

Assessment Development Committee

November 17, 2022

3:00 – 5:00 pm ET

North Carolina



AGENDA

3:00 – 3:05 pm	Welcome and Review of Agenda <i>Patrick Kelly, Chair</i> <i>Christine Cunningham, Vice Chair</i>	
3:05 – 3:50 pm	Project Update: 2028 NAEP Science Assessment Framework <i>Mark Loveland, WestEd</i> <i>Taunya Nesin, WestEd</i>	Attachment A
3:50 – 4:10 pm	Initial Plans and Next Steps for 2030 NAEP Writing Assessment Framework Update <i>Sharyn Rosenberg, Assistant Director for Assessment Development</i>	Attachment B
4:10 – 4:30 pm	Perspectives on Gradual, More Frequent Updates to NAEP Assessment Frameworks <i>Sharyn Rosenberg</i>	Attachment C
4:30 – 5:00 pm	Initial Results from Pretesting of NAEP Mathematics Items (CLOSED) <i>Dana Kelly, NCES</i>	Attachment D
Information Item	Item Review Schedule	Attachment E

Quarterly Progress Report

2028 NAEP SCIENCE ASSESSMENT FRAMEWORK UPDATE

Project Overview

In July 2022, the Governing Board awarded a contract to WestEd to conduct an update of the NAEP Science Assessment Framework and the companion Assessment and Item Specifications. The goal of the Science Framework project is to update the NAEP Science Framework documents through the work of a 30-person Steering Panel, a 20-person Development Panel, an 8-person Educator Advisory Committee (EAC), and a 6-person Technical Advisory Committee (TAC). This will be accomplished through an initial Steering Panel meeting, three subsequent Development Panel meetings, conducting ongoing and targeted outreach efforts to gather public comment on draft versions of the documents, and production of a final updated Science Assessment Framework and Assessment and Item Specifications for Science to be submitted to the Governing Board by October 2023.

The Science Framework update is to be conducted using a combination of external experts and Science specialists within WestEd. To complete this work, WestEd has partnered with Safal Partners, to assist with gathering and analyzing public comment feedback, and Cary I. Sneider Consulting, to assist with writing the narrative text of the framework (the statement of work indicated that the panel members should focus on developing a substantive outline of the assessment framework; WestEd staff and consultants will produce a draft of the narrative text based on the outline to be reviewed and edited by the panel members rather than having the panel serve as the primary authors).

Project Team

The Project Management Team consists of Steve Schneider, Mark Loveland, Taunya Nesin, Marianne Perie, and Megan Schneider. As project director, Steve Schneider provides day-to-day leadership, guidance, and liaising with the Governing Board. Dr. Schneider has over 40 years of science, mathematics, and technology education experience and led WestEd's four previous Framework development and update projects. Project co-director, Mark Loveland, and Science Content Lead, Taunya Nesin, have oversight for all programmatic activities. Dr. Loveland was project coordinator for the Technology and Engineering Literacy (TEL) Framework development project and project co-director for the Mathematics and Reading Framework updates. A panel leadership team of four (Aneesha Badrinarayan, Jenny Christian, Nancy Hopkins-Evans, and Joseph Krajcik) will work with WestEd to plan meetings, facilitate panel discussions, and represent the panel's work to the Governing Board. Together, they and Dr. Nesin will lead the Steering and Development Panel activities, and Dr. Nesin will also

coordinate the EAC. Measurement Lead, Dr. Perie, will coordinate the TAC. Ms. Schneider will serve as Project Manager, documenting all project activities. In addition to the project leaders, the broader project team includes additional science subject matter experts, members of the science measurement team, project coordinators, and research assistants. Additional information about the project team and participants in the framework update can be found at: www.naepframeworkupdate.org.

Project Timeline

The project timeline, first identified in the project kickoff meeting and updated as needed, describes WestEd's project management and coordination of panel, EAC, and TAC activities to update the NAEP Science Assessment Framework and Assessment and Item Specifications. The bulk of the framework update work will be carried out by the Framework Steering and Development Panels. Comprised of 30 individuals representing various stakeholder groups, the Framework Steering Panel will formulate recommendations for updating the Science Framework, based on the state of the field. Twenty members of the Steering Panel will constitute the Framework Development Panel. The Development Panel is charged with developing the draft outlines of the project documents and engaging in the detailed deliberations to determine how to reflect the Board charge and Steering Panel recommendations in an updated framework. Dates for the Steering Panel meeting and the three Development Panel meetings have been finalized for October and December 2022 and January and June 2023. Additional work will take place asynchronously and via webinars.

Preparatory work for the Framework Panel activities has been extensive. WestEd has prepared a project Design Document which serves as the blueprint for the project processes, describing outcomes and metrics, and as the touchstone for quality assurance monitoring. Additionally, a Technical Advisory Committee comprised of six technical experts will respond to technical issues raised during panel deliberations. A new addition to the framework update process, an Educator Advisory Committee, will provide additional guidance from teachers and administrators.

Panelist Selection

Using processes outlined in the Design Document, WestEd worked in consultation with Governing Board staff and Widmeyer/FINN Partners (communications contractors to the Board) to support the Assessment Development Committee (ADC) recommendation of 30 members of the Steering and Development Panels. WestEd staff worked with Board staff and the ADC to suggest criteria to evaluate the 120 nominations submitted to the Board in response to the open call for panelist nominations. The following factors were prioritized in constructing a balanced panel: individuals specifically nominated to represent a national organization, given the critical need to engage various constituencies; panelist role; experience and expertise overall and the specific sub-content areas covered by the framework; demographic characteristics, including race, gender, and geography; previous experience with and stance on the Next Generation Science Standards (NGSS), including both NGSS developers and critics, and practitioners in states that have adopted NGSS standards, NGSS-alike standards, and non-NGSS standards; and diverse perspectives on issues relevant to the Board charge. The Assessment Development Committee finalized their recommended slate of panelists on August 23, and the

recommended slate of panelists and potential alternates was unanimously approved by the Executive Committee on August 29. All 30 invited panelists agreed to participate on the framework panels.

Panelist Deliberation

The work of the panels, EAC, and TAC will be informed by the Governing Board charge to the panels and a compilation of resources. The Board charge is intended to serve as a springboard for discussion by the Framework panels and identify issues that are a priority for the Board in the update of the NAEP Science Framework. The Board charge and Resource Compilation draw from the current NAEP Science Framework and Specifications documents, national and international standards, state frameworks and standards, extant assessments, reports, research on science education and assessment, and other resources. These resources highlight reports, high-level presentations, and associated academic research papers, including a comparison study of the NAEP Science Assessment Framework with the current generation of state standards and assessments in science. In preparation for the Steering Panel meeting, WestEd project staff conducted an initial orientation of the panelists to their responsibilities in updating the NAEP Science Framework. A successful Steering Panel meeting was held on October 17-18, 2022.

Next Steps

Over the course of the Development Panel meetings, WestEd will facilitate the updating of the Science Framework documents, producing draft versions of the Assessment Framework and Assessment and Item Specifications. The Development Panel will be responsible for developing detailed outlines of the updated framework and specifications, with WestEd later developing draft documents based on the panel outlines. Outreach will be conducted primarily by WestEd and Safal Partners, in conjunction with Board staff, Widmeyer/FINN Partners, and with assistance from collaborating organizations. Feedback on the detailed outline and subsequent drafts will come from member organizations represented on the two panels, other organizations, and the public. Organizations may choose to convene meetings, gather feedback via a web-based portal, or have members contact a web site. In all instances, stakeholder groups will follow procedures for securing input and ensuring representation of diverse views. WestEd and Safal staff will tabulate feedback, make recommendations for revisions addressing the feedback, and coordinate the development of final versions of the framework documents to be submitted to the Governing Board.

Milestones

The major milestones of the project are summarized below.

Milestone	Dates
Project Kickoff Meeting	July 2022
Project Timeline Development	July 2022
Design Document Development	July – August 2022
Identification of Steering and Development Panelists and TAC Members	July – August 2022
Resource Compilation Development	September – October 2022
Steering Panel Meeting	October 2022
Development Panel Meetings	December 2022, January & June 2023
Convene TAC and EAC	10 meetings for each, 2-3 weeks prior to and after each panel meeting and prior to submission of draft framework documents
Draft Versions of Framework Outline and Other Documents	February – July 2023
Gather Public Comment on Framework Outline	March – April 2023
Develop Final Versions of Framework Documents	June – October 2023
Submit Final Process Report	December 2023

November 2022 Assessment Development Committee Meeting

During the upcoming ADC meeting on November 17, Project Co-Director Mark Loveland and Science Content Lead Taunya Nesin will brief the Committee on recent and upcoming project activities, including key takeaways from the Steering Panel meeting held on October 17-18.

Initial Plans and Next Steps for Updating the 2030 NAEP Writing Assessment Framework

According to the [NAEP Assessment Schedule](#), the NAEP writing assessment will next be administered in 2030 and updates to the framework will be considered for this administration. The NAEP U.S. History and Civics Assessment Frameworks are also scheduled to be updated for the 2030 administrations of those assessments, but implementation of the writing framework takes longer so that update needs to be launched first. The current contract #91995922C0001 with WestEd to develop recommendations for the 2028 NAEP Science Assessment Framework update includes an option to develop recommendations for one additional NAEP framework, beginning in summer 2023.

In accordance with the [Board policy on Assessment Framework Development](#), the first step in the process of updating a framework is to seek public comment on whether and how the existing framework should be changed. Board staff plans to launch this call for public comment shortly after the November Board meeting. Other initial steps include commissioning papers from content experts to provide guidance to the Board. Information from the initial public comment and expert papers would be presented to the Board in March 2023, for the purpose of informing whether and how to proceed with a Board charge to the framework panels in May 2023.

The current [NAEP Writing Assessment Framework](#) was adopted in 2007 for implementation in 2011. The Board made a policy decision at the time to begin new trend lines without attempting to perform bridge studies to determine the feasibility of connecting results based on the previous framework. The current framework focuses on “writing on computer,” replacing the previous framework which focused on writing by hand; the mode of administration in the current framework is not incidental but is conceptualized as being a central part of the construct.

In 2011, the NAEP writing assessment was administered at the national level at grades 8 and 12; results from that administration can be found at: https://www.nationsreportcard.gov/writing_2011/. In 2017, the NAEP writing assessment was administered at the national level at grades 4 and 8, but the results were not able to be reported due to technical concerns related to changes in the device and platform used to administer the assessment; more information is available at: <https://nces.ed.gov/nationsreportcard/writing/2017writing.aspx>. The technical issues encountered in 2017 make it infeasible to maintain trend lines with the 2011 results in the future, regardless of whether or not a new framework is adopted.

In this session, Assistant Director for Assessment Development Sharyn Rosenberg will present a brief overview of initial considerations and seek ADC feedback on what additional information is needed to inform the decision about whether and how to proceed with updating the 2030 NAEP Writing Assessment Framework.

Perspectives on Gradual, More Frequent Updates to NAEP Assessment Frameworks

One of the Governing Board's legislatively mandated responsibilities is to develop assessment objectives for NAEP, which is operationalized through assessment frameworks and test specifications. The National Center for Education Statistics (NCES) uses the frameworks and specifications to develop items and test forms for administering the assessments. The Board exercises its authority to develop and update the NAEP frameworks through its policy on [Assessment Framework Development](#). This policy was recently updated in March of this year, but there has been continued interest in re-examining the current policy to consider whether and how smaller changes to NAEP frameworks might occur on a more frequent basis rather than waiting 10 years and making larger changes all at once.

In preparation for the May 2022 Assessment Development Committee (ADC) meeting, Assistant Director for Assessment Development Sharyn Rosenberg prepared a [paper outlining various questions and considerations](#) that would need to be addressed to pursue this idea. The Committee discussed the paper and supported the Board staff proposal to commission consultant papers on this topic. Through a contract with the Manhattan Strategies Group (MSG) and subcontract with the Human Resources Research Organization (HumRRO), papers were commissioned from six consultants who were intended to represent different perspectives and experiences on this topic:

- Carol Jago, former Governing Board member and ADC Chair
- Andrew Ho, former Governing Board member and Chair of the Committee on Standards, Design and Methodology (COSDAM)
- Jessica Baghian, former state leader in Louisiana
- Stanley Rabinowitz, psychometrician with extensive experience working on state assessments and the national exams in Australia
- Ada Woo, psychometrician with extensive experience working on certification exams
- Alicia Alonzo, former member of the NAEP Science Standing Committee, and the committee that recently updated the 2023 TIMSS Science Framework using a process similar to what has been proposed for updating NAEP assessment frameworks

Independent of the consultant papers commissioned by Board staff, Lorrie Shephard of the NAEP Validity Studies (NVS) Panel has been working on a comprehensive white paper on this same topic for quite some time, and it is being published on the [NVS website](#) in early November. The NVS paper is included in this attachment, followed by the consultant papers.

Board staff plan to organize a technical panel meeting in January 2023 with the authors of the seven papers and NCES to discuss ideas raised in the papers for the purpose of informing recommendations for how to proceed with the Board policy and procedures for updating NAEP frameworks. Minutes summarizing the technical panel meeting will be shared with ADC in February 2023, and a webinar will be organized to provide an opportunity for ADC members to ask questions of panelists.

During the November ADC meeting, Committee members will discuss initial reactions to the papers and suggest questions to pose to the authors in the technical panel meeting.

White Paper: NAEP Framework and Trend Considerations

Lorrie Shepard
University of Colorado Boulder

October 2022
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:

Keena Arbuthnot
Louisiana State University

Peter Behuniak
Criterion Consulting, LLC

Jack Buckley
American Institutes for Research

Phil Daro
*Strategic Education Research Partnership
(SERP) Institute*

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Ina V.S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

James Pellegrino
University of Illinois at Chicago

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado Boulder

David Thissen
University of North Carolina, Chapel Hill

Gerald Tindal
University of Oregon

Sheila Valencia
University of Washington

Denny Way
College Board

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
1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
Email: naepvaliditystudies@air.org

CONTENTS

INTRODUCTION.....	3
NAEP’S CURRICULUM-NEUTRAL, BALANCED FRAMEWORKS	5
THE IMPORTANCE OF TREND DATA FOR MONITORING EDUCATIONAL PROGRESS.....	7
A BRIEF HISTORY OF NAEP FRAMEWORK AND TREND CHANGES	10
NAGB’S FRAMEWORK DEVELOPMENT PROCESSES	13
NAGB’S PRACTICES TO PROTECT TREND	14
ARGUMENTS FOR AN “EVOLUTIONARY” APPROACH TO FRAMEWORK REVISIONS.....	15
Subject-Matter Committees to Guide Framework Revisions and Updates	15
Processes to Inform and Name Construct Revisions	16
Recommendations for Implementing an Evolutionary Approach to Framework Revisions.....	20
What Are the Cautions or Downsides to an Evolutionary Approach?	20
HOW DECISIONS ABOUT FRAMEWORKS AND TREND CAN OBSCURE OR ILLUMINATE PROGRESS	21
Example: Long-Term Trend NAEP versus Main NAEP	21
Example: Reweighting of TUDA Mathematics Results to Align with State Assessment Content.....	27
ALIGNMENT AND BRIDGE STUDIES TO EVALUATE CONSTRUCT SHIFT	30
Recommendations Regarding Bridge Studies to Evaluate Construct Shift.....	31
ARGUMENTS FOR AND AGAINST “BREAKING TREND”	32
NCES SPECIAL STUDIES	33
CONCLUSION	35
REFERENCES.....	37

INTRODUCTION

Two critically important features of the National Assessment of Education Progress (NAEP) are its subject-matter *frameworks* and its reporting of *trends* or changes in achievement over time. The purpose of this white paper is to provide the National Center for Education Statistics (NCES), the National Assessment Governing Board (NAGB), and the NAEP research and policy community with a summary of issues and evidence affecting framework and trend policies. The central argument of the paper, consistent with the recommendations from several expert panels (National Academies of Sciences, Engineering, and Medicine [NASEM], 2022; NCES, 2012; National Academy of Education [NAEd], 1992), is that *NAGB should develop an explicit policy to enable a more “evolutionary” approach to framework revisions*. Such a policy would protect trend and, at the same time, ensure the relevance of construct¹ representation by providing for ongoing, incremental revisions to frameworks.

As NAGB and NCES know well, there is a fundamental tension between keeping frameworks up to date and, at the same time, maintaining the stability and comparability of achievement data over time. To be appropriate for the nation as a whole, NAEP requires curriculum-neutral frameworks that broadly reflect what is currently being taught in a subject-matter domain and that also have sufficient reach to anticipate disciplinary learning goals intended for the future. Thus, NAEP frameworks must be revised or they will become outdated, unable to capture learning goals at the forward edge of disciplinary standards. Yet, to measure change, assessments must stay the same.

The first several sections of the paper provide background information on the particular requirements of NAEP frameworks, the importance of trend data for monitoring “educational progress,” a brief history of NAEP framework and trend changes, and then NAGB’s framework development policy, which includes cautions and safeguards to protect trend. The middle part of the paper presents the more detailed argument in favor of an evolutionary approach to framework revisions. Who has recommended this approach and why? How would expanding the scope of work for standing subject-matter committees facilitate an evolutionary approach, and how might steps in an evolutionary approach articulate with NAGB’s existing review processes? Under what circumstances would conceptual recommendations for revisions be the impetus for NAGB to invoke its full-scale process for developing new frameworks?

Of course, there are potential downsides to an evolutionary approach to framework revisions, which are addressed in the final sections of the paper. Studies are reviewed to show how starting a new trend, when the definition of a construct changes, can heighten documentation of educational progress as assessed by the new construct. The reverse is also true: failing to change the definition of a construct in keeping with changes in the field can obscure or fail to detect evidence of educational progress. The use of bridge studies to evaluate construct shift is reviewed with particular attention to the greater difficulties that arise when new and old constructs are highly correlated and when instructional changes in response to new disciplinary standards occur gradually. The arguments for and against “breaking trend” are highlighted very briefly with responses from the evolutionary

¹ The word “construct” is a measurement term referring to the underlying competencies addressed by an assessment; it includes both the content dimensions and the mental processes elicited.

perspective, which presumes that comparisons in adjacent assessment cycles and over the short term are more important than comparisons spanning many decades. Lastly, an argument is made for special NCES Research and Development studies to address policy uses of NAEP data when framework and trend decisions affect construct definitions in ways that are likely to lead to inappropriate policy interpretations.

The concluding section of the paper recapitulates recommendations focused specifically on requirements for a new framework policy, in addition to the current policy—which would address the kinds of evidence to be collected and review processes needed to enable smaller and more frequent framework revisions. NAGB would continue to review proposed minor changes and would need to have a timely process for deciding when proposed minor changes were of sufficient import to warrant invoking the full-scale framework development process.

NAEP'S CURRICULUM-NEUTRAL, BALANCED FRAMEWORKS

Assessment frameworks are broad overview documents that serve as a blueprint to guide assessment development. They lay out the knowledge and skills to be covered by an assessment and provide examples of the types of questions that should be included. Typically, achievement constructs or subject-matter domains are represented as two-dimensional structures, showing both content strands and cognitive processes. In mathematics, for example, the content strands are Number Properties and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra, and cognitive processes are assessed with three different levels of mathematical complexity—ranging from simple recall to higher levels of mathematical reasoning and analysis.

Development of new assessment frameworks is governed by NAGB's *General Policy: Conducting and Reporting the National Assessment of Educational Progress* (2013) and by its specific *Assessment Framework Development Policy Statement* (2022). These policy documents wisely identify several key ideas that are important to highlight here:

- The domain of knowledge and skills represented by a framework must be broad, not favoring one particular curriculum over another.
- Frameworks must also encompass both current learning goals and advances occurring in the field.
- Item formats convey substance and are therefore part of framework descriptions.

NAEP is intended to serve as a national monitor, reporting on the educational achievement of the nation as a whole as well as for designated subgroups. When NAEP began in the late 1960s, there were no frameworks, as the intention was to report on individual test items of interest. Assessment blueprints were not developed until the early 1980s, when they were needed to support total score reporting. However, the more visible role for NAEP frameworks began in the 1990s with the federal legislation that created both NAGB and State NAEP. In the ethos of the time, NAEP was called upon to direct the nation's attention to important education goals but also—to forestall establishment of a national curriculum—should “assert the importance of instructional pluralism” (Glaser & Bryk, 1987). Reconceptualization of NAEP content also called for assessment of “higher-order thinking” skills as well as basic competencies. This new, more ambitious way of conceptualizing content domains was thought of as the *union* of multiple curricula rather than the lowest common denominator or *intersection* of possible curricula. Today, NAGB's current framework policy upholds this important principle of curriculum neutrality:

The framework shall focus on important, measurable indicators of student achievement to inform the nation about what students know and are able to do without endorsing or advocating a particular instructional approach (NAGB, 2022, p. 4)

With the challenge in the 1990s for NAEP to assess more complex levels of thinking and reasoning came the recognition that NAEP would need to develop new types of test questions and reduce the proportion of multiple-choice formats that often tapped only rote memorization. For this reason, illustrative items have continued to be an important aspect of

how assessment frameworks are explicated. At the same time, it was acknowledged that the visioning of the new NAEP—what came to be called “Main NAEP”—reached beyond what was then being taught in schools. Giving a test over material that had not been taught would be unfair in an end-of-course examination, but not for a national monitor. To be able to monitor progress toward important goals, it was critical that they be represented in the assessment. Thus, the National Assessment must play a leadership role by anticipating “important advances in the field.” At the same time, NAEP cannot swing wildly, focusing only on new learning goals, if those new ideas are not yet being taught, because it would miss measuring what students are, in fact, learning. A reason to keep assessments the same over time is not, then, just to maintain trend but also to continue to assess what is familiar and most prevalently being taught in the present moment. Thus, NAEP frameworks are guided by this balancing act, between what is and aspirations for the future, as summarized in NAGB’s policy:

The framework shall reflect current curricula and instruction, research regarding cognitive development and instruction, and the nation’s future needs and desirable levels of achievement. This delicate balance between “what is” and “what should be” is at the core of the NAEP framework development process. (p. 7).

THE IMPORTANCE OF TREND DATA FOR MONITORING EDUCATIONAL PROGRESS

Because they define assessment content, NAEP frameworks also affect the reporting of trend data. As suggested by the words “educational progress” in its name, NAEP’s primary purpose is to serve as an independent monitor—reporting on the status of achievement in the nation at the time of each assessment, but especially tracking trends in achievement over time.

