering literacy framework for the 2018 National Assessment of Educational Progress.* https://files.eric.ed.gov/fulltext/ED594359.pdf

National Research Council. (1996). *National science education standards.* The National Academies Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning.* The National Academies Press.

National Research Council. (2002). *Learning and understanding: Improving advanced study of Mathematics and Science in U.S. high schools.* The National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grade K-8.* The National Academies Press.

National Research Council. (2009). *Learning science in informal environments: People, places, and pursuits*. The National Academies Press.

National Research Council. (2012). *A Framework for K–12 Science Education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

National Research Council. (2013). *Next Generation Science Standards: For states, by states*. The National Academies Press.

The Nation's Report Card. (2019). *See how U.S. fourth-, eighth-, and twelfth-grade students performed in science*. https://www.nationsreportcard.gov/highlights/science/2019/

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*, 183–196.

Neidorf, T., Stephens, M., Lasseter, A., Gattis, K., Arora, A., Wang, Y., Guile, S., & Holmes, J. (2016). *A comparison between the Next Generation Science Standards (NGSS) and the National Assessment of Educational Progress (NAEP) Frameworks in science, technology and engineering literacy, and mathematics. Technical report*. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/subject/science/pdf/ngss_naep_technical_report.pdf

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*, 320–323.

Pellegrino, J. W. (2016). *21st Century science assessment: The future is now*. SRI International.

Pellegrino, J. W., Chudowsky, N, & Glaser, R. (Eds). (2001). *Knowing what students know: The science and design of educational assessment*. The National Academies Press.

Pellegrino, J. W., & Hilton, M. L. (Eds.) (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st Century*. The National Academies Press.

Pellegrino J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*(2), 119–134.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.) 2014. *Developing assessments for the Next Generation Science Standards*. The National Academies Press.

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science, 372*(6540), 338–340.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Kluwer Academic.

Smith, P. S. (2020). What does a national survey tell us about progress toward the vision of the NGSS? *Journal of Science Teacher Education, 31*(6), 601–609.

Way, D., & Strain-Seymour, E. (2021). *A framework for considering device and interface features that may affect student performance on the National Assessment of Educational Progress.* NAEP Validity Studies Panel.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

Wilson, M. R., & Bertenthal, M. W. (Eds.). (2005). *Systems for state science assessments*. The National Academies Press.

Zhai, X. (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology, 30*(2), 1–11.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459.

Zhai, X., & Pellegrino, J. W. (in press). Large-scale assessment in science education. *Handbook of research on science education* (Vol. III). Routledge.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151.