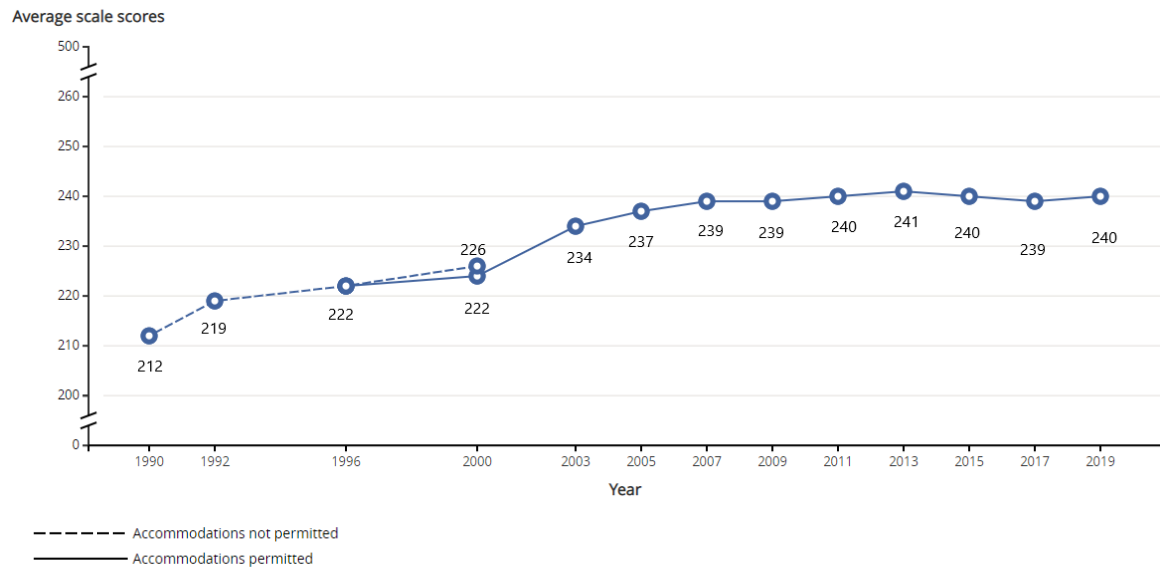
For example, results from the 2019 mathematics assessment showed substantial gains compared to results from 1990, an increase in average score of 27 points at grade 4 and 19 points at grade 8 (exhibit 1). At grade 4, this meant that students at the 50th percentile in 2019 were performing at roughly the same level as 75th percentile students in 1990. These increases occurred primarily in the decades from 1990 to 2009, however, with much flatter trends or even declines for lower performing groups of students from 2009 to 2019. In reading, the improvements from 1992 to 2019 were slight; average scores were four points higher at grade 4 and three points higher at grade 8. These small reading gains over the course of nearly three decades might have been stronger but for the flattening and downturn in performance from 2009 to 2019. This pattern of losing ground was observed across the performance continuum in reading, with the exception of students performing at the 90th percentile at grade 4.

Being able to provide technically accurate information on achievement improvements or declines illustrates the importance of NAEP trend data for policy purposes. Note that these comparisons of changes in student performance at different times require that the assessments stay the same. *Thus, there is a tension between desires to revise or update frameworks, based on curricular or technological innovations or new research, and the need to keep frameworks the same to “protect trend” and enable important comparisons over time.* As described further below, the NAGB Assessment Framework Development policy clearly acknowledges this tension. When soliciting “input from experts to determine if changes are warranted,” the Board’s Assessment Development Committee should make clear “the potential risk to trends and assessment of educational progress posed by changing frameworks” (2022, p. 7).

The Importance of Trend Data for Monitoring Educational Progress

Exhibit 1. NAEP score trends for math and reading grades 4 and 8, national public

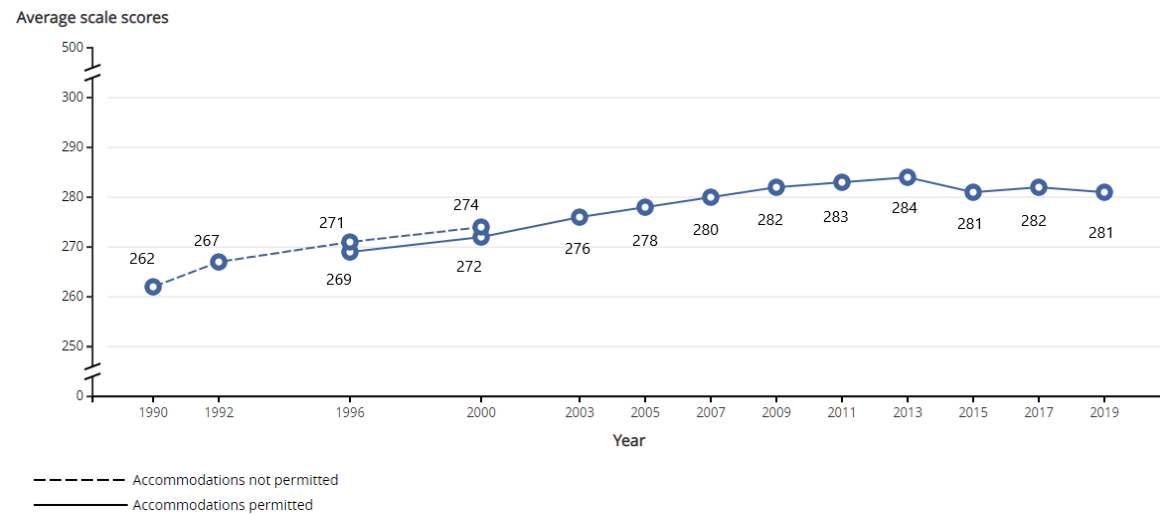
Mathematics Grade 4



NOTE: The NAEP Mathematics scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 1990, 1992, 1996, 2000, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, and 2019 mathematics assessments.

Mathematics Grade 8



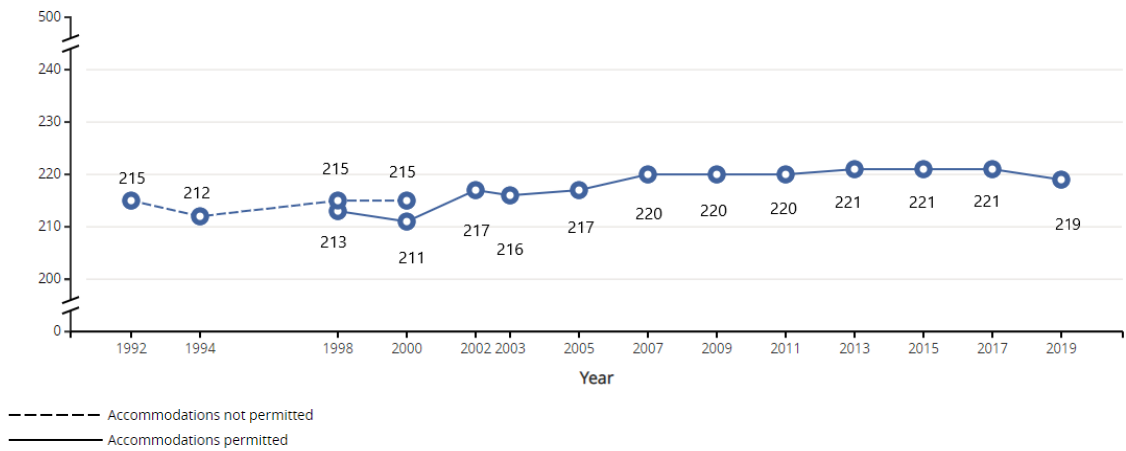
NOTE: The NAEP Mathematics scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 1990, 1992, 1996, 2000, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, and 2019 mathematics assessments.

The Importance of Trend Data for Monitoring Educational Progress

Reading Grade 4

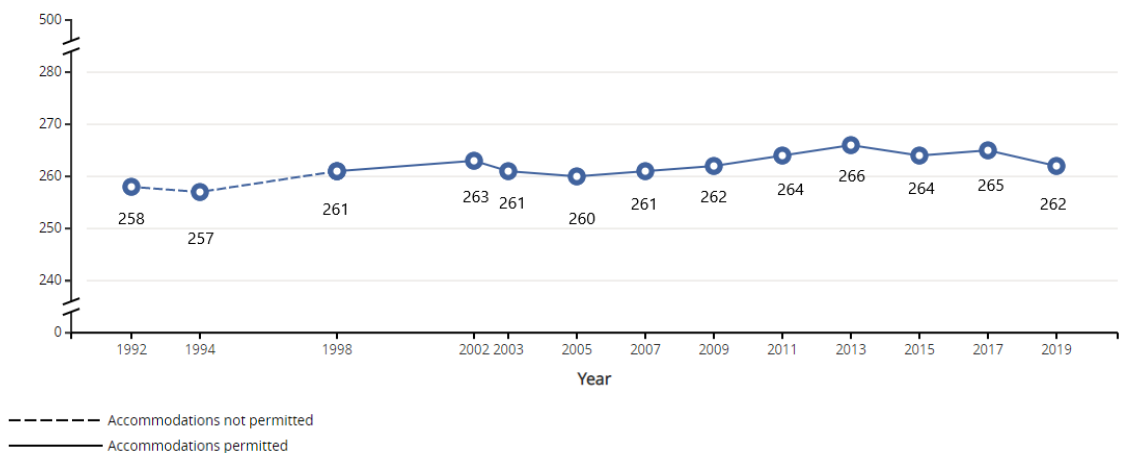
Average scale scores



NOTE: The NAEP Reading scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 1992, 1994, 1998, 2000, 2002, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, and 2019 mathematics assessments.

Reading Grade 8

Average scale scores



NOTE: The NAEP Reading scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 1992, 1994, 1998, 2002, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, and 2019 mathematics assessments.

A BRIEF HISTORY OF NAEP FRAMEWORK AND TREND CHANGES

New NAEP frameworks created in the 1990s under the newly formed Governing Board represented an intentional departure from prior assessments. As a result, new baseline measurements needed to be established and new trend lines begun. To understand this historical context, it is important to remember that the call to measure higher-order thinking skills was part of a nationwide educational excellence movement aimed at teaching to more challenging “world class standards.” The movement included policy leaders, chief executive officers and governors, cognitive scientists, and subject-matter experts. The 1988 law (Augustus F. Hawkins Act, 1988) that created NAGB and added State NAEP derived many of its main components from the Alexander and James report (1987) written by the study group chaired by Lamar Alexander, then-governor of Tennessee and chair of the National Governors Association (NGA). This revisioning of NAEP occurred at a time of intense, national political attention on education, as evidenced by the 1989 Education Summit led by President George H. W. Bush and subsequent NGA leader Governor Bill Clinton.

Also in 1989, the National Council of Teachers of Mathematics produced a set of mathematics standards aimed at developing students’ abilities to reason mathematically, engage in problem solving, and communicate mathematically. These standards, focused on sensemaking, were strongly influenced by the cognitive revolution and decades of research eventually brought together in the National Research Council’s (NRC’s) consensus report *How People Learn* (1999). Findings from cognitive research refuted ideas from prior behaviorist and hereditarian theories, which held that higher levels of reasoning had to be postponed until basics were mastered and that only an elite few were capable of mastering more advanced academic work. It was a breakthrough, widely shared, to acknowledge that intellectual abilities are developed through learning opportunities, including specific instructional practices, like talking with classmates about how to solve a math problem. Similar changes—focused on higher levels of thinking, reasoning, drawing connections, and inquiry—led to new content standards being developed in each subject area. For example, the *National Science Education Standards* were developed by the NRC in 1996. In this same milieu, NAGB convened experts and undertook a consensus process to develop new frameworks for NAEP, including the Mathematics Framework developed in 1990; Reading, in 1992; History and Geography, in 1994; and Science, in 1996.

Main NAEP frameworks developed in the 1990s were not identical to the content standards put forward by disciplinary professional societies, but they shared important ideas about levels of cognitive complexity needed in instruction and correspondingly in assessment. NAEP frameworks laid out subject-matter domains in ways that were clearly more challenging than had been true for prior NAEP blueprints and item pools. The idea was not simply to make assessments more difficult but, rather, to engage students with content in a way that deepened their understanding and knowledge use. Because of these intentional framework changes, new trend lines were begun, and a separate investment was made to maintain and monitor what then became known as NAEP Long-Term Trend (LTT), carrying forward earlier assessment methods. It is significant to note that different conceptualizations of assessment content can produce different assessment results and different pictures of educational progress. This point is illustrated in a later section with data comparing LTT and Main NAEP results in mathematics.

Exhibit 2 provides a summary of framework changes and trends in all of NAEP's subject areas. Since the 1990s, there have been only a few occasions when the Governing Board decided to "break trend." This happens when a new framework changes the definition of the test construct sufficiently to make comparisons to prior years inappropriate. For example, a new trend was started in 2005 for grade 12 mathematics. The grade 12 Mathematics Framework was substantially revised in 2005 and again in 2009, but a bridge study found sufficient similarities between 2005 and 2009 to maintain the trend from 2005 onward. (The methodology involved in bridge studies is described later in the paper.) For science, NAEP frameworks were substantially revised in 2009 for grades 4, 8, and 12. The changes made in the new 2009 Science Framework were judged to be substantial enough to warrant starting new trends. Content changes included the addition of space science and crosscutting concepts among the Life, Physical, and Earth and Space Sciences. The new Science Framework also explicitly elaborated on the assessment of scientific practices, including the ability to use scientific principles, scientific inquiry, and technological design.

Over the past 20 years, NAGB has also approved other updates to assessment frameworks that did not disrupt the reporting of trends. In most cases, minor revisions to frameworks are believed not to alter the underlying structure or meaning of the construct being assessed and, therefore, do not require breaking trend. In those instances in which a new framework development process is undertaken, as described in a later section, it becomes more likely that the construct could change. Bridge studies, then, provide empirical checks to ensure that old and new assessments are sufficiently similar so as to permit common scaling and comparisons over time. Note that bridge studies are also conducted when administrative changes might alter either the meaning or the difficulty of the assessment.

Two examples of administrative changes are the use of accommodations for English learners and students with disabilities and the switch from paper-and-pencil tests to digital test delivery. Bridge studies in 1996 indicated that accommodations had created a slight change in the difficulty of assessments but had not altered the construct being measured. As a result, trend data before and after accommodations are shown in the same trend graphs but are distinguished (usually with dotted and solid lines, respectively). Bridge studies for digitally administered assessments have not found construct shift, and small but consistent mode effects were adjusted psychometrically so that data could be reported as a continuous trend (Jewsbury et al., 2020).

A Brief History of NAEP Framework and Trend Changes

Exhibit 2. Framework changes and trend continuation, by subject

		1990	1992	1994	1996	1998	2000	2001	2002	2003	2005	2006	2007	2009	2010	2011	2013	2014	2015	2017	2018	2019
Reading	4		■●	□		□	□		■	□	□		□	■		□	□		□	□		□
	8		■●	□		□			■	□	□		□	■		□	□		□	□		□
	12		■●	□		□			■		□			■			□		□			□
Math	4	■●	□		■		□			□	■		□	□		□	□		□	□		□
	8	■●	□		■		□			□	■		□	□		□	□		□	□		□
	12	■●	□		■		□				■●			■			□		□			□
Science	4				■●		□				□			■●					□			□
	8				■●		□				□			■●		□			□			□
	12				■●		□				□			■●					□			□
Writing	4		■●			■●			□							■●				□		
	8		■●			■●			□				□			■●				□		
	12		■●			■●			□				□			■●				□		
U.S. History	4			■●				□				■				□						
	8			■●				□				■				□		□			□	
	12			■●				□				■				□						
Geography	4			■●				□								□						
	8			■●				□								□		□			□	
	12			■●				□								□						
Civics	4					■●						□			□							
	8					■●						□			□			□			□	
	12					■●						□			□							
TEL	4																	■●			□	
	8																	■●			□	
	12																	■●			□	

KEY: ■ = New framework; □ = Same framework; ■ = Modified framework; ● = New trend.

NOTE: TEL is Technology and Engineering Literacy.

SOURCE: Adapted (with additions and reformatting) from Nellhaus et al., 2009.

NAGB'S FRAMEWORK DEVELOPMENT PROCESSES

As summarized above, the NAGB *Assessment Framework Development Policy Statement* (2022) identifies principles that guide how content domains are to be laid out broadly and then further delineated with item specifications and example items. As described next, NAGB's policy statement also addresses the processes by which new frameworks are to be developed, including the comprehensive process by which stakeholders are involved, review mechanisms for deciding when a new framework is needed, and information resources that should inform framework development.

Principle 2 of NAGB's policy calls for an inclusive consensus process: "The Governing Board shall develop and update frameworks through a comprehensive, inclusive, and deliberative process that involves active participation of stakeholders" (NAGB, 2022, p. 5). The work of conceptualizing a new framework is undertaken by a broadly representative Framework Steering Panel. A subset of the Steering Panel then serves as the Framework Development Panel to draft the new framework along with more detailed assessment and item specifications. A balanced and widely supported final document is further ensured by public comment, Board review, and revision processes with special attention to feedback from teachers, curriculum developers, and researchers in the specific content area.

NAGB's review process is undertaken at least once every 10 years (NAGB, 2022) to determine whether existing frameworks remain sufficiently relevant or if changes are needed. The Assessment Development Committee (ADC), a subset of NAGB, commissions reviews by content experts and then makes a recommendation to the Board as to whether minor revisions are needed; or, if major revisions are called for, the full framework development process is invoked. For example, in 2018 the ADC solicited reviews of the 2017 Mathematics Framework by mathematics experts. Although disparities noted were varied—for example, between the existing NAEP framework and current research or compared to state standards, content experts agreed that substantial revisions were needed, which prompted the development process for a new Mathematics Framework for the 2026 assessment. It is worth noting the significant time interval between initial review and implementation of a new framework (almost a decade), due to the time involved in framework development and then assessment development in keeping with a new blueprint.

Once the decision has been made to develop a new framework, principle 2 of NAGB's policy identifies authoritative resources that should be taken into account:

- i. The framework panels shall consider a wide variety of resources during deliberations, including but not limited to relevant research, trends in state and local standards and assessments, use of previous NAEP results, curriculum guides, widely accepted professional standards, scientific research, other types of research studies in the literature, key reports having significant national and international interest, international standards and assessments, other assessment instruments in the content area, and prior NAEP frameworks, if available.

Typically, these materials have been compiled by NAGB staff and by the contractor charged with convening and supporting the Steering Panel.

NAGB'S PRACTICES TO PROTECT TREND

NAGB does not have a separate policy about maintaining trend. Instead, the tension between innovation and continuity with the past is addressed in NAGB's general policy and as part of the framework development policy. Then empirical bridge studies, as described later, are undertaken to ensure that old and new item pools can be scaled together. The general policy stance toward preserving trend is as follows:

For NAEP to measure trends in achievement accurately, the frameworks (and hence the assessments) must remain sufficiently stable. However, as new knowledge is gained in subject areas, the information and communication technology for testing advances, and curricula and teaching practices evolve, it is appropriate for NAGB to consider changing the assessment frameworks and items to ensure that they support valid inferences about student achievement. But if frameworks, specifications, and items change too abruptly or frequently, the ability to continue trend lines may be lost prematurely, costs go up, and reporting time may increase. For these reasons, NAGB generally maintains the stability of NAEP assessment frameworks and specifications for at least ten years. NCES assures that the pool of items developed for each subject provides a stable measure of achievement for at least the same ten-year period. In deciding to develop new assessment frameworks and specifications, or to make major alterations to approved frameworks and specifications, NAGB considers the impact on reporting trends. (NAGB 2013, pp. 6–7)

In addition, because of the importance of trend data to the mission of NAEP, the framework development policy emphasizes that the Governing Board's charge, when launching a development and update process, "shall explicitly address whether maintaining trends with assessment results from the previous framework should be prioritized above other factors" (NAGB, 2022, p. 5). As noted previously, when content experts are asked to review existing frameworks, they are to be warned of the risks of changing frameworks to the monitoring of trends; and the Board itself must "balance needs for stable reporting of student achievement trends against other Board priorities and requirements" (NAGB, 2022, p. 9).

ARGUMENTS FOR AN “EVOLUTIONARY” APPROACH TO FRAMEWORK REVISIONS

A central recommendation of this paper is that NAGB should develop a more explicit policy to protect trend and, at the same time, ensure the relevance of construct representation by providing for ongoing, incremental revisions to frameworks. The idea of making more frequent but smaller changes to frameworks and specifications was termed an “evolutionary” approach, in conversations among NAEP Validity Studies (NVS) Panel members, and is consistent with recent recommendations from the NASEM consensus study report *A Pragmatic Future for NAEP* (2022):

RECOMMENDATION 3-2: The National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES) should work both independently and collaboratively to implement smaller and more frequent framework updates. This work should include consideration of the possibility of broadening the remit of the standing subject-matter committees that already exist to include responsibility for gradual framework updates, participation in item model development, and working directly with both NAGB and NCES.

At its May 2022 meeting, NAGB’s Assessment Development Committee discussed NASEM’s Recommendation 3-2, recognizing that the reasoning behind “more frequent, gradual changes to NAEP assessment frameworks” is to address two important but competing goals: “Minimize the possibility of breaking trend” and at the same time “increase relevance by reflecting necessary changes in the field more quickly” (Rosenberg, 2022, p. 5).

An evolutionary approach to framework revision is implicitly what NAEP now does when it modifies frameworks without starting a new trend and is the methodology used by international assessments such as PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study). An evolutionary approach protects NAEP’s most precious asset: its ability to document changes in achievement over time, for the nation as a whole, for states, and for identified subgroups. It smooths relatively small changes in construct meaning between adjacent administration years and thus supports interpretation of gains and losses over relatively short time intervals (perhaps a decade). Use of evolutionary trend methodology presumes there is less concern about whether data points 20 or 30 years apart share the same construct meaning. This is analogous to vertical scaling used to measure individual student growth on standardized achievement tests. Vertical scales sometimes cover achievement ranges from kindergarten to grade 12, but it would be indefensible to compare gains in achievement between the two extremes. Instead, vertically equated achievement measures support interpretation of growth within relatively small spans of the scale.

Subject-Matter Committees to Guide Framework Revisions and Updates

Standing subject-matter committees (one for each subject area assessed) were first recommended in 1992 by the NAEP Panel, commissioned by Congress to evaluate implementation of NAEP’s Trial State Assessment. The purpose of standing committees would be to ensure greater continuity, “beginning with the development of the framework

and continuing through the various stages of item specifications, item writing, item scoring, and reporting of results” (NAEd, 1992, p. 30). Standing committees would be in contrast to ad hoc convenings of different content experts to participate at various stages of assessment development and review. Again in 2012, the expert panel, convened by NCES on the Future of NAEP (2012), specifically linked the need for standing committees to the need for oversight to facilitate incremental change: “With standing subject-matter panels, assessment frameworks for each subject-grade combination might be adjusted more frequently, defining a gradually changing mix of knowledge and skills, analogous to the Consumer Price Index” (NCES, 2012, p. 16).

While NCES did create standing committees in response to the 1992 NAEd panel recommendation, the purview of these committees has been limited to the item development stage. A benefit of the current committees is that the content experts who participate can become knowledgeable over time about NAEP’s purposes and structures and the complexities of implementing a given framework. At present, however, the charge to the committees does not ask experts to review updates needed to the framework, nor do they work to coordinate the meaning of assessment inferences from framework to reporting of results. Thus, the NASEM recommendation specifically suggests that consideration be given to broadening the remit or scope of work of the standing subject-matter committees to include responsibility for gradual framework updates and “participation in item-model development,” both of which have implications for how the meaning of constructs is clarified or revised. The NASEM recommendation also suggests that subject-matter committees work more directly with both NAGB and NCES. With this last statement, the NASEM panel raises the importance and purview of subject-matter committees while acknowledging that gradual revisions must not usurp the authority of the Governing Board to determine the content of NAEP assessments.

Processes to Inform and Name Construct Revisions

The point of making incremental changes in frameworks is to keep up with changes in the field and thus avoid having to make more drastic and potentially disruptive changes later. This is analogous to how the U.S. Bureau of Labor Statistics makes changes every 2 years to the components and weightings of goods and services in the “market basket” on which the Consumer Price Index (CPI) is based. The CPI tells us how much the cost of goods and services has increased or decreased—just like monitoring improvements or decreases in student achievement. At the same time, revising the underlying “market basket framework,” based on Consumer Expenditure Survey data, keeps up with changes in what consumers are actually purchasing.

To make this shift to a more evolutionary approach, NAGB would need to redeploy existing resources to inform the ongoing work of standing subject-matter committees and its own deliberations. NAGB’s policy already attends to all of the important sources of information needed to inform the development of new or revised frameworks, such as reviews of both state and international frameworks and assessments, relevant research, and widely accepted professional standards. However, in the past, much of this information was gathered after the decision had been made to launch development of an entirely new framework. With an evolutionary approach, relevant information would need to be available on a more ongoing basis. NAGB does not need costly new procedures, and it certainly does not need to invoke

the full Steering Committee process every 2 or 4 years; but it does need a systematic way to monitor potentially important changes in the field, possibly by regularly surveying state assessment coordinators and/or curriculum directors, in addition to the knowledge of professional standards provided by experts on standing subject-matter committees.

In thinking through process changes needed to support an incremental approach to framework revisions, NAGB will have to decide what responsibilities are assigned to standing subject-matter committees versus the deliberations and decisions that require the attention of the ADC and possibly the full board. For example, every 2 years, subject-matter committees might distill survey information collected from states and write a summary memo to NAGB describing the substantive nature of changes occurring in the field and making recommendations as to whether changes in the NAEP framework might need to be considered. Instead of a fixed period of 5 or 10 years specified for framework review, the ADC should initiate reviews in response to ongoing understandings about whether or not substantial changes are occurring in a field.

Recommendations from a standing committee to the ADC about possible framework revisions should not be made at the whim of subject-matter committee members. Rather, standing-committee experts should be able to identify patterns from multiple sources of evidence documenting how a field is changing. For example, the mathematical practices eventually agreed upon for the 2026 Mathematics Framework are clearly an amalgamation and a more generalized set of reasoning and problem-solving practices, responsive to both Common Core State Standards (CCSS) and Common-Core-influenced state standards. However, Common Core authors did not invent these ideas. This shift in what it means to know and do mathematics has been brought about by learning and disciplinary research over the past two decades, as summarized in *How People Learn II* (NASEM, 2018). Recognizing these shifts, and if authorized, a standing committee for mathematics could have proposed to the ADC revisions to NAEP’s cognitive complexity dimension in the direction of mathematical practices several assessment cycles earlier. Recommendations for change from a standing committee could be proposed first conceptually, based on evidence as to the need for change; and then, if further work were approved by the ADC and NAGB, a detailed draft of proposed revisions could be developed with stakeholder review, etc.

Note that the politics surrounding CCSS a decade ago likely weighed heavily against considering revisions to NAEP Reading and Mathematics Frameworks (concurrent with the new Science Framework) to reflect features of Common Core-like standards adopted and still in place in a large majority of states. While the politics of CCSS were complex and changed over time, importantly, the political backlash was not as much about the learning principles per se and more about the idea of a “Common” curriculum and ties to the allocation of federal recession recovery funds. As was emphasized previously, a national monitoring assessment should not swing wildly in the direction of new learning goals or innovative curricula. Too great a change in the direction of the CCSS would have been unfair to the non-CCSS states and would have violated NAGB’s policy to seek a balance between current practice and advances in the field. However, it is also the case that NAEP cannot be too late to the table; otherwise, it will lack the content necessary to detect progress toward new goals. Unfortunately, the political surround prevented attention to the genuine research-based advances reflected in the CCSS. If ongoing monitoring practices had been in place to evaluate important substantive changes in the field, perhaps it would have been

possible to consider the research basis for development of new frameworks in reading and mathematics separately from the politics.

In addition to continuous monitoring of changes in the field by standing subject-matter committees, NAGB will need to consider how its processes could allow for small, medium-sized, or major changes to existing frameworks and what the relationship should be between its old and new policies. Smaller scale changes—as when new item types are developed to implement more fully an already approved framework—would not require much scrutiny from NAGB. Rosenberg (2022) acknowledged, for example, that not all requirements of a framework can be fully operationalized at the outset and lessons are often learned from the first administration that prompt revisions for subsequent administrations. By contrast, if a conceptual summary of changes in a field implied real changes in how the subject-matter construct should be defined, then the initial recommendation from the standing committee would signal the need for additional work, and NAGB would need to specify (by policy or ad hoc deliberations) what additional information was needed and who should do the work. For medium-sized revisions—as illustrated by the 2009 Reading Framework discussed below—NAGB might augment the membership of the standing subject-matter committee. For more substantial changes—for example, possibly adding the explicit assessment of mathematical practices described above—initial recommendations from a standing committee could be deemed by NAGB to be “major changes” and thus could provoke the launch of NAGB’s full Steering Committee framework development process. In this way, NAGB could adopt an incremental or evolutionary approach to framework revisions but could reserve the right to invoke its full-scale framework development policy whenever the potential changes were substantial enough that additional information gathering, participation of a new Steering Committee, and broader review from stakeholders were needed.

In addition to considering appropriate review processes and sources of evidence, it is important to emphasize that substantive changes in how a construct is defined and operationalized (by item types) must be clearly and publicly documented. Good examples of what this might look like can be found in the documentation of past framework changes. As an illustration of a “medium-size” shift in construct definition, exhibit 3 is taken from the 2019 Reading Framework summarizing the differences between the 1992–2007 NAEP Reading Framework and the 2009–2019 NAEP Reading Framework. I call this a medium-size change because substantive changes were clearly identified, but content expert studies of both frameworks and items did not warrant starting a new trend line. Changes summarized in exhibit 3 were consistent with evolving changes in research on reading comprehension. For example, the cognitive targets in the 2009 framework were specified in much greater detail, as to both level and text type, than had been the case for the more general understanding, interpretation, and drawing connections between reader and text in the 1992–2007 framework. A few examples from a page-long matrix include the following cognitive targets for Literary Text items: Locate/Recall items might ask the reader to identify character traits or a sequence of events; Integrate/Interpret items require the reader to compare or connect ideas, problems, or situations; Critique/Evaluate items could ask the reader to “evaluate the role of literary devices in conveying meaning” (NAGB, 2019, p. 43).

Exhibit 3. Similarities and differences: 1992–2007 and 2009–2019 reading frameworks

1992–2007 NAEP Reading Framework			2009–2019 NAEP Reading Framework		
Content	Content of assessment: <ul style="list-style-type: none">LiteraryInformationalDocument	Contexts for reading: <ul style="list-style-type: none">For literary experienceFor informationTo perform task	<ul style="list-style-type: none">Literary textFictionLiterary nonfictionPoetry	<ul style="list-style-type: none">Informational textExpositionArgumentation and persuasive textProcedural text and documents	
Cognitive Processes	Stances/aspects of reading: <ul style="list-style-type: none">Forming general understanding.Developing interpretation.Making reader/text connections.Examining content and structure.		Cognitive targets distinguished by text type		
			Locate/ recall	Integrate/ interpret	Critique/ evaluate
Vocabulary	Vocabulary as a <i>target</i> of item development, with no information reported on students’ use of vocabulary knowledge in comprehending what they read.		Systematic approach to vocabulary assessment with potential for a vocabulary subscore.		
Poetry	Poetry included as stimulus material at grades 8 and 12.		Poetry included as stimulus material at all grades.		
Passage Source	Use of intact, authentic stimulus material.		Use of authentic stimulus material plus some flexibility in excerpting stimulus material.		
Passage Length	Grade 4: 250–800 words Grade 8: 400–1,000 words Grade 12: 500–1,500 words		Grade 4: 200–800 words Grade 8: 400–1,000 words Grade 12: 500–1,500 words		
Passage Selection	Expert judgment as criterion for passage selection.		Expert judgment and use of at least two research-based readability formulas for passage selection.		
Item Type	Selected-response and constructed-response items included at all grades.		Selected-response and constructed-response items included at all grades.		

SOURCE: National Assessment Governing Board, 2019, exhibit 2, p. 15.

Recommendations for Implementing an Evolutionary Approach to Framework Revisions

- NAGB should develop a more explicit policy to protect trend and, at the same time, ensure the relevance of construct representation by providing for ongoing, incremental revisions to frameworks.
- Standing subject-matter committees should have greater responsibility to ensure continuity and integration across stages of the assessment development process and to make recommendations for gradual framework revisions.
- To support a more evolutionary approach, some sources of evidence documenting changes in the field—such as relevant research, state and local standards and assessments, or widely accepted professional standards in the disciplines—may need to be collected on an ongoing basis and distilled by standing subject-matter committees to anticipate needed revisions.
- An evolutionary approach presumes that the utility of NAEP depends primarily on the relevance and comparability of frameworks over shorter time intervals; therefore, NCES and NAGB will need to caution policy researchers that subject-matter constructs have not necessarily been held constant over longer time periods.
- All revisions would require approval by the Governing Board, but NAGB will need to specify how a new evolutionary policy would provide for external review and/or articulate with its existing policy for the development of new frameworks. For example, medium-size revisions might require external review by stakeholders before submission for Board consideration; or some revised construct definitions recommended conceptually by a standing committee might be deemed major enough that NAGB would decide to invoke its existing full-scale process for development of new frameworks.

What Are the Cautions or Downsides to an Evolutionary Approach?

What happens if there really are meaningful changes occurring in the field that are blurred by taking an evolutionary approach that doesn’t change enough? In the next section, studies are considered that illustrate how starting a new trend with new frameworks makes it possible to see a more dramatic picture of educational progress and, likewise, how decisions not to revise the construct can obscure progress. Then, the subsequent section summarizes the use of alignment and empirical bridge studies to evaluate whether changes in the meaning of a construct have been sufficient to warrant beginning a new trend. Lastly, the need for special studies is considered as a safeguard for those occasions when protecting trend and maintaining continuity with the past could make NAEP less valid for evaluating the effects of curricular or instructional reforms.

HOW DECISIONS ABOUT FRAMEWORKS AND TREND CAN OBSCURE OR ILLUMINATE PROGRESS

The content of a test—what topics it “covers” and what thinking processes it requires—clearly affects test score results. While the emphasis in the preceding sections has been on the importance of keeping assessments the same or making small, incremental revisions to enable comparisons with the past, there is a competing consideration whereby failure to make major framework changes could prevent NAEP from documenting progress appropriately. Being able to “see” important changes might also depend on whether or not a new framework is accompanied by the beginning of a new trend.

Example: Long-Term Trend NAEP versus Main NAEP

The new NAEP or Main NAEP, begun in 1990 (or 1992, depending on subject area), was based on new frameworks and consensus processes along with important changes in administration procedures, the participation of states on a voluntary basis, and so forth. The framework development process did not involve explicit consideration as to how the 1990s frameworks were expected to depart from prior content specifications. It was generally understood, however, in the context of Goals 2000 and standards-based reforms, that new frameworks would likely be more challenging and would address higher-order thinking skills. Subsequently, studies have been done that enable us to examine more specifically how the new assessments differed substantively from prior assessments and to see how, in turn, content differences affect results.

In 2006, the National Center for Education Statistics (NCES) commissioned the Human Resources Research Organization (HumRRO) to conduct a content analysis or “alignment” study comparing LTT and Main NAEP item pools in both reading and mathematics at grades 4 and 8. HumRRO researchers Dickinson et al. (2006) used the Webb (1997) alignment methodology to examine both the breadth and depth of knowledge covered as well as categorical concurrence with the major content strands in the respective 2003 Main NAEP frameworks. (A cross-framework comparison was not conducted because original documents were not available for LTT assessments.) The methodology involved item reviews by panels of content experts trained to reliably apply study criteria.

Study findings regarding content alignment are summarized in exhibit 4, reproduced from the HumRRO report. The 100 percent Full Categorical Concurrence, for both assessments in both grade levels and subject areas, means that every standard was represented by at least six test items. This is a minimal requirement; a more exacting evaluation was provided by the Range and Balance criteria, which examine coverage at the level of objectives within each strand. For grade 4, Main NAEP had high or full coverage (80–100 percent), where the percentage refers to the percentage of objectives covered by at least one test item. At grade 8, in both subject areas, the range of knowledge covered by Main NAEP was partial; i.e., between 60 percent and 67 percent of objectives were assessed by one or more test items. LTT is very weak on this criterion, 0–20 percent coverage, which reflects the differences between the specific content objects measured by the two assessments (Dickinson et al., 2006). Except for math at grade 4, both Main NAEP and LTT do less well on the more stringent balance criterion, which requires uniform distribution of items across objectives within content strands.

Exhibit 4. Summary of alignment results for Main NAEP 2003 and LTT

Content Area	Categorical Concurrency		Range of Knowledge		Balance of Knowledge	
	Main NAEP	LTT	Main NAEP	LTT	Main NAEP	LTT
Reading – 4th	FULL (100%)	FULL (100%)	FULL (100%)	WEAK (0%)	WEAK (0%)	PARTIAL (67%)
Reading – 8th	FULL (100%)	FULL (100%)	PARTIAL (67%)	WEAK (0%)	WEAK (30%)	PARTIAL (67%)
Math – 4th	FULL (100%)	FULL (100%)	HIGH (80%)	WEAK (20%)	FULL (100%)	FULL (100%)
Math – 8th	FULL (100%)	FULL (100%)	PARTIAL (60%)	WEAK (0%)	WEAK (0%)	WEAK (20%)

NOTE: LTT is Long-Term Trend.

SOURCE: Dickinson et al., 2006, table 19.

As part of the alignment study, expert reviewers were also asked to rate the Depth of Knowledge (DOK) or cognitive processing demands of each test item in the two assessments. The four-point rubrics differed slightly for the two subject areas. In general, items at DOK level 1 measure simple recall; level 2 items require some mental processing, such as comprehension (Reading) or interpreting graphs (Math); level 3 items involve strategic thinking, requiring students to synthesize ideas from text or use reasoning and evidence; and level 4 items involve more complex and extended thinking, planning, and abstract reasoning. The results of the DOK comparisons for the two assessments are presented in exhibits 5–8 from the HumRRO report.

In all four comparisons—for both reading and mathematics, at both grades 4 and 8—Main NAEP items tend to require higher levels of cognitive processing than LTT. The shift in the distribution of items across the four levels of the rubric is best captured by the mode or most frequent category, which is level 2 for Main NAEP and level 1 for LTT. For reading at both grades 4 and 8, the vast majority of Main NAEP items are rated 2 or 3, with even 9–10 percent at level 4. This is in contrast to LTT, for which the vast majority of reading items are rated 1 or 2, with only 2 percent of items at level 4. For mathematics, the difference is not quite so stark but is still evident. At grade 4, 56 percent of LTT items are simple recall items as compared to 45 percent for Main NAEP. At grade 8, 70 percent of LTT math items are level 1 versus 42 percent for Main NAEP.

Exhibit 5. Reading 4th grade: Depth of knowledge comparison

	2003 NAEP		LTT	
Depth of Knowledge Level	Percent of Items Rated This Level		Percent of Items Rated This Level	
	Mean	SD	Mean	SD
1	24.23	6.42	49.90	26.11
2	40.44	7.44	40.06	18.65
3	27.89	7.95	9.24	8.45
4	10.07	4.91	1.83	0.64
	Mode: 2	Range: 2.86	Mode: 1	Range: 2.42

NOTE: LTT is Long-Term Trend. SD is standard deviation.

SOURCE: Dickinson et al., 2006, table 14.

Exhibit 6. Reading 8th grade: Depth of knowledge comparison

	2003 NAEP		LTT	
Depth of Knowledge Level	Mean % of Items Rated This Level		Mean % of Items Rated This Level	
	Mean	SD	Mean	SD
1	23.63	17.50	58.36	21.25
2	43.53	7.55	32.87	18.41
3	39.45	8.59	7.43	2.91
4	8.63	5.13	2.35	1.24
	Mode: 2	Range: 2.86	Mode: 1	Range: 2.57

NOTE: LTT is Long-Term Trend. SD is standard deviation.

SOURCE: Dickinson et al., 2006, table 15.

Exhibit 7. Math 4th grade: Depth of knowledge comparison

	2003 NAEP		LTT	
Depth of Knowledge Level	Mean % of Items Rated This Level		Mean % of Items Rated This Level	
	Mean	SD	Mean	SD
1	45.49	18.78	55.58	29.85
2	46.43	17.56	39.21	24.27
3	13.41	19.06	8.36	9.34
4	1.4	1.13	X	X
	Mode: 2	Range: 2.13	Mode: 1	Range: 1.63

NOTE: LTT is Long-Term Trend. SD is standard deviation.

SOURCE: Dickinson et al., 2006, table 16.

Exhibit 8. Math 8th grade: Depth of knowledge comparison

	2003 NAEP		LTT	
Depth of Knowledge Level	Mean % of Items Rated This Level		Mean % of Items Rated This Level	
	Mean	SD	Mean	SD
1	42.01	17.17	69.74	17.84
2	47.97	15.26	29.59	17.73
3	9.51	6.92	1.5	1.15
4	0.95	0.37	X	X
	Mode: 1	Range: 2.44	Mode: 1	Range: 1.44

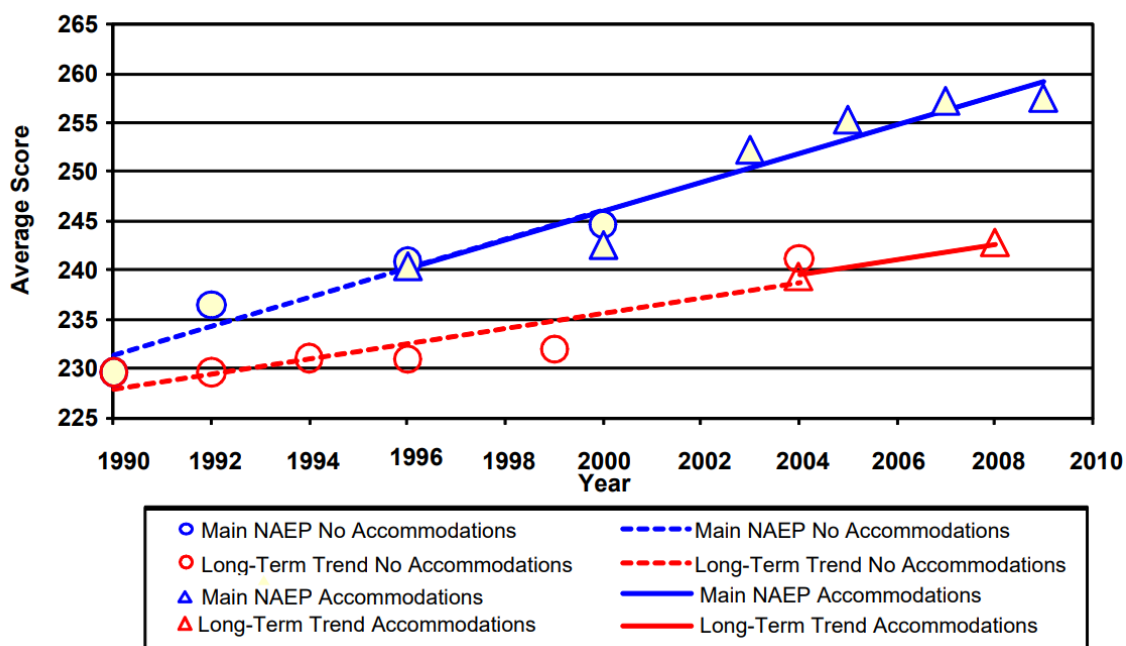
NOTE: LTT is Long-Term Trend. SD is standard deviation. The source reports "Mode: 1" for 2003, however, based on the reported means, it should say "Mode: 2."

SOURCE: Dickinson et al., 2006, table 17.

An important question, then, is whether differences in the test construct, especially differences in levels of cognitive complexity, result in differences in test score outcomes. Do different assessments of student achievement result in different pictures of "educational progress" over time? NCES cautions against trying to make direct comparisons between LTT and Main NAEP because the two assessments are not on the same scale and there are not straightforward ways to convert age cohorts from LTT (9-, 13-, and 17-year-olds) into grade cohorts for Main NAEP. For this reason, Beaton and Chromy (2010) undertook a study on behalf of the NAEP Validity Studies (NVS) Panel to examine the relationship between the two trends based on changes in the definition of the test constructs.

Beaton and Chromy (2010) investigated numerous differences between the two assessment programs, such as the introduction of accommodations and greater inclusion and private school coverage. They also provided in-depth analyses of differential trends by racial and ethnic groups, of trend components attributable to population shifts, and of interaction effects of age and grade distributions associated with the respective population differences. For our purposes here, the primary comparisons of the trend data from the two assessments are sufficient. These results are summarized in four graphs, exhibits 9–12, showing the two trends over nearly a 20-year period for reading and mathematics at both grades 4 and 8.

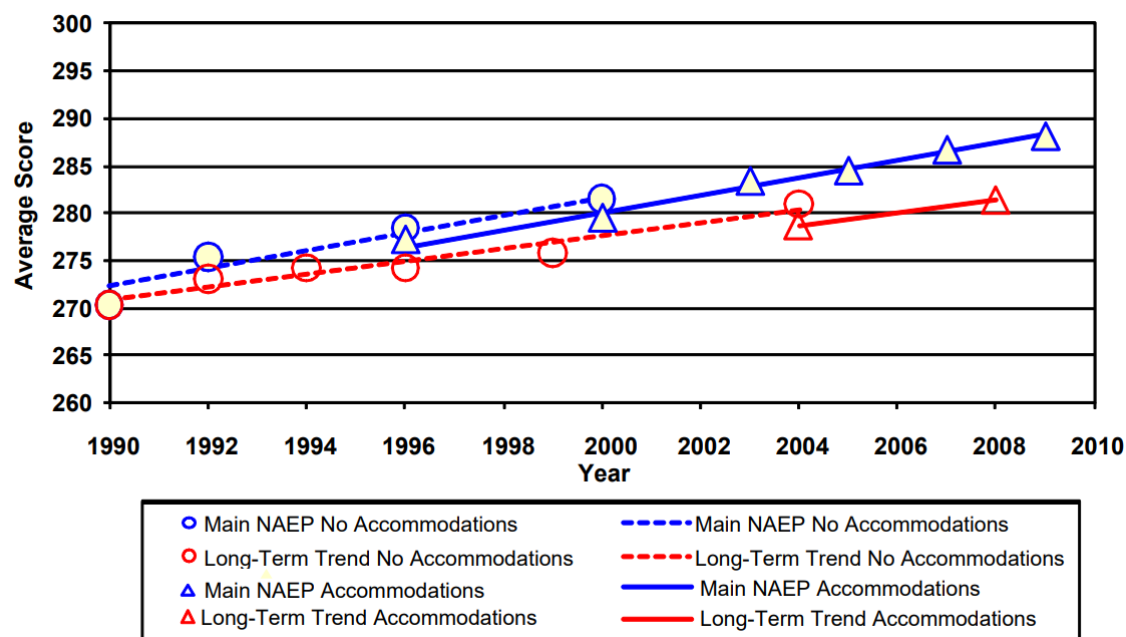
Exhibit 9. Average mathematics scores by assessment year: Main NAEP grade 4 (transformed) and LTT age 9



NOTE: The 1990 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

SOURCE: Beaton & Chromy, 2010, figure 2.1.

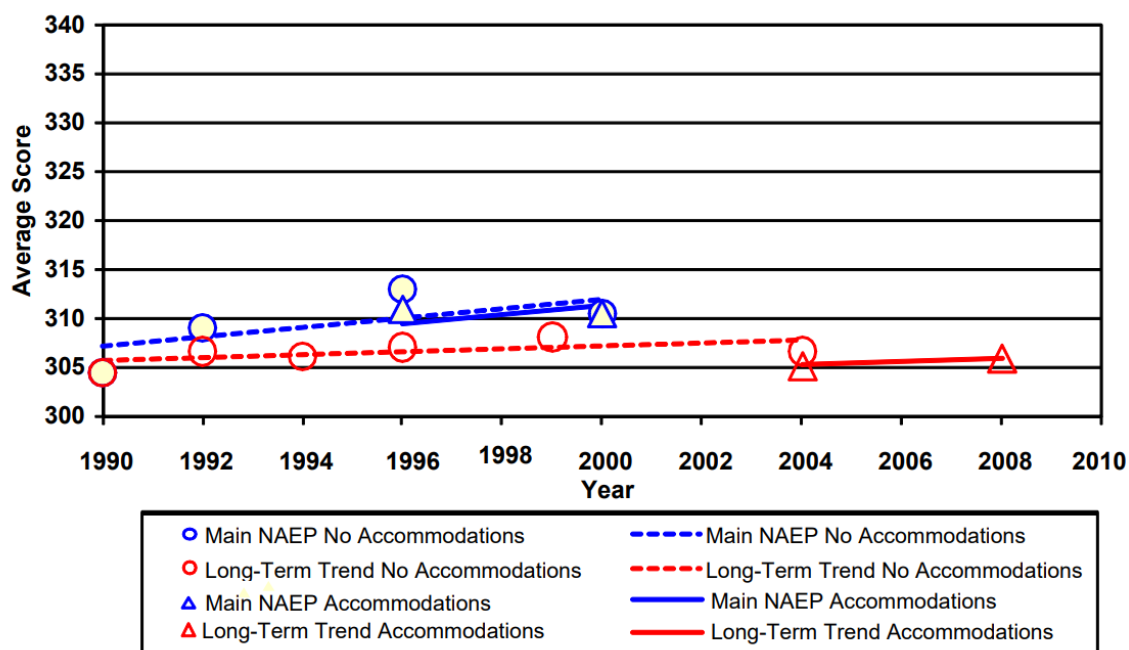
Exhibit 10. Average mathematics scores by assessment year: Main NAEP grade 8 (transformed) and Long-Term Trend age 13



NOTE: The 1990 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

SOURCE: Beaton & Chromy, 2010, figure 2.2.

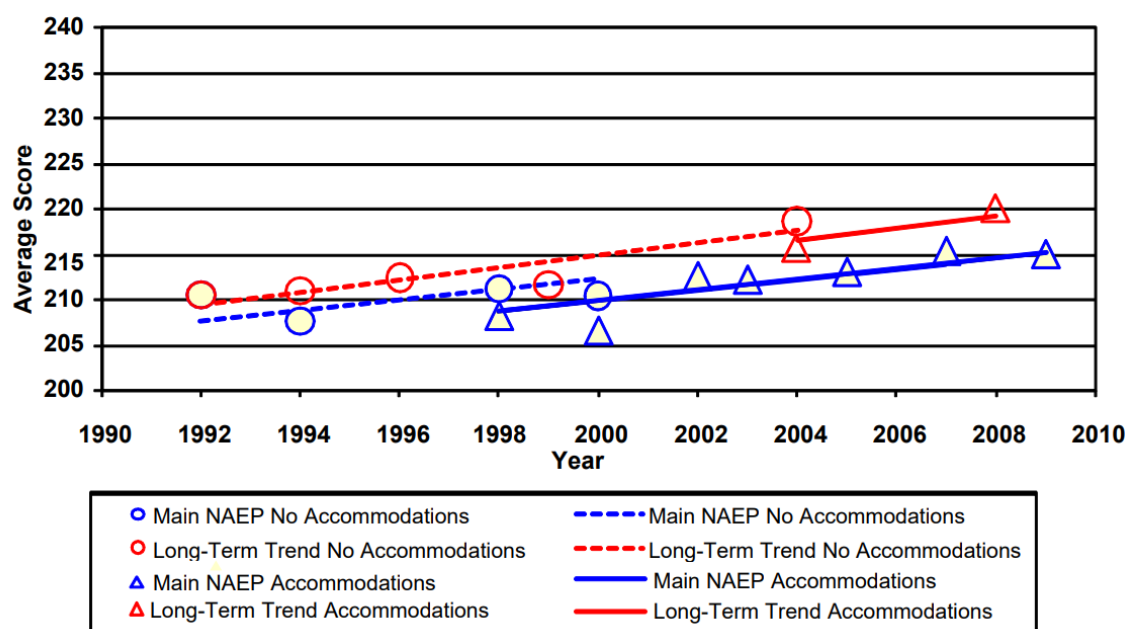
Exhibit 11. Average reading scores by assessment year: Main NAEP grade 4 (transformed) and Long-Term Trend age 9



NOTE: The 1992 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

SOURCE: Beaton & Chromy, 2010, figure 2.4.

Exhibit 12. Average reading scores by assessment year: Main NAEP grade 8 (transformed) and Long-Term Trend age 13



NOTE: The 1992 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

SOURCE: Beaton & Chromy, 2010, figure 2.5.

The differences between the trends in mathematics are statistically and practically significant and are especially pronounced for grade 4. Main NAEP shows a greater gain than LTT. The steeper gain for Main NAEP compared to LTT Mathematics at grade 8 is statistically significant, although not as dramatic as the steeper gain for Main NAEP at grade 4. This picture of differential gains is what might have been expected given the greater cognitive challenge of Main NAEP compared to LTT. There was simply more room to grow on a test with a higher ceiling. Diverging trends in mathematics also illustrate how changing constructs as part of framework development can change assessment results. In the policy world and in the world of mathematics education reform, gains on Main NAEP are consistent with what might have been hoped for with the introduction of the 1989 Curriculum Standards by the National Council of Teachers of Mathematics (which the Main NAEP frameworks mirrored) and with subsequent efforts by states and districts to improve mathematics instruction. Although making such causal inferences from NAEP data certainly would not be possible without much more extensive evidence and analyses, the point here is that *these gains could not have been detected if only LTT data were being collected or if the two assessments had been blended in one revised framework and continuous trend.*

In reading, the trend comparisons tell a different story. There were no significant differences between the LTT and Main NAEP Reading trends for either grade 4 or 8. The no-difference finding occurred despite the content evidence from the HumRRO study (Dickinson et al., 2006), which showed greater cognitive complexity for Main NAEP Reading—on the same order as was found for Main NAEP Mathematics. Both reading assessments showed significant improvement over time for grade 4 and no change over time for grade 8. There has been considerable speculation among researchers and policy analysts as to why it has been so difficult to improve reading achievement as measured by NAEP. One possibility is that reading achievement is influenced by both in-school and out-of-school factors, whereas mathematics achievement is more directly tied to schooling and therefore more sensitive to curricular changes.

Example: Reweighting of TUDA Mathematics Results to Align with State Assessment Content

A second and very recent example helps to illustrate how framework dimensions shape assessment results and can thereby enhance or dampen the appearance of educational progress. The Common Core State Standards (CCSS) developed under the auspices of the National Governors in 2010 became politically suspect despite having broad support initially from governors and chief state school officers. Staying aloof from the political fray, NAGB did not undertake a frameworks revision process to consider whether NAEP should be made more consistent with CCSS. However, key features of the CCSS have a strong research base, especially the idea of (1) covering fewer topics in more depth, with less repetition and more continuity over time, and (2) developing higher-order thinking processes more explicitly through disciplinary practices. This research base may be one reason that so many states have adopted new standards over the past decade that resemble the “college and career ready” standards of the CCSS even when they are not explicitly identified as Common Core standards.

In the two most recent main NAEP Mathematics administrations, fewer Trial Urban District Assessment (TUDA) districts experienced significant improvement and more experienced no change, or even significant decreases, compared to the entire first decade of TUDA

NAEP Mathematics results. Between 2003 and 2017, of 112 comparisons between adjacent years across participating TUDAs, there were 12 significant decreases at grade 4; 11 of these were observed in 2015 or 2017. Similarly, of the five significant decreases during the same period between adjacent years at grade 8, four were observed in 2015 or 2017. Some superintendents in TUDA districts raised concerns as to whether these relatively negative trends could be due to their states' adoption of new college- and career-ready standards in mathematics, followed by corresponding shifts in their state assessments and instructional emphases, which were not reflected in the NAEP mathematics assessments.

Enis Dogan (2019) undertook a study for NCES to evaluate whether 2017 Mathematics grades 4 and 8 TUDA mean scores would change if the NAEP subscales were reweighted to correspond to the content distributions of the respective state assessments. The study also examined how reweighting would affect TUDA trend data from 2013 through 2019. Based on a separate study by the NVS Panel (Daro et al., in press) proportional allocation of items by content strand was available for three state assessments, including both national assessment consortia (PARCC and Smarter Balanced). Data in exhibits 13 and 14 show the relative weighting of items by content strand for NAEP and three state assessments at grades 4 and 8.

As shown in exhibit 13, the weight of the Numbers subscale at grade 4 was much greater on state assessments compared to its 40 percent weight on NAEP; weights for Numbers on the respective state assessments were 54 percent, 73 percent, and 71 percent. There was also an across-the-board decrease in the weights assigned to Data, from 10 percent on NAEP to 0 or 1 percent on state assessments. Additionally, two state assessments gave substantially less weight to Measurement and Geometry compared to NAEP. These shifts reflect the research-based intention of college- and career-ready standards to teach fewer targeted subjects in greater depth, which for grade 4 are Numbers and, secondarily, Algebra.

Nine TUDA districts fall within the states whose assessments were analyzed by Daro et al. (in press). When Dogan (2019) recomputed 2017 grade 4 NAEP Mathematics results using weights consistent with each district's state assessment, all showed improvements, from 1.1 to 4.6 points on the NAEP scale. To test whether these improvements produced by reweighting were consistent across years, the reweighting calculations were completed for each of the TUDA administration years starting in 2013. The median difference across these districts was 0.49 in 2013, 2.18 in 2015, 2.08 in 2017, and 2.3 in 2019. Starting in 2015, all districts showed improvement with reweighting to better match their respective state standards.

Exhibit 13. Subscale weights relative to state assessments and according to the NAEP framework: Grade 4 mathematics

	Numbers	Measurement	Geometry	Data	Algebra
<i>Weight in NAEP framework</i>	40%	20%	15%	10%	15%
Weight relative to SA2	54%	18%	15%	0%	14%
Weight relative to SA3	73%	9%	2%	1%	14%
Weight relative to SA4	71%	8%	3%	0%	18%

NOTE: States included in the study were not named. SA2 = State Assessment 2, SA3 = State Assessment 3, and SA4 = State Assessment 4.

SOURCE: Dogan, 2019, table 2.

Relative weights for NAEP and state assessments for grade 8 mathematics are presented in exhibit 14. Here, the increased importance of algebra in college- and career-ready standards can be seen in the state assessments' proportional allocations of 39 percent, 42 percent, and 45 percent compared to NAEP's 30 percent. Conversely, Data was less heavily weighted on the state assessments, 2 percent or 7 percent, compared to NAEP's 15 percent. Note that the differences between NAEP and these state assessments was not so great at grade 8 as the differences at grade 4. Not surprisingly, reweighted 2017 NAEP results still showed improvements for all nine TUDA districts, but the improvements were much smaller, ranging from 0.9 to 2.6 NAEP scale score points. The median difference across these districts was 0.41 in 2013, 0.76 in 2015, 1.02 in 2017, and 1.28 in 2019. Again, all nine districts showed improvement with reweighting from 2015 onward.

Exhibit 14. Subscale weights relative to state assessments and according to the NAEP framework: Grade 8 mathematics

	Numbers	Measurement	Geometry	Data	Algebra
Weight in NAEP framework	20%	15%	20%	15%	30%
Weight relative to SA2	16%	19%	19%	7%	39%
Weight relative to SA3	14%	18%	19%	7%	42%
Weight relative to SA4	21%	16%	16%	2%	45%

NOTE: States included in the study were not named. SA2 = State Assessment 2, SA3 = State Assessment 3, and SA4 = State Assessment 4.

SOURCE: Dogan, 2019, table 4.

Findings from the Main NAEP versus LTT NAEP differences in mathematics and from the TUDA district reweighting results illustrate the point that progress toward valued learning goals may be obscured if those learning dimensions are insufficiently represented in the assessment. The change from LTT to Main NAEP made it possible to document progress in mathematics on an assessment that required greater depth of knowledge in response to more cognitively complex items. However, a similar relative gain was not observed in reading. In the TUDA comparisons, the Common Core-influenced decision in states to emphasize deeper mastery of numeracy skills at grade 4 and of algebra at grade 8 necessarily meant less instructional attention to other NAEP topics, especially Data. Reweighting results to reflect state standards showed achievement gains that otherwise would have been missed.

The differences observed in these comparisons are large enough to have policy consequences. Leaders in TUDA districts might, for example, fear that their investments in curricular reforms geared to state standards were ineffective, when in fact reweighted results are moving in the desired direction. Looking at the differences between Main NAEP and LTT NAEP in mathematics, some might conclude that it was unnecessary to start a new trend because improvement over time would still be visible, just with a rising trend line that was not quite so steep. However, this reasoning fails to recognize that *the relationship between the old and new assessments would not necessarily be the same across jurisdictions*. For policy purposes, it is important to be able to see which states or urban districts are showing the greatest progress toward new and more ambitious goals. A blended assessment, without a separate Main NAEP trend line, would make it harder to see these important differences.

ALIGNMENT AND BRIDGE STUDIES TO EVALUATE CONSTRUCT SHIFT

In the past, the substantive processes involved in developing a new framework have themselves been informative as to how the underlying construct measured by the assessment might be changing. Discussions at each stage of the process involve naming the changes being contemplated. For example, the 2005 Mathematics Framework for grades 4 and 8 changed the mathematical ability and power dimension to the dimension of mathematical complexity but was not expected to alter the nature of assessment items. When changes could be more substantial so as to preclude maintaining trend, NAGB may conduct an alignment study, convening content experts to review how well old and new assessment items align with their respective frameworks and how they compare with each other. If they are judged to be similar—as was true, for example, for the 2005 and 2009 Mathematics Frameworks at grades 4 and 8—then the next step is to proceed with empirical bridge studies.

Bridge studies, also called “trend studies,” involve administering the old and new assessments to three randomly equivalent samples of students. One group takes the old assessment; another takes the new assessment. Students in the third group are administered a mixed or “braided” version of the assessment. By administering both assessments to the same sample of students, it is possible to put the old and new items on the same scale and to determine the empirical relationship between the two. To evaluate whether the old and new assessments are similar enough to support maintaining trend, several comparisons are made:

- Are blocks of old and new questions comparable in terms of difficulty, nonresponse rates, and reliability?
- Do the old and new assessments produce similar results (average scores and achievement level percentages) for major reporting groups?
- Are IRT item parameters similar between the scaled together and scaled separately scenarios?
- Are correlations between blocks and subscales similar in both the “scaled together” and “scaled separately” conditions?

In virtually all cases in which bridge studies have been performed, the decision has been that old framework and new framework assessments are sufficiently similar to support maintaining trend. *Thus, it is important to note that all of the major decisions to start a new NAEP trend have been made on substantive grounds when there has been a clear understanding of important changes being made in the framework that necessitate the start of a new trend.*

Two important concerns can be raised about the adequacy of empirical methods for making decisions about trend:

1. Bridge study methods may not be sensitive enough to detect important differences, especially given that old and new construct definitions are expected to be strongly correlated. Moreover, old and new assessment blocks are built with substantial proportions of items from the prior assessment. Bridge studies do not test construct shift implied by only the new item types that were added.
2. It is also important to note that construct differences, not yet reflected in current instructional practice, are not likely to be detectable at the outset using empirical

methods. Altered constructs are not likely to be manifest until they are actually being learned by sufficient numbers of students.

The effects of instruction (or other learning opportunities) on construct shift can be seen in the Beaton and Chromy (2010) study, for example, where Main NAEP and LTT mathematics *were set equal initially* and diverged increasingly over time. Similarly, the reweighted trends in the Dogan (2019) study showed a pattern of steadily increasing improvement and divergence from reported NAEP over time. When new items are introduced, reflecting content and reasoning skills not yet taught, it is reasonable to expect that those items would be relatively more difficult, accessible only to relatively few students who had had instruction or who had experience with the content outside of school. Substantial instructional shifts, which do not happen quickly, would likely affect both means and correlations among test dimensions. Items representing new goals would most probably load on the first, general factor in a factor analysis and only later be discernible as an identifiable, but still correlated, separate factor. An understanding of the limitations of empirical methods for detecting construct shift gives support to NAGB's practice of relying primarily on substantive evidence to decide when a new framework is different enough to break trend.

A third issue regarding bridge studies has to do with the proportional weights assigned to construct strand or subtests, as was examined in the Dogan (2019) study. Much like ecological correlations, which show stronger relationships between aggregates than for individual items, reweighted subtests (or large disproportions in the allocation of items by substrand) will have much bigger effects on total scores than making more subtle changes in the mix of items. To our knowledge, bridge studies have always been done with the same *proportional allocation* of items to both old and new versions of the assessment and thus would not address the issue of construct shift in cases like the CCSS, where the decision was made to focus on particular content strands in much greater depth.

Recommendations Regarding Bridge Studies to Evaluate Construct Shift

- NAGB should consider making an explicit policy, consistent with its long-standing practice, that the decision to start a new trend will be based primarily on substantive evidence of intentional changes in the definition of the assessment construct.
- NAGB or NCES should conduct studies to test whether empirical checks like those currently used would be sensitive enough to detect the kinds of trend divergences observed in the Beaton and Chromy (2010) study. If empirical checks lack this kind of sensitivity, then the community of NAEP researchers should resist relying on the phrase “it’s an empirical question.”
- The proportion of items allocated by substrand within a construct like reading or mathematics is an important substantive decision and should continue to be addressed as part of Steering Panel and ADC deliberations.

ARGUMENTS FOR AND AGAINST “BREAKING TREND”

Having approved new Mathematics and Reading Frameworks for 2026, NAGB faces a critical decision about whether or not to start new trends for its two most widely administered and frequent assessments. Several “now or never” arguments can be made for starting new trends in 2026:

- The existing frameworks and trends will be 36 and 34 years old, respectively. If trends are not restarted for these new frameworks, they would be unlikely to be disrupted for at least another decade.
- The changes in cognitive research and research on disciplinary learning are as great in the past 30-plus years as in the decades preceding the beginning of Main NAEP in the 1990s; and these changes, as summarized in *How People Learn II* (NASEM, 2018), are represented sufficiently in the new frameworks.
- The addition of disciplinary practices in mathematics and across disciplines means that there are potentially much more explicit ways to test for higher-order thinking skills, in the new Mathematics Framework, than when these learning goals relied on vague descriptors about cognitive complexity. Similarly, the change in reading contexts in the 2026 Reading Framework—from general literary and informational texts to discipline-specific literature, social studies, and science contexts—is a change in the definition of reading that could (and should) be reflected in systematic differences in item types between the old and new assessments.

The opposing argument in favor of maintaining trends, consistent with an evolutionary approach to framework revision, does not dispute these claims but focuses instead on the utility and sufficiency of comparability across short time intervals. The construct of reading or mathematics may well have shifted compared to 30 years ago, but what matters more is the smoothing and overlap in construct meaning from 2024 to 2026 and then to 2028. Although the research base has changed significantly over the past two decades, changes in instructional practices happen much more gradually, consistent with an evolutionary approach. With the new NAEP Science Framework in 2009 and the beginning of a new trend, NAGB did well to anticipate and reflect the research-based changes evident in NRC’s *A Framework for K-12 Science Education* (2012) and subsequent Next-Generation Science Standards (2013). However, in reading and mathematics, the inflection point associated with the CCSS and college- and career-ready standards-based reforms beginning in 2010 has already been missed. As a result, it is not so clear how starting a new trend in 2026 with an only slightly changing item pool would provide a more accurate picture of educational progress. One lesson to be learned from the Dogan (2019) study is that changing the mix of items to be consistent with changes in frameworks may not affect the reporting of assessment results. It was only when *aggregations of items in subtest scores were reweighted* to reflect framework differences that important differences were seen in NAEP Mathematics results. When a decision has been made *not* to change NAEP frameworks despite changes in state standards, disciplinary research, and content area professional standards, then NAGB and NCES should undertake special studies, as discussed in the next section, to evaluate the consequences of differences in construct meanings for various uses of NAEP data.

NCES SPECIAL STUDIES

The Dogan (2019) study provides striking, if not definitive, evidence that the decision *not* to create a new mathematics framework in 2011 or 2013 might have been a scientific mistake, threatening the validity of NAEP for evaluating changes in student achievement associated with major curricular reforms, even though it was understandable why NAGB chose to stay out of the politics of the Common Core per se. A larger replication study involving states as well as more TUDA districts is needed to determine whether NAEP Mathematics can be used as a policy research tool in the Common Core era.

In the 1990s and 2000s, NAEP was the “gold standard” for evaluating the effects of state accountability policies pre-NCLB (No Child Left Behind) as well as the effects of NCLB. NAEP Reading and Mathematics scores have been the outcome measure in many hundreds of studies by economists and education policy researchers. In the research literature on test-based accountability, the problems of test-score inflation or nongeneralizable achievement gains caused by a narrow focus on teaching to state accountability tests are well known (Figlio & Loeb, 2011). When evidence of achievement gains or closing of gaps differed on state tests versus NAEP, NAEP was considered the more trustworthy indicator of actual achievement trends (Klein et al., 2000). The importance of NAEP as a policy research tool would no longer hold true, however, in the most recent decade, if there is evidence that the content of NAEP Mathematics is importantly different from reform learning goals.

Today, the most pressing standards-based reform question is no longer the effects of high-stakes accountability but, rather, the effectiveness of the CCSS. NAEP results are being used to call the Common Core standards themselves—or Common Core implementation—a failure (Polikoff et al., 2020). The technically sophisticated analysis by Song et al. (2019) includes an acknowledgment that “our measures of student achievement—NAEP scores—are not perfectly aligned with the CCR standards” (p. 25). However, researchers do not have a way to estimate the magnitude of effects from 20 percent tested-but-not-taught outcome measures. Policy interpretations based on the most recent NAEP assessments conclude without caveat, “With the release of the 2019 National Assessment of Educational Progress results in math and reading, it became clear that standards-based reform has not moved the needle on student achievement” (C-SAIL, 2020, p. 2). It is important to note that the median improvements—2.3 points at grade 4 and 1.28 points at grade 8—found from reweighting 2019 math scores in the Dogan (2019) study are the same order of magnitude as the “losses,” compared to predicted, attributed to CCR standards by Song et al. (2019): -1.49 at grade 4 and -2.47 at grade 8.²

NCES is a federal statistical agency. According to the principles outlined by NASEM (2017), to maintain the credibility and trustworthiness of the data it provides, NCES “must avoid even the appearance that its collection, analysis, or dissemination processes might be manipulated for political or partisan purposes” (p. 3). Therefore, the recommendation here is consistent with the NAEP Validity Studies (NVS) Panel response to the Dogan study (Hughes et al., 2019), which recommended that NCES *not consider score adjustments or any kind*

² Note that this comment is only about the similar magnitude of effects. The Dogan study was conducted with large urban districts, whereas Song et al. calculated effects for “treatment” states, defined as those states that made the biggest changes in their standards in response to the CCSS.

of state or district “customizing” as part of official reporting of NAEP results. However, the principles guiding statistical agency decisions also stress the importance of maintaining credibility among data users, especially noting that “few data users are in a position to verify the completeness and accuracy of statistical information” (p. 2).

It would be well within NCES’s responsibilities to conduct a one-time study extending the Dogan (2019) analysis. Rather than producing a competing version of NAEP results, such a study could be part of the NCES Research and Development series, which in addition to studies on the “cutting edge” of methodological developments, include studies that contribute to “discussions of emerging issues of interest to educational researchers, statisticians, and the federal statistical community in general” (Bandeira de Mello et al. 2009, p. iii). A reweighting study to inform policy researchers would be similar in purpose to NCES Research and Development reports comparing NAEP and state assessment results in reading and mathematics that began in 2003 at a time when there was considerable policy debate about whether differences in state proficiency rates were due to differences in test content or differences in the stringency of proficiency cut points (McLaughlin et al., 2008a, 2008b). These important, policy-relevant reports have been continued with every assessment cycle to the present day (Ji et al., 2021).

- Replication of the Dogan (2019) study is the most urgent immediate need, given the current policy context and conclusions being drawn about the CCSS based on NAEP data.
- More generally, separate research and development studies will also be needed whenever major decisions are made that could affect the meaning of the construct being assessed and could, in turn, lead to differences in policy conclusions. As was noted previously, bridge studies may not be able to detect construct changes in the short term when two versions of a construct are strongly correlated.

CONCLUSION

The National Assessment Governing Board (NAGB) already has a thoughtful and comprehensive Assessment Framework Development Policy (2022) that attends to the importance of curriculum-neutral frameworks—encompassing what is currently taught and anticipated future learning goals, based on research and other sources of evidence. NAGB’s policy also clearly recognizes the fundamental tension between the need to keep up with changes in a subject-matter discipline and, at the same time, the need to maintain stability and comparability with the past, so as to report accurately on gains or declines in achievement. This conceptual and substantive tension between up-to-date frameworks and protecting trend can be exacerbated logistically because of the scale of effort required to develop new frameworks and the amount of time that has elapsed between the start of framework review and implementation of a new or even revised assessment.

The central recommendation of this paper is that NAGB should develop a new policy to enable smaller and more frequent updates to existing frameworks. This idea, termed an “evolutionary” approach by members of the NAEP Validity Studies (NVS) Panel, is consistent with recommendations from the recent NASEM (2022) consensus study report and the earlier expert panel report on the future of NAEP (NCES, 2012). One agreed-upon recommendation in these reports, about how to support this change, is to expand the responsibilities of standing subject-matter committees. This would mean that experts, already familiar with NAEP’s purposes and structures, would monitor evidence from the field and propose needed framework changes to NAGB.

Making this shift to an incremental or evolutionary approach will be more complicated, however, than merely broadening the charge to standing subject-matter committees. As outlined in this paper, NAGB will need to redeploy resources to gather information—for example, on state and international frameworks, relevant research, and professional standards—on an ongoing basis, rather than only after the launch of a new framework development process. An ongoing process will necessarily need to be more streamlined; thus, NAGB will have to determine the appropriate extent of stakeholder reviews for small or medium-size revisions. Most importantly, *NAGB will need to decide how a new policy for evolutionary revisions should articulate with its existing policy for new frameworks*. For example, NAGB implicitly reserves the right to decide that a significant revision proposed conceptually by a standing subject-matter committee is major enough to warrant invoking NAGB’s existing, full-scale process for development of new frameworks.

There are, of course, potential downsides to an evolutionary approach to framework revisions. When differences between versions of a construct are blurred or smoothed by incremental changes, evidence of progress on the new construct could be dampened or obscured. Preference for an evolutionary approach is based on the greater utility associated with short-term comparisons. It is also true that major conceptual shifts—like the 1990s difference between Long-Term Trend and Main NAEP and the new NAEP Science Framework and new trend in 2009, in anticipation of the Next-Generation Science Standards (2013)—are rare. More often, subject-matter committees and NAGB will be trying to respond to changes in the field that are themselves gradual and do not have a clear inflection point.

Conclusion

If NAGB adopts an evolutionary approach, bridge studies would not be needed for every modification. However, because bridge studies will still be needed for medium-size and major construct changes, studies should be undertaken to determine how robust this methodology is for detecting real differences; or is there a confirmatory bias when there is substantial overlap in item pools between adjacent assessments? Such overlap supports joint scaling and short-term trend interpretations, but it means that bridge studies do not really address questions relevant for research purposes about changes in construct meaning over longer term framework revisions.

NCES and NAGB also need to be alert to those occasions when framework decisions could be affecting policy inferences from NAEP data. As an example, the Dickinson et al. (2006) and Beaton and Chromy (2010) studies comparing LTT and Main NAEP were undertaken because of conflicting interpretations of NAEP data by both researchers and journalists. More recently, for good reasons (recapitulated above), NAGB did not consider revising NAEP Reading and Mathematics Frameworks in response to the CCSS, adopted by some but not all states. Now, however, NAEP data are being used to draw conclusions about the effectiveness of the CCSS despite the Dogan (2019) study, which found substantial misalignment between NAEP and CCSS in mathematics. As emphasized in the NVS Panel response (Hughes et al., 2019), NCES should neither make score adjustments nor allow states or districts to “customize” NAEP results. But NCES should undertake separate studies as part of its Research and Development series to inform policy researchers when differences in construct definitions could be shaping policy conclusions.

NAEP has rightly been regarded as the gold standard for measuring educational achievement. Its subject-matter *frameworks* by which disciplinary content domains are defined and its *trend* data are two of its most precious assets. The recommendation to gradually revise frameworks will enable NAEP to stay at the cutting edge in representing content domains and at the same time maintain sufficient comparability to enable monitoring of achievement gains and losses over time.

REFERENCES

- Augustus F. Hawkins-Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988, Title III, Pub. L. No. 100-297, U.S.C. § 3401-3403 (1988).
<https://www.govinfo.gov/content/pkg/STATUTE-102/pdf/STATUTE-102-Pg130.pdf>
- Alexander, L., & James, H. T. (1987). *The nation's report card: Improving the assessment of student achievement*. National Academy of Education.
- Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. H. (2009). *Mapping state proficiency standards onto NAEP Scales: 2005-2007* (NCES 2010-456). U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences.
- Beaton, A. E., & Chromy, J. R. (2010). *NAEP trends: Main NAEP vs. long-term trend*. American Institutes for Research.
- Center on Standards, Alignment, Instruction, and Learning (C-SAIL). (2020). *A 20/20 vision for standards-based reform*. C-SAIL Publications, 23. University of Pennsylvania.
<https://repository.upenn.edu/c-sail/23>
- Daro, P., Hughes, G. B., Stancavage, F., Shepard, L., Webb, D., Kitmitto, S., & Tucker-Bradway, N. (in press). *A comparison of the 2017 NAEP mathematics assessment with current-generation state assessments in mathematics: Expert judgment study*. American Institutes for Research.
- Dickinson, E. R., Taylor, L. R., Koger, M. E., Deatz, R. C., & Koger, L. E. (2006). *Alignment of long term trend and main NAEP*. HumRRO.
- Dogan, E. (2019). Appendix: Analysis of recent NAEP TUDA Mathematics results based on alignment to state assessment content. In G. Hughes, P. Behuniak, S. Norton, S. Kitmitto, & J. Buckley (Eds.). *NAEP Validity Studies Panel responses to the reanalysis of TUDA Mathematics scores*. American Institutes for Research.
- Figlio, D., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (pp. 383-421). North-Holland.
- Glaser, R., & Bryk, A. S. (1987). Commentary by the National Academy of Education. In L. Alexander & H. T. James (Eds.), *The nation's report card: Improving the assessment of student achievement*. National Academy of Education.
- Hughes, G., Behuniak, P., Norton, S., Kitmitto, S., & Buckley, J. (2019). *NAEP Validity Studies Panel responses to the reanalysis of TUDA mathematics scores*. American Institutes for Research.

References

- Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., Rust, K., & Burg, S. (2020). *2017 NAEP transition to digitally based assessments on mathematics and reading at grades 4 and 8: Mode evaluation study*. National Center for Education Statistics.
<https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional-whitepaper.pdf>
- Ji, C. S., Rahman, T., & Yee, D. S. (2021). *Mapping state proficiency standards onto the NAEP scales: Results from the 2019 NAEP reading and mathematics assessments* (NCES 2021-036). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
<https://nces.ed.gov/nationsreportcard/subject/publications/studies/pdf/2021036.pdf>
- Klein, S. P., Hamilton, L., McCaffrey, D. F., & Stecher, B. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives*, 8(49), 1–20.
- McLaughlin, D., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., & Wolman, M. (2008a). *Comparison between NAEP and state reading assessment results: 2003*. (NCES 2008-475). National Center for Education Statistics.
- McLaughlin, D., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., & Wolman, M. (2008b). *Comparison between NAEP and state mathematics assessment results: 2003*. (NCES 2008-475). National Center for Education Statistics.
- National Academy of Education (NAEd). (1992). *Assessing student achievement in the states: The first report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1990 Trial State Assessment*. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.).
- National Academies of Sciences, Engineering, and Medicine (NASEM). (2013). *Next Generation Science Standards: For states, by states*. National Academies Press.
<https://nap.nationalacademies.org/catalog/18290/next-generation-science-standards-for-states-by-states>
- NASEM. (2017). *Principles and practices for a federal statistical agency: Sixth edition*. National Academies Press. <https://doi.org/10.17226/24810>
- NASEM. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.
- NASEM. (2022). *A pragmatic future for NAEP: Containing costs and updating technologies*. National Academies Press. <https://doi.org/10.17226/26427>
- National Assessment Governing Board (NAGB). (2013). *General policy: Conducting and reporting the National Assessment of Educational Progress*.
- NAGB. (2019). *Reading framework for the 2019 National Assessment of Educational Progress*.
- NAGB. (2022). *Assessment framework development policy statement*.
- National Center for Education Statistics (NCES). (2012). *NAEP: Looking ahead—leading assessments into the future*.

References

- National Governors Association. (2010). *Common Core State Standards*.
- National Research Council (NRC). (1996). *National Science Education Standards*. Coordinating Council for Education, National Committee on Science Education Standards and Assessment. National Academies Press.
- National Research Council (NRC). (1999). *How people learn: Brain, mind, experience, and school*. National Academies Press.
- NRC. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
- Nellhaus, J., Behuniak, P., & Stancavage, F. B. (2009). *Guiding principles and suggested studies for determining when the introduction of a new assessment framework necessitates a break in trend in NAEP*. American Institutes for Research.
- Polikoff, M. S., Petrilli, M. J., & Loveless, T. (2020). A decade on, assessing the impact of national standards has Common Core failed? *Assessing the Impact of National Standards. Education Next*, 20(2), 72–81.
- Rosenberg, S. (2022, April). *Considerations for smaller, more frequent changes to NAEP assessment frameworks*. National Assessment Governing Board.
- Song, M., Yang, & Garet, M. (2019). *Effects of states adoption of college- and career-ready standards on student achievement*. American Institutes for Research.
- Webb, N. L. (1997). *Research monograph No. 6: Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers.

NAEP Framework Development Reaction Paper

Carol Jago, Consultant

September 2022

The decision regarding whether or not to implement more frequent, incremental changes to NAEP frameworks will have an enormous impact on the future of NAEP and the Governing Board. After reading and pondering the Assessment Framework Development policy statement, the 2022 recommendations of the National Academies of Sciences, Engineering and Medicine, and the paper developed for the Assessment Development Committee's consideration by Sharyn Rosenberg, I come down in favor of this change.

First, all NAEP frameworks are not created equal.

In one category are the Reading and Mathematics Frameworks for assessments that must be conducted every two years. A second category includes science, civics, and U.S. history — frameworks that apply to assessments scheduled with regularity.

The Writing Framework and Technology and Engineering Literacy Framework fall, in my opinion, into a third category on account of how rarely the Governing Board can afford to assess these subject areas. TEL will next be assessed in 2028 and Writing in 2030. Geography, Economics, Foreign Language, and Arts are currently labeled “inactive” on the Governing Board framework website.

I believe it is essential that frameworks for subject areas in the first two categories be attuned to developments in the field and with changes to state content standards. Reading, mathematics, science, civics, and U.S. history are content areas of great interest to the public, making the argument for maintaining trend for these assessments a strong one. Taxpayers want to know how students are performing in these subjects over time.

There is no escaping the fact that maintaining a balance between content relevance and trend will be a continuing challenge.

Second, technological advancements in how assessments are delivered will likely require periodic (both incremental and maybe not-so-incremental) updates to NAEP frameworks.

The shift from paper and pencil to device delivery of an assessment has been absorbed into the NAEP testing environment while at the same time maintaining trend in reading and mathematics. As additional technological developments potentially alter how items are presented to students — I am thinking here particularly of further developments in the presentation to test-takers of scenario-based items — there could be a need to make further changes to the Reading and Mathematics Frameworks even though they both have been revised and adopted by the Governing Board quite recently.

There may also be a need to make changes to the Reading and Mathematics Frameworks based upon lessons learned from the first administrations of assessments based upon the new

frameworks. I cannot imagine that anyone wants to revisit these two frameworks in their entirety any time soon and expect the same will be true for the Science Framework that is in development. Adjustments rather than wholesale revision seems much the more practical solution.

Considering NAEP frameworks “living documents” may be going too far, but frameworks should be able to adapt when:

1. Objectives in a new framework cannot be measured in the manner intended.
2. Advances in technology that make possible more finely tuned assessments.

Who should decide when revisions to a framework are needed?

The ultimate decision, of course, rests with the National Assessment Governing Board, but I believe NAEP standing committees are ideally positioned to recommend to the Board when developments in the field and/or in assessment technology necessitate incremental framework updates.

These standing committees, made up of individuals not only knowledgeable in their fields but also deeply knowledgeable about the NAEP assessments, could also be charged with identifying and recommending to the Governing Board when disruption in the field is so great that a full-scale revision process is needed.

Standing committee recommendations must, of course, be grounded in research regarding changes to standards, curricula, instruction, and assessment. Regular reports to the Assessment Development Committee and to the Governing Board should keep Board members apprised of the issues under discussion and the progress of such discussion. Surprises should be kept to a minimum. Support staff could be charged with this task.

While it is important not to be distracted by every shiny new thing, to retain their place as authoritative measures of student performance, NAEP assessments must be relevant.

Responses to questions regarding changes to NAEP Assessment Framework policy and processes:

1. Defining what is meant by making smaller, more frequent updates to frameworks

After every administration of each assessment, the relevant Steering Committee should examine results and reflect upon whether changes to the framework are needed. As a result of their more frequent administrations, this would occur more often for Reading and Mathematics than in other subject areas. It should be noted that this will place a heavier workload on those two Steering Committees as well as upon their support staff, in addition to the Board staff.

I don't think reducing the recommended interval for framework revision to 4-6 years from 10 years would make much of a difference. More important than any predetermined time frame is

careful consideration of the extent of change within a field and the extent to which advances in assessment technology are affecting student performance.

2. Clarifying whether a new approach is intended to augment versus replace the existing policy and processes for updating frameworks

Adding new policy on top of existing policy only makes for more work for staff and more detailed reports (that too few people read). The weeds just keep getting thicker and thicker.

I recommend that the Governing Board clarify to stakeholders and the public that they are instituting a new policy for updating frameworks that replaces pre-determined intervals for revision with a process for evaluating the need for changes following every assessment.

It is likely that such a policy change will have trickle down effects. A few that I envision include:

- a. Increased responsibility and workload for Steering Committees
- b. Refocusing the charge for the Assessment Development Committee to include making recommendations to the full Board when incremental changes to the framework are needed and when full revision is necessary.

This change in the charge to the ADC is in accord with Principle 5 of the March 2022 Assessment Framework Development policy: “The ADC shall be responsible for monitoring framework development and updates that result in recommendations to the Governing Board on the content and format of each NAEP assessment.”

- c. Shifting funding from contractors who manage large-scale framework development processes to staff supporting Standing Committees. It would be folly to suggest that the Reading, Mathematics, Science, U.S. History, and Civics Steering Committees could take on this additional work without funding to support their work.
3. Exploring a process for updating frameworks more often with smaller changes each time

Updating frameworks more often with incremental revisions should be more efficient, but the Governing Board will need to be careful how changes to policy are communicated to the public. We don’t want it to appear that the process is becoming less transparent. It might be a good idea to explore how PISA and TIMSS have handled this issue. Currently the process for NAEP framework development includes extended periods of time for public comment. Inevitably this step slows down the process of keeping frameworks updated to glacial speed.

Taking a more positive spin on this issue, the Governing Board could publicize the smaller changes to an adopted framework in succinct, “Good News!” updates. It will be important to make clear to all that the ultimate purpose of any change in a framework is to improve the assessment thereby improving education.

NAEP frameworks, particularly in Reading and Mathematics, are often used by state departments of education and publishers as they develop curriculum, instructional materials, and state tests. While extended, open public comment feels very democratic, input from key stakeholders is more valuable (CCSSO, Council of the Great City Schools, professional organizations, etc.). It should be possible to garner input from these groups in a more timely fashion than what has been done in the past.

The recently adopted changes to the NAEP Assessment Framework Development policy (March 2022) call for increased Board involvement. The Board must be kept abreast of discussion regarding incremental changes to frameworks but doesn't need to be involved in every discussion. As indicated in item G of Principle 5: Role of the Governing Board:

“At the conclusion of the framework development or update process, the Governing Board shall take final action on the recommended framework and specifications. The Governing Board shall make the final decision on the content and format of NAEP assessments. In addition to the panel recommendations, the Board may take into account other pertinent considerations on the domain and scope of what should be assessed, such as the broader policy context of assessment in the subject area under consideration.”

In terms of basing a decision about framework updates upon the extent to which a field is unified, in my experience, the fields of reading and mathematics are and are likely to remain polarized. Civics and U.S. History will always be political battlegrounds. Compromise, for the sake of the best possible assessment, may be a more reasonable goal than consensus.

4. Analyzing the potential impact of new processes on important factors such as costs and the ability to maintain trend

Other commentators on this team will provide much more informed views regarding the ability to maintain trend than I ever could. From a lay person's point of view, I don't think every incremental change to a framework should require an expensive bridge study. Common sense (I know, I know, common sense isn't measurable) can often be a guide to sound decision-making.

The current, 2-year process for framework development is costly in many ways, including in precious Board meeting time.

PISA and TIMSS staff should be able to provide the Governing Board with broad estimates of what their incremental change processes have cost the organizations. They should also be able to offer insight into how they have managed to maintain trend while implementing incremental changes to their assessments.

Maintaining trend is directly related to the value of NAEP, particularly to policy makers. I fear that breaking trend every time the wind shifts could be seen as a reason to abandon the national assessment altogether.

5. Estimating what resources and/or structural changes may be needed to implement new processes

If more responsibility for framework upkeep is shifted to NAEP Standing Committees, there is likely to be a need for increased stipends/honorariums for members of those committees along with the increased cost of support staff time.

If technological advances in the administration of NAEP result in cost savings, this funding could be redirected to the highest-profile Standing Committees' work and support.

6. Anticipating potential unintended consequences, such as reopening settled debates

The “debates” in reading and mathematics never really end; they only settle down for short periods of truce and then resurface with renewed vehemence.

NAEP results play a role in this endless tug-of-war, particularly when student performance is disappointing. The pendulum metaphor is clichéd but apt.

Alas, we are likely to be revisiting certain issues again and again. Nothing is “settled” for long.

7. Understanding what other potential solutions to the identified problems might exist

NAEP frameworks are currently voluminous documents. What if they were reconceived as much less detailed guidelines for a national assessment? More like a roadmap for item development than a description of the field.

Currently NAEP frameworks read something like national pronouncements. For example, these two paragraphs from the 2018 U.S. History Framework:

“This framework identifies the main ideas, major events, key individuals, and unifying themes of American history as a basis for preparing the 2018 assessment. The framework recognizes that U.S. history includes powerful ideas, common and diverse traditions, economic developments, technological and scientific innovations, philosophical debates, religious movements, and the interconnection of all these forces.

The teaching of history should introduce students to the process of historical inquiry. This process requires critical examination of evidence, thoughtful consideration of conflicting claims, and careful weighing of facts and hypotheses. Historical inquiry provides experience in the kind of reasoned and informed decision-making that should characterize each citizen's participation in our American democracy.”

Reflecting upon Board discussion of the Reading Framework, disagreement came down to philosophical differences.

Maybe NAEP frameworks try to do too much.

Measuring Change in a Changing World: Toward Efficient Measurement of Aggregate Educational Progress

Andrew Ho, Harvard Graduate School of Education

October 2022

In this short paper for the National Assessment Governing Board, I address Assistant Director for Assessment Development Sharyn Rosenberg’s list of questions related to recommendations for improving the framework update process for the National Assessment of Educational Progress ([Rosenberg, 2022](#)). My recommendations are:¹

- 1) only task framework panels for new subjects or rarely administered subjects that require a relaunch;
- 2) for all other subjects, create (or revise the charge of) standing framework committees to advise the Board and consult with the National Center for Education Statistics (NCES) on necessary incremental changes to existing frameworks and specifications; and
- 3) to adopt different perspectives on trend reporting and validation, including
 - a. a “moving window” perspective on trend validation,
 - b. three different levels of “bridge studies,” and
 - c. differences in validation for developing an index vs. a scale.

The challenge and importance of measuring educational progress

Measuring educational progress requires embracing an apparent paradox. We cannot measure change unless we hold our measures constant. But what students learn and how students learn are always changing. If we change our measures, we weaken our basis for measuring progress. And if we do not change our measures, we weaken our basis for measuring what students are currently learning. Rosenberg ([2022](#)) nicely captures this tension and the decades-long debate between measuring change and changing measures.

I have long seen the reporting of meaningful trends as NAEP’s most important responsibility ([Ho, 2020](#)). No other entity has the history, authority, or capability of NAEP to report meaningful, representative trends. And no state- and national-level educational question is as important as whether and how we are making educational progress. As Rosenberg ([2022](#)) suggests, we should be clear what goal we are trying to achieve. My goal is to report meaningful trends. I believe incremental framework updates achieve this goal. If the goal were to update and adopt consensus frameworks, incremental framework updates may not be the best strategy. I

¹ My recommendations here are compatible with two reports that I coauthored that Rosenberg ([2022](#)) cites in her background document: the Future of NAEP Report ([2012](#)) and the Pragmatic Future for NAEP Report ([2022](#)). They are also compatible with opinions and writing that I contributed as a member of the National Assessment Governing Board from 2012 to 2020. I am grateful to my coauthors and colleagues for informing my opinions through years of discussion and debate. However, the opinions in this paper are mine and not necessarily those of my colleagues, and I am responsible for all errors and misconceptions.

recommend them because incremental framework updates are the best way to report meaningful trends. What are incremental framework updates, and why are they necessary?

Recommendation 1: Reserve framework panels for subjects that are new or require relaunch

Current framework panels are better suited for revolution than evolution. As the Assessment Framework Development Policy presents the panels ([NAGB, 2022](#)), they are appointed like a Task Force to accomplish the task of creating or updating a framework. Panelists will not sign up for the position unless such a task interests them. A common scholarly instinct in authorship of a document is to frame the document around the question, “what is new or novel?” The creation and composition of framework panels is poorly suited to incremental progress.

The existing Principle 2 in the policy ([NAGB, 2022](#)) does not sufficiently distinguish between framework creation and framework updates. It remains a holdover from the original 2002 policy when the primary goal was development of new frameworks ([Orr, 2021](#)). Under sub-principle 2d, the policy suggests that the process can be less substantial for revisions, but there is insufficient distinction. The process is likely to result in panelists motivated toward substantial revision. Framework panels remain a useful tool but should be reserved for new subjects.

The Governing Board may also wish to convene framework panels for infrequently assessed subjects measured only at a national level, particularly those that might require a “relaunch.” For example, for subjects offered at a greater-than-4-year interval, like Technology and Engineering Literacy after 2018, the Governing Board may discuss with a standing committee, shortly after a release, whether a “relaunch” is necessary based the time interval and on current and forecasted changes in how students learn the subject matter. The concern would be that, say, a 6-year interval between assessments would prevent an end-to-end comparison of achievement across that interval given changes in what and how students learn. At the 1- or 2-year mark in that interval, the Board, on the advice of a standing committee as constructed below, could move to relaunch the framework with a framework panel, as necessary.

Subjects that the Governing Board assesses more frequently (every 4 years or more frequently) should not, in my opinion, require relaunch nor require convening a framework panel. Shorter time intervals improve the likelihood of stable constructs across these intervals. For these subjects, as I describe below, I recommend standing committees that advise the board on updating content and specifications incrementally, to maximize the likelihood of reporting meaningful trends.

The longer the existing trendline and the more granular (state- and district-level) the aggregation for reporting, the more important it is to avoid a new framework panel and rely instead on standing committees tasked with incremental adaptation. This improves the likelihood of preserving and extending the historical record of educational progress and enabling helpful long-run comparisons among multiple jurisdictions. I do not believe that the Reading and

Mathematics frameworks should ever be relaunched in a manner that suggests a sudden and discontinuous “new Reading” or “new Mathematics.”

Recommendation 2: Create (or revise the charge of) standing framework committees to update frameworks for existing subjects incrementally

My recommendation to create or revise the charge of existing standing framework committees is compatible but in sum more specific and unreserved than the guidance from the Future of NAEP Report (2012) and the Pragmatic Future for NAEP Report (2022) to which I contributed. For reasons I explain above, current framework update panels threaten rather than advance the goal of reporting meaningful trends. Standing framework committees can build norms and expertise that better serve the goals of incremental progress and discourage seismic reinvention. This use of standing committees should replace, not augment, the existing process for framework update panels.

To serve these aims, standing committees should meet regularly with NAGB and advise on or direct framework revisions for every administration of their assessment. Under current assessment schedules, this implies revisions could affect assessments every two years for Reading and Mathematics and every 4 years for other subjects like Civics, U.S. History, and Science. These revisions may be prospective due to extended timelines for implementation given NCES constraints on item and task development and field testing.

Rosenberg (2022) suggests a possible solution whereby framework update panels develop new frameworks as illustrated by existing policy but leave NCES to implement them more incrementally, achieving the same incrementalism that enables reporting meaningful trends. While this approach is appealing and superior to the sudden implementation of new frameworks, it does not address my core concern that framework panels are inclined toward ambitious reinvention. Resulting changes may appear jarring or controversial to constituents even if NCES and the Governing Board implement the changes incrementally.

In contrast, the Governing Board should charge standing committees to maintain and improve frameworks. Membership terms that overlap and rotate, like Governing Board terms, can preserve institutional knowledge. Standing committees can also serve a useful bridging role between the Governing Board and NCES that deepens coordination and communication between NAEP governance and NAEP operations.

Recommendation 3: Adopt different perspectives on trend validation and reporting

I base my recommendations on a set of perspectives and principles that others may not necessarily share. Making these perspectives explicit may help to explain if not resolve disagreements about the best way to report and protect meaningful trends.

a) A “moving window” perspective on trend validation

As content and contexts shift over time, the change is gradual, not static and then suddenly different. Accordingly, I consider the goal of scale maintenance to be the goal of enabling scale meaning in any “moving window” of 4-8 years along a timeline, rather than ensuring rigid comparability of the most recent year of results to an originating baseline year. One metaphor that illustrates this comes from the proverb, “the bamboo that bends is stronger than the oak that resists.” The bamboo tree’s flexibility is akin to an equating design like NAEP’s, where developers deploy common items and tasks through a window and then retire them to allow new items to become common items. The oak tree’s inflexibility is akin to a reference set of items that stays constant over time, where new items are regularly linked back to this reference set. As the metaphor suggests, the former design enables flexible trend estimation within moving windows, whereas the latter design is destined to break as the content and context of learning changes.

Figure 1. Illustrating the proverb, “the bamboo that bends is stronger than the oak that resists,” to motivate continuous updating of common items rather than equating to an historical reference set. Images from [here](#) and [here](#).



A similar heuristic that Jack Buckley and I discussed in our conversations developing the Pragmatic Future for NAEP Report (2022) was the thought experiment known as the “Ship of Theseus.” The Greek historian Plutarch describes the Athenian preservation of the Ship of Theseus, where preservationists replaced planks of the ship whenever they decayed. The analogy to trends is that the replacement of items, content, and context is like the replacement of planks. The question is whether the ship is still the Ship of Theseus even after every plank is replaced. My perspective on this is, whether or not the ship is still the same as it originally was, it is similar enough to recent ships. It is sufficiently similar in a “moving window” of time.

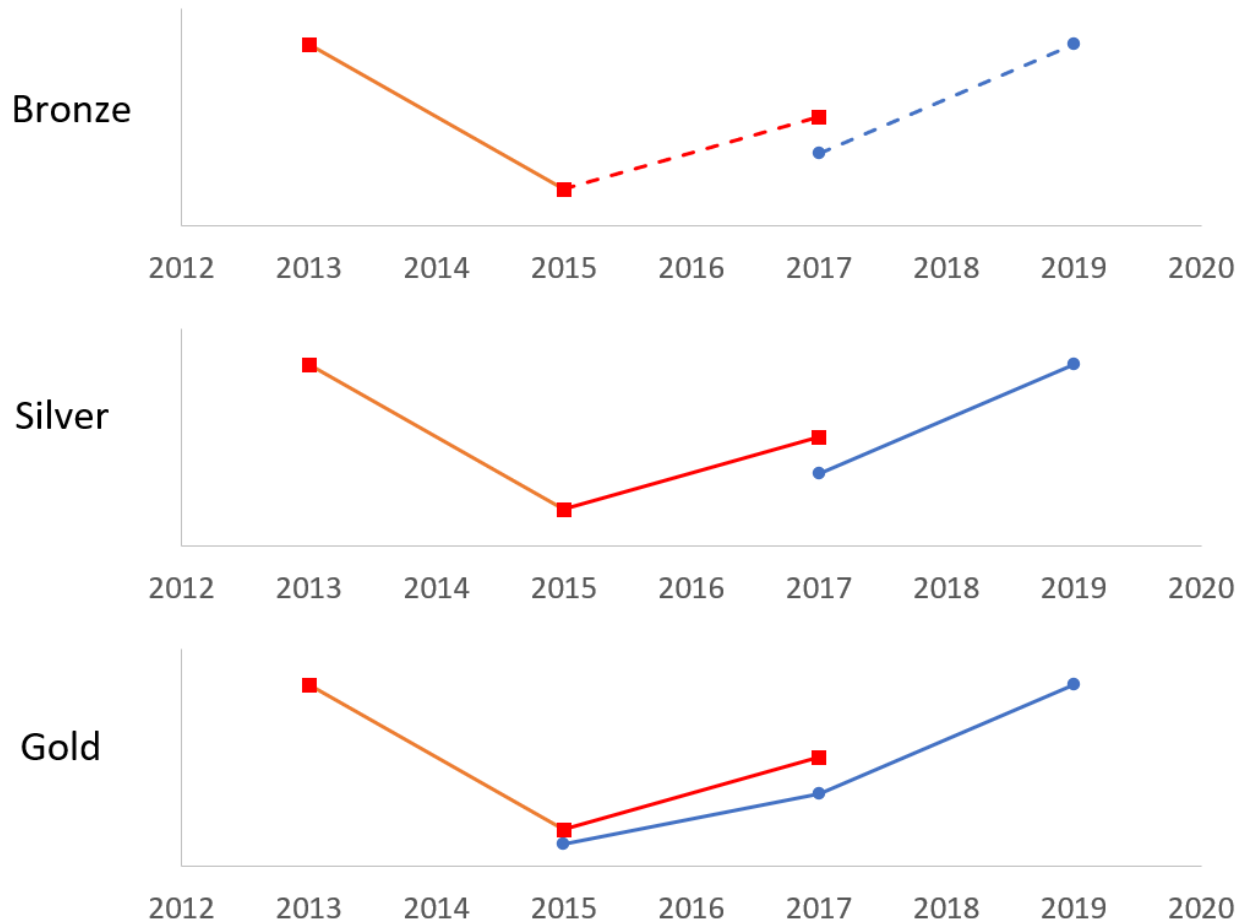
b) Three different levels of “bridge studies”

Bridge studies provide valuable evidence for trend validation. In cases where the Governing Board and NCES anticipate that trend maintenance will be difficult due to shifting content and contexts, high-quality bridge studies can act as a safeguard against a broken trend. Bridge studies

can support proposed transitions from existing to newer assessment content and context. The Governing Board can deploy higher quality bridge studies when maintaining trend is particularly challenging. I distinguish these bridge studies here as bronze, silver, and gold.

The bronze bridge study takes a relatively small number of common or calibrated items on the time 1 scale and embeds them into a time 2 administration. This is a standard test equating procedure where there is no anticipated substantial shift in construct, mode, population, or method over the intervening time period. Standard psychometric methods can evaluate whether items are functioning differently over time. If the old and new items are not comparable on the same scale, then developers may have to salvage an imprecise trend from the time 1 common items alone. It is also possible that the trend breaks altogether because items change inconsistently in their location or format. This trend may be more imprecise. It only guarantees a tentative measure of progress from time 1 to time 2 and only on the time 1 scale. The time 2 items establish a new baseline that is not comparable to the time 1 scale. The time 2 items may also form an imprecise baseline due to limited deployment. Reports sometimes show results from a bronze bridge study as staggered points shown in Figure 2 below. This is misleading because it implies that a simple shift can equate the trends across the break.

Figure 2. Illustrating bronze, silver, and gold bridge studies from 2015 to 2017. The bronze bridge study has relatively few common items embedded in 2017 and guarantees only a tenuous trend from 2015 to 2017. A silver bridge study administers both old and new tests in 2017 and guarantees a trend from 2015 to 2017 but not necessarily from 2015 to 2019. A gold bridge study offers two trends across a 2-year span and two separate 4-year trends with a 2-year overlap.



The silver bridge study administers two different modes, frameworks, or methods at timepoint 2. For example, if NAEP aimed to bridge from a paper-based test mode to a digitally based test mode, it could administer both at timepoint 2. Comparison of measurement properties, both within timepoint 2 and for the common administration from timepoint 1 to timepoint 2, provides additional reassurance that mode A is stable from timepoint 1 to timepoint 2 and that mode B is comparable to mode A in timepoint 2. If the bridge fails, then there is a common pretrend for mode A that ends at timepoint 2, and a new trend baseline that begins at timepoint 2 in mode B. Thus, if the bridge fails, there is a measure of progress from timepoint 1 to timepoint 2, but only for mode A.

The gold bridge study administers two different modes, frameworks, or methods, at two timepoints, in both timepoint 1 and timepoint 2. This was the approach that the Governing Board and NCES adopted for the transition to digitally based assessments. This “gold” or “double bridge” study allows for not only the comparison of mode A and mode B at both timepoints 1 and 2, but also whether the mode effect itself differs across time. If the bridge fails, there are ultimately two trends, one reportable in mode A from timepoints 1 to 2, and another trend reportable in mode B from timepoints 1 to 2. If the bridge fails, transparent index weights could enable users to combine the trends in terms of standard deviation units or report them side by

side. Either approach achieves the goal of trend interpretation. In this way, the gold bridge study can protect trend interpretations even if the bridge fails on psychometric grounds.

The bronze, silver, and gold bridge studies are tools in the NAEP toolkit whose costs should be anticipated in the investigation and implementation of possible improvements. Significant shifts may demand higher quality bridge studies. If standing committees recommend a moderate framework update, the Governing Board may protect trend with a silver bridge study. If standing committees believe a major framework update or mode transition is necessary, the Governing Board may decide that a gold bridge study is necessary.

If there are multiple major shifts required that may not be time-sensitive, the Governing Board may choose to stagger them with different bridge studies to support each transition. Methodological improvements may require only a bronze bridge study, if simulations and theory can show equivalence. However, if methodological improvements must coincide with changing administration conditions, threats to trend may warrant higher quality bridge studies. Ultimately, the job of NAEP is in its name: assess educational progress. NAEP innovation and NAEP improvement should not come at the expense of NAEP's primary purpose. Anticipating mode effects and construct shifts will help to plan for bridge studies that can fulfill the Governing Board's charge to measure educational progress.

c) Differences in validation for developing an index vs. a scale.

The Future of NAEP Report (2012) made the apt comparison to the Consumer Price Index (CPI) that monitors inflation. The CPI rotates goods through a “market basket” as goods become antiquated; for example, a rotary phone is eventually replaced by mobile phone. Validation of this “linkage” is not supported by checking what a rotary phone costs now—it is irrelevant. Instead, the goal is to track inflation over a “moving window” that is always relevant in its era. The same principle can hold for NAEP's monitoring of educational progress.²

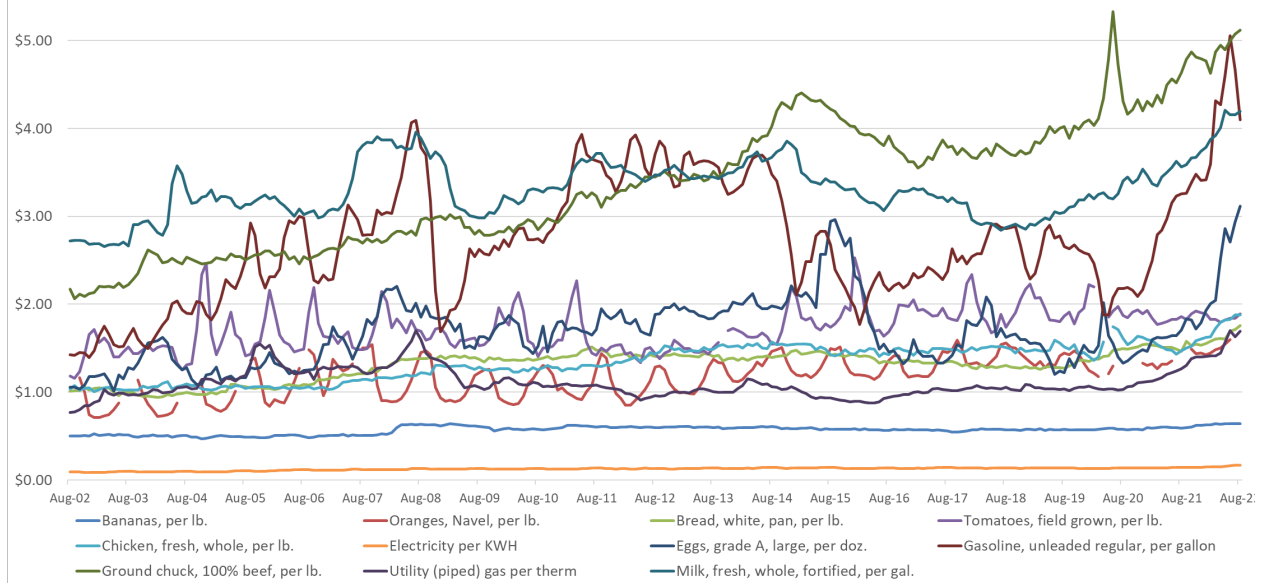
Traditional psychometric perspectives on scale maintenance argue that indexing is insufficient because it renders score interpretations incoherent. For example, the Governing Board's rich performance level descriptors suggest that students scoring in a particular region of a score scale can answer certain kinds of items correctly (NAGB, 2017). If item difficulties change in a nonlinear fashion, indexing can still result in an average score calculation, but the score will not maintain the same meaning, and performance level descriptions can become inaccurate descriptions of knowledge and skills for students in that score range.

From the perspective of indexing, such incoherence may not be consequential. Consider the goal of measuring the change in average consumer prices, as Figure 3 illustrates below. If gasoline becomes much more expensive but bananas do not, there may in fact be some incoherence,

² A National Research Council discussed NAEP Market Basket Reporting in 2001 to evaluate whether common items could be embedded in test forms to support cross-jurisdictional comparisons. Rather than evaluate linking between jurisdictions, my goal here is to support linking over time.

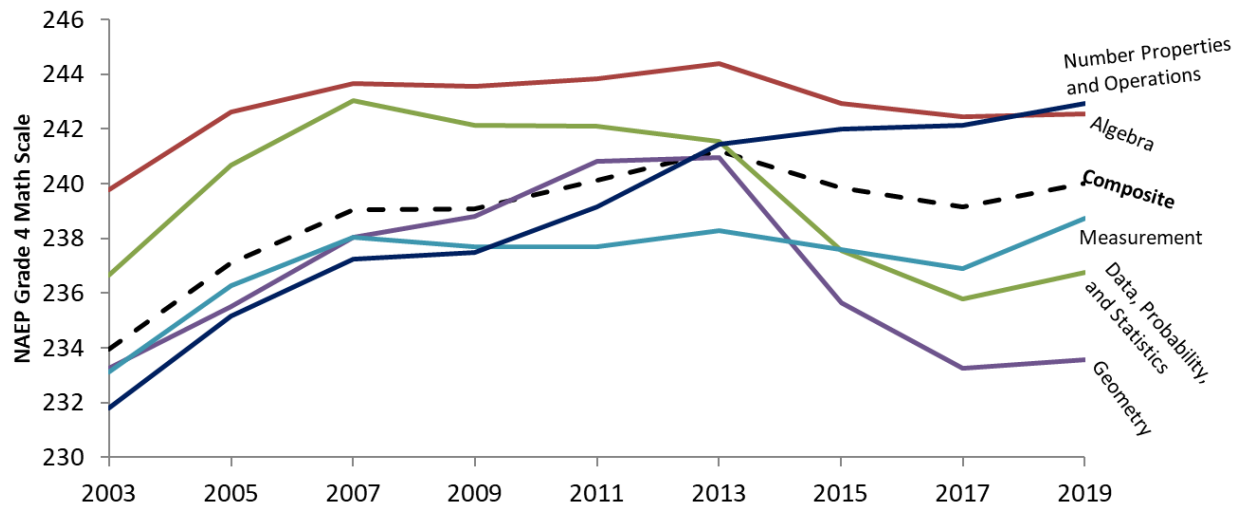
because people who eat more bananas and take public transportation may not be experiencing the same price increases as people who drive and eat beef. However, the CPI's target inference is at an aggregate level, and inflation has important economic implications. Coherence is a small sacrifice to answer an important economic question.

Figure 3. Prices of goods in U.S. dollars, selected items. Source: [Bureau of Labor Statistics](#).



The NAEP scale, too, functions as an index, a weighted average across multiple subscales. NAEP subscales reveal similarly inconsistent trends across subscales. Figure 4 shows that the average public-school student has made steady progress in Number Properties and Operations since 2003. Recent declines in the NAEP composite score have been driven by Algebra, Geometry, and Data, Probability, and Statistics.

Figure 4. NAEP Grade 4 Mathematics Subscale Trends, 2003-2019. Source: [NAEP Data Explorer](#).



Because NAEP is a low-stakes measure of aggregate educational progress that never reports scores at an individual- or even a school-level, I am unconcerned by potential incoherence in performance level descriptions over time. Such descriptions can be revised in a manner recommended by a recent National Academies Evaluation (2017). The Governing Board should not be distracted from reporting educational progress by short-sighted psychometric principles that assume incorrectly that there are high-stakes interpretations of student-level scores. These do not exist. Prioritizing psychometric coherence amounts to misconceiving the job of NAEP as testing students when its job is monitoring educational progress.

The Governing Board and NCES can and should deploy bridge studies to evaluate and inform interpretations of educational progress. However, by adopting an indexing perspective, a failed bridge study may not necessarily prevent the NAEP program from reporting meaningful progress. Transparent weights across different constructs can enable calculation of an index of educational progress.

In economics, scholars and policymakers understand the societal importance of measuring the change in average prices, even as market baskets change. In education, scholars and policymakers should similarly recognize the societal importance of reporting the change in average student knowledge, skills, and abilities, even as that collection, too, may change.

By adopting “moving window” and “indexing” perspectives and deploying bridge studies strategically, the Governing Board and NAEP can continue to serve its singular and essential mission of reporting meaningful educational progress.

References

- Haertel, E., Beaugard, R., Confrey, J., Gomez, L., Gong, B., Ho, A., et al. (2012). NAEP: Looking Ahead – Leading assessment into the future. Recommendations to the Commissioner. National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/future_of_naep_panel_white_paper.pdf
- Ho, A. D. (2020). Departing remarks to the National Assessment Governing Board. Retrieved from <https://scholar.harvard.edu/files/andrewho/files/andrewhodepartingnagb.pdf>
- National Research Council. (2001). NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting. Washington, DC: The National Academies Press. <https://doi.org/10.17226/10049>
- National Academies of Sciences, Engineering, and Medicine. (2017). Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress. Washington, DC: The National Academies Press. <https://doi.org/10.17226/23409>
- National Academies of Sciences, Engineering, and Medicine. (2022). A Pragmatic Future for NAEP: Containing Costs and Updating Technologies. Washington, DC: The National Academies Press. <https://doi.org/10.17226/26427>
- National Assessment Governing Board (2017). Mathematics Framework for the 2017 National Assessment of Educational Progress. Retrieved from <https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/mathematics/2017-math-framework.pdf>
- National Assessment Governing Board (2022). Assessment Framework Development. Policy Statement. Retrieved from <https://www.nagb.gov/content/dam/nagb/en/documents/policies/assessment-framework-development.pdf>
- National Research Council. (2001). NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting. Washington, DC: The National Academies Press. <https://doi.org/10.17226/10049>.
- Orr, C. S. (2021). History, Policy, and Decision Points for Developing NAEP Frameworks. Retrieved from <https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/History-of-NAEP-Frameworks-Report-Final.pdf>

Rosenberg, S. (2022). Considerations for smaller, more frequent changes to NAEP assessment frameworks. Retrieved from https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/quarterly-board-meeting-materials/2022-05/6-Assessment_Development_Committee.pdf

NAEP Framework Development
Jessica Baghian, Watershed Advisors
September 2022

Executive Summary

Across the world, some countries have national assessments to determine the academic success and struggles of their students. The United States does not. For Americans, NAEP is our best and most trusted view into student academic success at scale. It allows educators, researchers, politicians, and taxpayers to understand the national educational attainment trends of American students.

In particular, for state leaders, NAEP results matter deeply. As a former state leader overseeing measurement and academics in Louisiana, our state relied on NAEP to (1) ensure the rigor of our state-defined proficiency benchmarks, (2) compare our ranking relative to peer states, and (3) understand our state's trends in comparison to the rest of the nation.

Now, the National Assessment Governing Board (Governing Board) is contemplating updating assessment frameworks on a more frequent basis. A few reasons have been offered for more frequent change - cost; relevance; operational adjustments based on lessons learned; and smaller, more frequent changes instead of infrequent, bigger changes. As a former state leader, I was asked to provide a response to the Governing Board's considered change and the reasons offered for the potential change. In the memo that follows, I offer responses and feedback to each.

Overall, more frequent consideration of changes to NAEP carries significant risk. Every time test makers change a test, it risks breaking the national trend. Should the NAEP trend line break, America does not have a strong fall back. Even if the trend is maintained, frequent changes risk sowing doubt in the assessment's reliability. Therefore, the NAEP trend line and the trust it garners must be protected above all else.

Of the reasons offered for more frequent changes, only the last stands up to scrutiny. If test makers need to change six things, for example, it seems reasonable that making two changes per cycle for three cycles (or, six years) is preferable to six changes all in one year. It is easy to imagine this "gradual operationalization" of changes protecting the trend, instead of introducing more risk.

However, tests are forever imperfect tools made by professionals known for their perfectionist tendencies. If framework committees meet and changes are allowed, changes will almost inevitably be made. Every such change to the framework, no matter how small, inevitably creates some noise in the results.

Therefore, I recommend the following:

- The Governing Board should prioritize maintaining stable trend lines and, therefore, review the framework only once every 10 years.
- When frameworks are updated at the ten-year mark, gradual operationalization should be allowed. The timeline for such operationalization should be set at that time.

Background

Often called “The Nation’s Report Card,” NAEP is a critical tool for state leaders because it provides a common measure of student achievement across states and over time.¹ The United States does not have national standards, national curricula, or a national assessment. Instead, each state develops its own unique assessment aligned to state-specific academic standards and those assessments change, often significantly, with turnover of state education leaders. As a result, it is not possible to use state summative assessments to compare state achievement.² NAEP is the best proxy measure for understanding comparability across states and over time, a critical need for state education leaders and advocates.

Why NAEP Matters to State Leaders

As a former state executive overseeing measurement and academics for more than 700,000 students in Louisiana, NAEP served as a key checkup on the health of our K-12 system. My colleagues and I valued NAEP because it provided a reliable measure of how well the state’s K-12 system was performing relative to rigorous grade-level benchmarks, to other states, and to national trends in achievement. Specifically:

- NAEP holds a high bar for proficiency, against which state test proficiency markers should be compared. NAEP allows states to compare their own performance against a national bar for excellence, ensuring that all states are driving toward a rigorous K-12 experience. Ultimately, NAEP keeps states honest in their own evaluations of achievement and protects historically underserved students from a “race to the bottom” on proficiency determinations.
- NAEP allows for comparisons across states. NAEP data are hugely important for states to understand their ranking relative to other states. States can use NAEP rankings to contextualize their own performance. Because state summative assessments measure different curricula and standards, NAEP is the best available tool for comparing achievement across states.
- NAEP offers insight into a state’s performance in the context of national trends. It is not enough for state leaders to know how their students are achieving in the current year. To

¹ IES National Center for Education Statistics, “[An Introduction to NAEP](#).” (2010)

² IES National Center for Education Statistics, “[An Introduction to NAEP](#).” (2010)

truly understand the health of their K-12 system, state leaders need to know which direction their students are moving and how that compares to national trends.

Recommendation

NAEP has the unique ability to provide a reliable, accurate measure of trends in achievement nationwide. State leaders value NAEP because it delivers what no other assessment tool can—it provides a yardstick for states to measure their performance against rigorous, grade-level benchmarks, both in comparison to other states and in the context of national trends.

NAEP data are hugely important for state leaders to understand the health of their K-12 systems. That said, any noise in the test results or disruptions in the trend lines makes it difficult for state leaders to know how best to use the data. To ensure NAEP continues to deliver valuable information to state leaders, the Governing Board should prioritize maintaining trend lines and minimizing noise above the perceived need for increased relevance. To do this:

- The Governing Board should prioritize maintaining stable trend lines and, therefore, review the framework only once every 10 years.
- When frameworks are updated at the ten-year mark, gradual operationalization should be allowed. The timeline for such operationalization should be set at that time.

Recommendation Rationale

The paper prepared by Dr. Sharyn Rosenberg offers four primary rationales as to why NAEP may be updated more frequently: (A) reducing costs; (B) increasing relevance by reflecting changes in the field in a more timely manner; (C) greater opportunities to adjust operations based on lessons learned in the first administration of a new framework; and (D) minimizing the risk to trend by making smaller changes, rather than an accumulated larger set of changes. This memo offers reactions and solutions to address the concerns implied in these rationales for more change.

A. Reducing Costs

As Dr. Rosenberg notes, more studies are needed to assess cost. That said, as a former test maker, I am extremely skeptical that allowing test makers to more frequently change assessments will reduce cost. More change translates to more item creation, more committees, more standard setting, and more reporting adjustments. It is very hard to imagine how this reduces cost.

Even if more frequent updates would reduce cost, this is still an insufficient justification for risking the ability to fairly compare results across states and over time. America spends trillions on education. It is hard to imagine savings sufficient to justify risking the national value of NAEP.

B. Increasing relevance

As a state policymaker, NAEP was relevant for many reasons, but it did not dictate the academic standards and content learned by Louisiana’s children. And in a country driven by federalism, many would contend it should not. Therefore, year-over-year framework reconsiderations and related tweaks, as are made in many state assessments, are not necessary for NAEP. In fact, they risk NAEP’s greatest value—the trend line and the valid state comparisons.

One could imagine a few scenarios where instructional responsiveness becomes more important. For example, if research-based consensus emerged about a serious flaw in NAEP’s design or if the academic experience of children across America significantly changed, NAEP may need to adjust in order to maintain validity. In any such scenario that I can imagine, the changes or research findings would not happen overnight and could be addressed at the ten-year mark.

Should one of these changes happen overnight, it is still wise to proceed cautiously with changes. Making abrupt changes risks NAEP jumping into the political fray, as noted in Dr. Rosenberg’s paper (e.g., CCSS). Under the current ten-year process, because of the multiple committees and consensus building that process entails, changes to the framework avoid short-term fads that are ultimately not worth risking the historical trend line.

In sum, NAEP should not be adjusting to the newest trend of the day; doing so carries great risk and very little value. Should a true crisis occur, the policy allows for adjustments to address extenuating circumstances. Therefore, seeking “relevance” is insufficient to justify more frequent changes.

C. Adjusting to lessons learned

If there is a serious flaw or issue with the test design, test items, etc., responsible test makers should absolutely respond. However, in my former role, operationalizing the framework (e.g., writing items, data analysis) was different than creating the framework.

The Governing Board need not reconsider the overall NAEP framework more frequently than every ten years in order to allow for reasonable adjustments that respond to the framework directive approved by the Board. This is technical execution, not framework development.

Even if the Governing Board classified such technical adjustments as a reconsideration of the framework, the current policy allows for responses to extenuating circumstances. Committees need not default to meeting more frequently. Instead, these technical concerns should be addressed only if and when they arise.

D. Smaller changes over time to minimize the risk to the trend

Of all rationales offered, minimizing the possibility of breaking trend by incorporating smaller, more frequent updates to the NAEP framework is the most compelling. As a state policymaker during the transition to online NAEP, I had (and have) concerns about the maintenance of trends during that time period. As Dr. Rosenberg notes, retrospectively, it appears that a gradual shift to online would have been preferable.

When a framework is reconsidered each decade, part of the consideration should be – how different is the updated framework from the previous framework, and what is the safest way to transition while maintaining trend? If the answer is to gradually operationalize the updated framework over a two- or four-year period, for example, then such gradual changes to the format and assessment items should be adopted as part of the once-every-ten-years framework change adoption process.

I would urge a “decide once” principle here. Change the substance and determine the operational strategy in tandem. Such a practice allows for gradual adjustment without revisiting and re-questioning the framework every few years—a practice that assuredly risks the trend.

Conclusion

NAEP is valuable for comparability across states and for keeping states honest on what true proficiency looks like for children across America. NAEP is not political, and it needs to stay that way. NAEP is not an instructional driver, and it should not try to be.

For state leaders, NAEP results remain extremely important. But any questions about the validity or historical trend of the assessment dampens states’ ability to leverage NAEP results. To continue delivering valuable data to states, the Governing Board should prioritize maintaining trend lines and minimizing unnecessary noise above all else.

For better or worse, American education does not evolve quickly enough to necessitate changing the NAEP assessment framework more often than once every ten years.

To allow for best practice test making and to support trend maintenance, the Governing Board should allow for a gradual rollout of framework changes over several years, but any such gradual rollout should be pre-determined at the time of the once-every-ten-years framework update.

Allowing more frequent reconsideration for anything other than “extenuating circumstances” will almost certainly lead to more changes—and every such change increases the risk to the trend.

Keeping NAEP Relevant: Considerations for Smaller, More Frequent Changes to NAEP Assessment Frameworks¹

Stanley Rabinowitz, EdMetric LLC

September 2022

For five decades, information on what American students know and can do has been informed by the Nation’s Report Card—the National Assessment of Educational Progress (NAEP). The purpose of NAEP is “to improve the effectiveness of our nation’s schools by making objective information about student performance in selected learning areas available to policymakers at the national, state and local levels” (<https://nces.ed.gov/nationsreportcard/about/>). NAEP’s primary goals are to measure the status of the educational attainments of American students and to report changes and long-term trends of those attainments. NAEP is known as the gold standard in student assessments.

All assessments must evolve to meet changing societal and educational landscapes. Some changes are part of routine operations, others are more singular, such as moving to digital NAEP. The revision of NAEP frameworks has been more of the former—the National Assessment Governing Board (Governing Board) policy mandates a review for relevance of assessments and their frameworks at least once every 10 years. The Governing Board is conducting a review of this policy to determine whether and how more frequent revisions may impact NAEP’s primary goals: measuring achievement and trend. The goal of this review includes ensuring NAEP’s relevance in an ever-changing environment and maintaining its validity and value.

The intent of this paper is not to recommend a single pathway for the revision of NAEP frameworks but rather to identify and describe a series of policy, technical, communication, and fiscal considerations for the Governing Board to consider as it reviews its framework policy as part of the overall, ongoing evolution of NAEP.

Relevance in the NAEP Context

How might an assessment like NAEP define and maintain its relevance? To answer this, it is important to understand the role NAEP plays in the assessment ecosystem. Consider a continuum that runs from formative (classroom-based) assessment, to interim (periodic) assessments, to state summative assessments, to NAEP, through to international assessments such as PIRLS, PISA, and TIMSS.



Each of these assessments differs across two key factors germane to our discussion of NAEP relevance:

- instructional feedback

¹ A commissioned reaction to Rosenberg, S. Considerations for Smaller, More Frequent Changes to NAEP Assessment Frameworks (April, 2022)

- system accountability

As we move across the continuum, from formative on one end to international on the other, there is a clear decrease in the value of the assessments to guide day-to-day classroom instructional decisions. Moreover, the ends of the continuum (formative and international) do not play major roles in any formal accountability systems. In contrast, state annual summative assessments are featured prominently in these accountability systems by design and statute. Interim assessments are expected to inform midcourse corrections to instructional practice; several states are currently exploring their potential role as accountability indices as part of the Innovative Assessment Demonstration Authority (IADA) authorization.

With respect to the two key factors, I would argue that NAEP can support both to an important degree. On the instructional side, NAEP results “present a broad view of students’ knowledge, skills, and performance over time” (NAGB, 2022). High NAEP scores reflect high levels of academic achievement, much of which could be attributed to classroom supports and practice. (The opposite might be said on the lower end of the achievement scale.) And while NAEP results are not timely enough or fine-grained enough to support individualized instruction, they can be used to suggest, particularly through analysis of *item map* reports, content domains that are well taught or that might require more attention or a different pedagogical approach. To the extent that NAEP subgroup sampling is successful, results can also highlight and pinpoint systemic inequities at the national and state levels.

NAEP can also play a less formal but more independent accountability role relative to state summative assessments, helping to identify the extent to which state systems of standards and assessments are rigorous, technically sound, and equitable. Just as local educators can benefit from analyzing local academic achievement indices relative to state assessment results, so should state education officials value the check provided by NAEP on the progress towards the goal of universal proficiency as measured by the state assessment system. NAEP can also drive improvement of both local and state assessments via innovation of content (both frameworks and assessment items), administration procedures (especially with the move to digital), data analysis techniques, and creative and accessible reporting.

There is an important caveat to the benefits described above: *NAEP’s interpretive value is only as high as its real and perceived relevance*. As Rosenberg states, “concerns were raised about the validity of interpretations for NAEP results due to a perceived lack of alignment between content in some of the NAEP frameworks and state content standards.” Both instructional feedback and accountability fairness require a sufficient degree of content alignment between programs. To the extent that NAEP frameworks and state content standards differ, many students will not be exposed to NAEP assessment content, creating Opportunity to Learn (OTL) challenges, resulting in a decrease of the validity and value of NAEP results.

I am not arguing that NAEP should simply mirror state standards and assessments. Historically, NAEP has been the driver of improvements in both content and assessment practices, and it is crucial it maintains this status. This paper focuses specifically on the role NAEP frameworks can play in maintaining NAEP’s relevance and how changes in the processes by which NAEP

frameworks are revised may both maintain NAEP’s leadership role and enhance its relevance and reputation among its stakeholders, educators, and the general public.

Developing and Refining Frameworks

From my experience across a range of roles and contexts², the success of content—whether it is called frameworks, content standards, state/national curricula—in driving assessment is based on a number of factors related to its development and revision. Getting them right is both art and science, especially understanding the most vexing questions: *which content is taught, how much is enough (or, more likely, too much), and who gets to decide?* Equally important is understanding how and when to update, revise, and/or replace existing content standards.

Common content-related concerns include:

- there is too much content to be taught and assessed,
- the content is too vague or too complex to support instruction and assessment development,
- the content is not connected to real-world needs and expectations,
- some content does not translate well into the mode of assessment, especially higher order thinking skills and non-traditional content such as speaking and listening (moving to digital has helped to address this concern),
- some content is inaccessible to traditionally underserved student populations such as Black, Indigenous, and People of Color (BIPOC), Students with Disabilities (SWDs), English Learners (ELs), and
- there may be a lack of support (e.g., professional development) for teachers to properly teach the content.

If content is not properly developed, revised, validated, communicated, and supported, teachers may not teach the content, assessment items may not align to the content, and large numbers of students may not have access to the content, leading to inequities, invalidity, and irrelevance.

The process for framework refinement/revision is especially complex with many potential missteps and lost opportunities. As noted above, Governing Board policy mandates a review for relevance of assessments and their frameworks at least once every 10 years. Most state programs have a set (somewhat arbitrary) time for review of content standards, typically about seven or so years. Changing the frameworks too soon may upset implementation activities and result in educators who are caught between understanding the previous content while moving to support the new content. Not changing the frameworks soon enough can lead to stale and obsolete

² I have served as a state assessment director (New Jersey), managed a national assessment program (Australia), served as PMO director for the development of the Smarter Balanced Assessment Consortium, and performed several curricular and assessment development and validation functions including TAC member in more than a dozen programs, content standards development and review (e.g., CCSS, NGSS), item development, equating, alignment studies, and standard setting for many state and national assessments (including NAEP). The comments in this paper are informed by this experience.

content, which applies more in some domains than others. Changing too often is costly both monetarily and in system disruption, affecting local curriculum, pedagogy, and teacher lesson plans; assessment content/alignment; existing performance standards; accountability indices (both growth and status); and trend lines. Under these circumstances, it is imperative that the Governing Board carefully examine its framework development and revision policy and processes, and not institute any changes without the evaluation of the value, burden, and unintended consequences of any new approach, including as they relate to the potential of a move to a more incremental framework revision strategy.

The remainder of this paper focuses on several considerations that might impact when and how a NAEP framework may be revised. This discussion necessarily involves balancing complex trade-offs and competing priorities. These various considerations encompass the issues and questions raised in *Considerations for Smaller, More Frequent Changes to NAEP Assessment Frameworks* (Rosenberg, 2022), which includes discussion of “potential rationales for more frequent, gradual changes to the NAEP assessment frameworks” and “understanding possible solutions.”

This reexamination is especially timely given the current NAEP and larger societal context: the move to digital NAEP with its array of benefits and technical/logistical challenges; understanding and addressing the impact of COVID-19 on student achievement and well-being; and advancing our shared mission for equity and inclusion in all aspects of the education experience of students, teachers, policy makers, and stakeholders.

Considerations, Trade Offs, and Competing Priorities

- *What role does NAEP want to be play?* NAEP’s objective to be the gold standard in assessment may be interpreted to mean that evolutions in state and international frameworks and assessments should be monitored but not typically trigger any response. On the other hand, if the relationship and consistency between state and NAEP results is important (e.g., for OTL or convergent validity expectations), the Governing Board may need to pay greater attention to trends in state content standards and assessment practices and modify its frameworks to ameliorate any large differences.
- *What events require a new/revised framework?* There may be some factors that require substantial review and potential significant changes to a framework such as: moving to digital NAEP with the possibility of assessing content formerly not available on a paper assessment; the introduction or rejection of the Common Core State Standards in states and its impact on state content standards; and, obviously, adding a new content area to the NAEP program. Defining the characteristics of such events and anticipating future occurrences will facilitate planning and budgeting.
- *How does NAEP operationalize the reality that “Nothing is ever perfect”?* We need to accept that all frameworks can be improved—the Governing Board needs to develop a protocol to define “good enough” for now and when good enough is no longer “good enough.” Any change has both the potential for improvement and disruption. As

Rosenberg (2022) warns: “It is possible that revisiting frameworks too frequently...might actually encourage unnecessary changes” or reopen acrimonious debates. Each framework review creates the opportunity for controversial issues considered resolved to become “unresolved,” increasing rancor and uncertainty in an already highly politicized environment.

- *Are there inherent differences within and across content areas?* Some content areas are more likely to be impacted by external factors than others. For example, the world of IT is constantly evolving while reading and mathematics may be relatively more stable. Thus, the Technology and Engineering Literacy (TEL) Framework may require more ongoing review and revision to stay relevant. Aspects of the Science Framework may be impacted by new discoveries or procedures or modes of assessment (e.g., the opportunities for performance items created by digital administrations). The Civics Framework needs to be nimble enough to meet the changing social environment, for example, the timely commitment to Diversity, Equity, and Inclusion (DEI). While the centrality of “*standardization*” is deeply embedded in large-scale assessment, the Governing Board needs to determine when strict adherence to standardization of process potentially exacerbates invalidity and inequity by decreasing access and relevance.
- *What constitutes a “change?”* Change, somewhat like beauty, may be in the eye of the beholder. We need to build consensus defining a minor versus a major change as well as how to measure the cumulative impact of “minor” changes over time. Defining and measuring change in the NAEP framework context has technical, fiscal, and communication challenges. Key to this process is the determination of when a framework has evolved to such an extent that the item bank is no longer in alignment and the maintenance of trends is suspect. It remains an empirical question of if/when small changes over time impact trendlines relative to wholesale framework revisions. Bridge studies are part of the solution to the technical questions, but they alone will not fully resolve any potential cumulative impacts of even minor changes.
- *When does devotion to trend work against the interests of NAEP?* The supremacy of the maintenance of trend in NAEP planning cannot be overstated. NAEP’s enabling statute states the purpose of state assessments involves: “... fair and accurate measurement of student academic achievement and reporting of trends in such achievement.” However, it is possible that this devotion may work against the modernization and relevance of the NAEP program. The Governing Board needs a protocol for deciding the conditions where trend is threatened (e.g., when do a series of minor revisions grow into a major revision?) and when preservation of trend is no longer desired or possible. Since content is such a large component of trend, the review and revision of NAEP frameworks must be a key factor in this trend-review protocol.
- *What is the impact of framework revisions on NAEP’s validity and equity?* The Governing Board needs to examine whether the existing framework process and resulting

content might be exacerbating sources of invalidity and introducing bias and inequity. For example, if state content standards differ greatly from the NAEP frameworks, traditionally under-resourced student populations are likely to be disparately impacted on NAEP assessments relative to more affluent cohorts (who may have greater access to rich content outside of school). Inequities may manifest differentially across content areas (e.g., access to state-of-the-art IT resources may differ by community more than access to reading or civics resources).

- *Is the framework process (the NAEP “way”) healthy?* The Governing Board needs to examine whether the current process for the development and revision of frameworks meets NAEP’s needs. According to Rosenberg (2022): “it currently takes NCES approximately 4-5 years to implement any changes to the framework in the assessment, due to the need for item development, reviews, cognitive labs, pilot testing, etc.” While the desire for inclusivity is commendable and essential, the current process may result in over-engineering, creating barriers to innovation and relevance. A more streamlined approach, perhaps with simultaneous reviews, may support the dual goals of inclusion and efficiency. It is also possible that a model of more frequent, less dramatic revisions to frameworks may decrease the pressure to get everything “right” the first time, giving NAEP more leeway to move ahead with conditional approval across interest groups and stakeholders. Each administration may be viewed as a learning experience for the Governing Board, with procedures in place to determine whether the alignment of content to the framework is working as intended and how the overall linkage can be improved. As Rosenberg (2022) states, the Governing Board will need to determine “what resources and/or structural changes may be needed to implement new processes.”
- *How can NAEP avoid fads (or hoopla)?* Though relevance is a desirable goal, NAEP needs to avoid being a player (or casualty) in the latest “culture war.” The challenge here, as “veterans” of the phonics, calculator, Common Core, civics/CRT “wars” can attest, one person’s (or stakeholder’s) requisite is another’s “passing fancy.” The Governing Board needs to distinguish between the latest “fad” and a true positive enhancement and be able to respond quickly once that determination is made.
- *How can NAEP balance dollar costs versus opportunity costs?* Assessment budgets are limited and often stretched thin. Prudent framework planning involves known time frames for framework revision with sufficient reserves and processes to address unplanned “unknowns,” be they positive in nature (e.g., a greater commitment to equity and inclusion) or negative (e.g., the impact of pandemics). Revising frameworks to meet these opportunities will almost certainly require a nimbler process than currently exists. As Rosenberg (2022) suggests, smaller framework revisions over time (as opposed to re-creations) may be cost effective, allowing new content to be developed and implemented gradually with potential minimal impact on trend.

- *What image does the Governing Board wish to project: Microsoft versus Apple?* Finally, a useful metaphor may be for the Governing Board to decide what is the image it wishes to convey to its various stakeholders and constituencies. Is it “steady” or “trendy”? There is no single, obvious right answer—either image can produce high-quality, reliable, and valid products (in NAEP’s case, frameworks, assessments, reports). The process for developing and revising frameworks will impact this image. The Governing Board might find it beneficial to be able to project either image depending on the circumstances. “Big-bang” changes may be necessary to meet some challenges (e.g., Digital NAEP, COVID-19) while smaller, incremental revisions might be the course the Governing Board charts in more ordinary times. Different circumstances may benefit from one image rather than its alternative. Rarely can an organization embody both simultaneously, or nimbly switch from one to the other.

Final Thoughts

In conclusion, the review of the NAEP framework development and revision policy and processes is timely, necessary, and complex. The Governing Board should begin by determining whether its goal is to update current practice or create a new model. The debate should focus not just on the pros and cons of various approaches, but on the likelihood that unintended, unanticipated consequences will compete with expected enhancements. NAEP has built and earned its reputation for integrity and trust across its many stakeholders. Its place as the “gold standard” of assessment will be challenged by the relevance of its frameworks, a concern that is both substantive and perception driven. Careful attention to the issues raised by Rosenberg (2022) and the considerations discussed in this paper (and the others in this series) create a pathway for the Governing Board to continue to achieve its singular critical mission in these complex, challenging times.

Feedback and Recommendations on NAEP Assessment Development Framework Update

Ada Woo, Ascend Learning

September 2022

Background

Historically, a NAEP assessment framework remained unchanged for a minimum of ten years. The purpose was to ensure that assessment content would be stable and trends in student achievements could be measured longitudinally. In March 2018, the National Assessment Governing Board policy on Framework Development was revised to enable more frequent framework updates from a cadence of no more than once every ten years to at least once every ten years.

In April 2022, Governing Board staff further defined the problem and identified potential solutions of more frequent, incremental updates to the NAEP assessment frameworks (Rosenberg, 2022). Board staff summarized that a shift to gradual framework updates could (1) enable the program to maintain and track performance trends, (2) allow changes to be implemented in the field more quickly, (3) allow assessment developers to make revisions after the first administration of a new framework, and (4) lower the cost of framework development. In the assessment industry, the certification and licensure testing sector faces many of these challenges. The current paper will discuss several relevant examples in licensure testing and key takeaways for the NAEP program.

Continuously Assess the Relevance of Assessment Frameworks

According to the National Assessment Governing Board Assessment Framework Development Policy Statement, the framework “shall determine the extent of the domain and the scope of the construct to be measured” in the NAEP assessments (National Assessment Governing Board, 2022). This included *what* should be measured, *how* to measure the construct, and *how much* of the construct should be demonstrated at each achievement level. Analogous to this in certification and licensure testing, examination programs conduct practice (or job) analyses to ascertain the knowledge, skills, and abilities required to perform at a competent level in a particular profession (e.g., what and how much would an examinee need to know in order to be a safe and competent physician?). Historically, professions that have licensure or certification requirements conduct practice analyses once every three to ten years. In the past decade, technological advances and societal changes have accelerated changes in practice for many professionals. In response, several licensure and certification testing programs have adopted a continuous practice analysis model in addition to conducting the once few years large scale studies. Data obtained from these continuous practice analyses enabled better monitoring of changes in practice and emerging trends.

The Federation of State Boards of Physical Therapy (FSBPT), developer of the National Physical Therapy Examination (NPTE), conducts a formal analysis of practice once every five

years. In this process, FSBPT identified the work requirements for entry-level physical therapy professionals and used the data to guide the development of test specifications. To keep pace with the evolution of physical therapy practice, FSBPT adopted a new survey methodology in 2018 and began analyzing practice annually. The new surveys are conducted in addition to the once every five years studies. According to the FSBPT, this continuous practice analysis methodology allows the organization to “monitor ongoing and emerging trends in entry-level requirements and to respond quickly to changes in the profession that necessitate adjustments to the licensure examinations” (Federation of State Boards of Physical Therapy, 2022).

FSBPT summarized the process and results of its continuous practice analysis studies in its most recent report (Rogers & Caramagno, 2020). In February 2020, FSBPT gathered input on existing list of physical therapy work activities (WAs) and knowledge and skill requirements (KSRs) from its Examination Committee. A survey was then sent to a large sample of physical therapists and physical therapy assistants to learn how important these professionals view each WAs and KSRs is to their practice. After survey data were analyzed, FSBPT convened a one-day meeting with its research vendor and Examination Committee Chairs to review the results and evaluate next steps.

The 2020 survey revealed that physical therapy practice characteristics are largely consistent with those in 2018 and 2019. Majority of respondents indicated no change in practice status on most WAs and KSRs; they recommended that the NPTE assessment framework to remain unchanged. Several emerging areas of physical therapy practice were identified from the survey data (e.g., laser light therapy, LED therapy, and telemedicine). The FSBPT indicated that it will continue to track WA and KSR trends overtime and use the data to inform updates of its assessment framework.

Emerging Knowledge and Expansion of Content Domain

As technology advances and environment changes, relevant knowledge in a domain may expand or evolve over time. This is true for both K-12 education and certification/licensure professions. In most subject matters, there are knowledge and skills considered core or foundational upon which new knowledge was built. These core knowledge and skills generally remain stable over time, allowing for continuity and the ability to track learner performance longitudinally. In this section, the methodologies of how two licensure organizations handled domain changes are highlighted.

The National Council of State Boards of Nursing (NCSBN), developer of the National Council Nurse Licensure Exam (NCLEX), is expanding its licensure examination to include the assessment of nursing clinical judgement and decision making. This process began in 2013, with a strategic practice analysis focused on emerging trends. Once the need to expand the NCLEX construct (i.e., entry-level nursing competency) to include clinical judgement and decision making was identified, additional research was conducted to determine how these new knowledge and skills would be measured on the NCLEX. In 2017, NCSBN began field trials of new item formats and research scoring methodologies. Based on its research findings, NCSBN

will add several new item formats (case studies/testlets and technology enhanced single items) to the NCLEX for the purpose of assessing clinical judgement and decision making. The updated NCLEX assessment framework will include an additional three case studies (18 items) on each exam and a range of additional clinical judgement single items. Majority of the exam will be consisted of traditional (i.e., non-clinical judgement) items. The NCLEX is a variable length computer adaptive test (CAT), the range of items on a particular exam will be dependent on the test-taker's ability. NCSBN will launch its new NCLEX assessment framework in 2023 (Betts, Muntean, Kim, & Kao, 2021).

With the updated assessment framework, the NCLEX will change its scoring model to partial credit (as opposed to the current dichotomous/correct-incorrect scoring). Major categories of the NCLEX test blueprint and much of the CAT item selection criteria will remain unchanged. The updated score report will provide performance feedback in the same major test blueprint categories as in prior NCLEX assessment frameworks, with an additional section on clinical judgement diagnostics. Thus, preserving the ability for the testing programs and nursing educators to track longitudinal trends across the test blueprint categories (Betts, Muntean, Kim, & Kao, 2021).

The Certified Public Accountant (CPA) Exam is undergoing similar changes. The Association of International Certified Professional Accountants (AICPA), developer of the Uniform CPA Exam, is in the process of updating and expanding its assessment framework. The process began with its 2020 practice analysis that informed the organization of changes in CPA practices. As a result, the CPA Exam will be realigned to include a core component and three disciplines. Examinees will be required to take the core component and select one discipline component. The core section included knowledge of accounting, auditing, and taxation/regulation. The three disciplines are: (1) tax compliance and planning, (2) business analysis and reporting, and (3) information systems and controls. Data and technology concepts will be assessed in all core and discipline exam sections. The AICPA is currently requesting comments from stakeholders. The comment period closes on September 30, 2022. The AICPA plans to launch the new assessment framework in 2024 (Association of International Certified Professional Accountants, 2022).

Develop a Consistent Assessment Framework Across Multiple Subject Areas

Many testing programs encompass a wide range of subject areas and stakeholder demographic characteristics. The NAEP program, for example, assesses student knowledge in mathematics, reading, science, and several other subjects. In the licensure and certification testing sector, the National Board for Professional Teaching Standards (NBPTS), issuer of the National Board Certified Teacher (NBCT) certification, develops assessments in over 30 certificate areas along learner ages and subjects (e.g., Early Adolescence English Language Arts and Middle Childhood Generalist). While specific standards are developed for each certificate areas, the NBPTS follows the same assessment framework for all certifications.

All NBPTS certificate assessments begin with the Five Core Propositions, developed based on What Teachers Should Know and Be Able to Do (National Board for Professional Teaching Standards, 2016). The five core propositions articulated the vision for accomplished teaching, the construct on which the NBPTS certifications are based. The five core propositions are (1) Teachers are committed to students and their learning, (2) teachers know the subject they teach and how to teach those subjects to students, (3) teachers are responsible for managing and monitoring students learning, (4) teachers think systematically about their practice and learn from experience, and (5) teachers are members of the learning community. From these five core propositions, NBPTS developed standards specific to each certificate area.

The NBPTS employs one assessment framework across all its certificate areas. All certification areas are based on the same five underlying core propositions. NBPTS also uses the same assessment formats and scoring design across all certifications. All certifications contain four components: content knowledge, differentiation in instructions, teaching practice and learning environment, and effective and reflective practitioner. Each component is assessed with different formats, ranging from multiple choice questions to video portfolios. The same scoring design is used across all certification areas, allowing for trend monitoring and comparisons both within a certification area and across multiple areas.

Recommendations for the NAEP Assessment Framework Development

As discussed in the sections above, examples of more frequent and incremental framework updates can be found among several high-stakes licensure and certification programs. The first step of assessment framework updates generally involved a stakeholder survey or practice analysis study to ascertain whether the content domain of interest has changed. Once changes in content domain are identified and an update of the assessment framework is deemed necessary, the testing organization should decide on how much (if any) the current and new frameworks will overlap. As seen in the NCLEX (nursing) and Uniform CPA Exam examples, one strategy is to retain core components of the framework that link the old and the new. Doing so will allow the testing program to track test-taker performance across time and monitor trends. This will also allow the new score report to maintain some level of consistency with the old, promoting ease of interpretation among stakeholder groups.

One important point to consider is the level of details the testing organization will track or provide in its score reports. Unlike educational tests, licensure and certification exams are created to generate one decision (i.e., pass/fail). Most licensure and certification programs do not provide detailed score reports or diagnostic feedback to their examinees. Many licensure and certification testing programs do not provide score reports to passing examinees at all. The need for licensure and certification programs to maintain detailed or granular test-taker performance tracking may be lower than that for the NAEP program.

Another hurdle of implementing more frequent and incremental assessment framework updates is to maintain a sufficient item inventory that meets test specifications. The NAEP program is in a unique situation where two organizations manage the assessment framework

creation and item development processes. The success of more frequent assessment framework updates will depend on the collaboration and partnership between the Governing Board and NCES. Parallel to the assessment framework development process, the team should review item inventory and conduct item bank gap analysis. This will provide the team with a realistic sense of how frequently and to what extent the assessment framework may be updated, as well as the cost of each update.

Finally, the Governing Board should review the policies around development and review of the NAEP assessment frameworks. As observed in the NBCT certifications, the NBPTS opted to apply a standardized assessment framework to all its certificate areas. This standardization streamlined the development, review, and stakeholder engagement processes. With changes to more frequent assessment framework updates, the Governing Board should evaluate and update the NAEP development process to ensure that time dedicated to committee and stakeholder reviews/approvals is manageable and supports a more rapid framework update cadence.

References

- Association of International Certified Professional Accountants. (2022). *Maintaining the Relevance of the Uniform CPA Examination - Aligning the Exam with the CPA Evolution Licensure Model*. Durham, NC: AICPA.
- Betts, J., Muntean, W., Kim, D., & Kao, S.-c. (2021). *Next Generation NCLEX News*. Chicago, IL: NCSBN.
- Federation of State Boards of Physical Therapy. (2022, September 15). *Ensuring Validity: Understanding the Process of Practice Analysis*. Retrieved from The Federation of State Boards of Physical Therapy: <https://www.fsbpt.org/Free-Resources/NPTE-Development/Ensuring-Validity>
- National Assessment Governing Board. (2022). *Assessment Framework Development Policy Statement*. Washington, DC: NAGB.
- National Board for Professional Teaching Standards. (2016). *What Teachers Should Know and Be Able to Do*. Arlington, VA: NBPTS.
- Rogers, A. P., & Caramagno, J. P. (2020). *Analysis of Practice for the Physical Therapy Profession: Report Memo 2020*. Alexandria, VA: HumRRO.
- Rosenberg, S. (2022). *Considerations for Smaller, More Frequent Changes to NAEP Assessment Frameworks*. Washington, DC: NAGB.

NAEP Science as a Context to Consider Options for NAEP Framework Revision

Alicia C. Alonzo, Michigan State University

September 2022

In this paper, I take up the question of how NAEP frameworks should be revised to reflect shifts in their respective fields. In doing so, I draw on my perspective as a science educator, using consideration of shifts in science education and in the NAEP Science Framework over the past 25 years to think more broadly about opportunities and tensions for the NAEP assessment program as a whole.

Context: NAEP Science Framework & Shifts in Science Education

The current framework for NAEP Science was approved in 2005 and formally introduced in 2008 (NAGB, 2008) as the [2009 NAEP Science Framework](#). It follows an [assessment framework](#) used for the 1996, 2000, and 2005 NAEP science assessments (NAGB, 2005). The 2009 NAEP Science Framework contained enough substantial differences from its [predecessor](#) to warrant a break from the 1996-2005 trend ([National Center for Educational Statistics \[NCES\], 2021](#)). The new trend (2009-2024) reflects the use of the 2009 framework in 2015 ([NAGB, 2014](#)) and 2019 ([NAGB, 2019](#)), as well as its future use in 2024 (Grade 8 only).

A challenge for NAEP is that its work to accommodate shifts in consensus views of student learning occur concurrently with further evolution of these views. For example, the break in NAEP Science Trend in 2009 represents, among other considerations, changes in response to the first set of national science standards: [National Science Education Standards](#) (NSES; National Research Council [NRC], 1996) and [Benchmarks for Scientific Literacy](#) (*Benchmarks*; American Association for the Advancement of Science, 1993) ([NAGB, 2008](#)). However, shortly after the establishment of this new trend, a new consensus view of science learning emerged in the form of [A Framework for K-12 Science Education](#) (NRC, 2012) and the [Next Generation Science Standards](#) (NGSS; NGSS Lead States, 2013). A similar overlap can be seen in the establishment of a new NAEP trend in 1996 and release of national standards in the same year.¹ As a result, since at least 1996, with the exception of 2009, NAEP Science has reported student achievement relative to consensus that had since been replaced (at least aspirationally) by substantively different expectations for students' learning. Although shifts in practice occur more slowly, NAEP Science loses some relevance when a 2024 assessment reflects 1996 standards.

A recent report ([Pellegrino, 2021](#)) found that “neither framework [NAEP Science or NAEP Technology and Engineering Literacy] is reflective of the more dramatic shifts found in the NRC

¹ Of course, it is likely that new views of science learning were discussed in the development of both the assessment framework and the standards, perhaps due to overlap in committee membership and likely due to consensus views being shared widely across the field. However, the NAEP assessment could not be purely aspirational and, thus, reflected an older view of science learning than was represented in the NSES. This is evidenced by the need to revise the 1996 assessment framework in 2009 to reflect reforms prompted by the NSES.

framework and NGSS” (p. 15). This is significant, given similarities between NAEP Science and the NGSS in terms of the inclusion of science practices and “the idea of performance expectations that involve the intersection of content and practice” (p. 15). Such similarities may lead to varying interpretations of the NAEP framework, with the potential for shift over time towards interpretations more consistent with the NGSS and less consistent with the intended NAEP framework. This may occur both naturally, as Standing Committee members bring evolving perspectives about key terms or about what items “should” look like, and more formally, e.g., as item development includes additional attention to NGSS-like practices by identifying “knowledge, skills, and abilities” (KSAs), as KSAs have stronger alignment to the NGSS than the four science practices laid out in the 2009 Science Framework (NAGB, 2008).

A new assessment framework—based on updates to the 2009 NAEP Science Framework—is currently being considered for the 2028 NAEP Science Assessment ([NAGB, 2022b](#)), based in part on efforts to explore differences between NAEP and the NGSS (Pellegrino, 2021). In the sections below, I use the opportunities and tensions inherent in revising the NAEP Science Framework in response to shifts in the field to consider broader questions of NAEP framework revision. In doing so, I respond to the four rationales for changing the current process (NAGB, 2022a) identified in Sharyn Rosenberg’s 2022 report to the NAEP Assessment Development Committee: Increasing relevance, maintaining trend, responding to “lessons learned,” and cost.

Nature of Shifts Necessitating New Assessment Frameworks

Multiple considerations may prompt a given framework revision. For example, two important factors in the development of the 2009 NAEP Science Framework were the introduction of new standards and availability of innovative methods of assessment (NAGB, 2008).

Shifts in Consensus & Resulting Practice

Several types of shifts in the field are relevant to changes in students’ achievement in science (or other subjects):

- 1) Gradual shifts that occur over long periods of time within the field as consensus is formed—e.g., the addition of new research or changes in an ongoing debate might cause a gradual shift in perspectives on student learning
- 2) Seismic shifts (e.g., new standards) that represent a consensus view that differs substantially from an earlier consensus view—i.e., the culmination of the gradual shifts taking place “behind the scenes”
- 3) Gradual shifts in the implementation of a new consensus view, as curriculum, assessment, and teacher supports are developed and begin to reach science classrooms around the country.

NAEP frameworks are intended to reflect a balance between current (and future) standards (“the nation’s future needs and desirable levels of achievement”) and “current curriculum and

instruction” (NAGB, 2022a, p. 2). A NAEP assessment framework may lack relevance due to misalignment with new standards (reflected in seismic shifts that result from gradual shifts “behind the scenes”) and/or with current practice (reflected in gradual shifts towards implementation of new standards).

Especially in the transition to new standards, when ambitious goals for student learning have been set but infrastructure (e.g., curricula, teacher professional development) are not yet available to fully support that learning, the tension between current standards and current practice may be large. Framework revision necessarily requires choices about whether current standards or current practice are prioritized when the two are in conflict. These choices, in turn, determine whether revisions to a given assessment framework should be seismic or gradual. A more standards-oriented revision would require a larger revision. Indeed, the necessity of new standards signals a large shift, quite possibly requiring a new trend. In contrast, a revision more attuned to what is happening in the nation’s classrooms (whether reflective of new standards or not) may result in small changes that add up to larger shifts over time.

Assessment Innovations

An assessment framework specifies both what content should be assessed and how it should be assessed. Innovations in assessment may suggest new ways of assessing an existing framework and/or new possibilities for what can be assessed and, thus, what content can be included in a new framework. While neither require changes to the framework, the latter may open up possibilities for an assessment to more fully represent the domain being assessed.

For example, new item types have permitted small changes to the role of science practices in TIMSS science assessment frameworks. In contrast to PISA, TIMSS assessment frameworks are closely tied to the curricula of the participating countries. Although recent TIMSS assessment frameworks (including the TIMSS 2023 Science Framework) have included “minor updates to reflect countries’ evolving science curricula, frameworks, and learning goals” ([Mullis et al., 2021](#), p. 19), the need to obtain consensus across 64 participating countries ([NCES, n. d.](#)) means that change is quite slow. However, in 2015, these updates included the addition of science practices, in recognition of “increasing emphasis... on science practices and science inquiry in many countries” (Mullis et al., 2021, p. 31). Moving from the 2015 ([Mullis & Martin, 2013](#)) to the 2019 ([Mullis & Martin, 2017](#)) to the 2023 (Mullis et al., 2021) frameworks, the role of the science practices has not changed significantly; however, the 2023 TIMSS Assessment Framework specifies that the science practices will be assessed using a new item type (Problem-Solving and Inquiry Tasks). The new item type allows students to engage with much more authentic forms of science practices and clarifies the role of the practices in the framework.

Options for Responding to Shifts in the Field

In this section, I consider three options for ways that NAEP assessment frameworks could be updated to reflect current standards and/or current practice and assessment innovations: current

practice ([NAGB, 2008a](#)); the proposal to conduct more frequent, smaller revisions to NAEP frameworks; and an alternative that combines the two (Rosenberg, 2022).

Current practice: Less frequent, larger shifts in assessment frameworks

This is current practice in NAEP, with shifts in assessment frameworks potentially occurring far after relevant shifts in the field have occurred. For example, an updated 2028 Science Assessment Framework would be the first to reflect a major shift in the field denoted by the introduction of the *Framework* and NGSS 15-16 years prior.

When should less frequent, larger shifts occur?

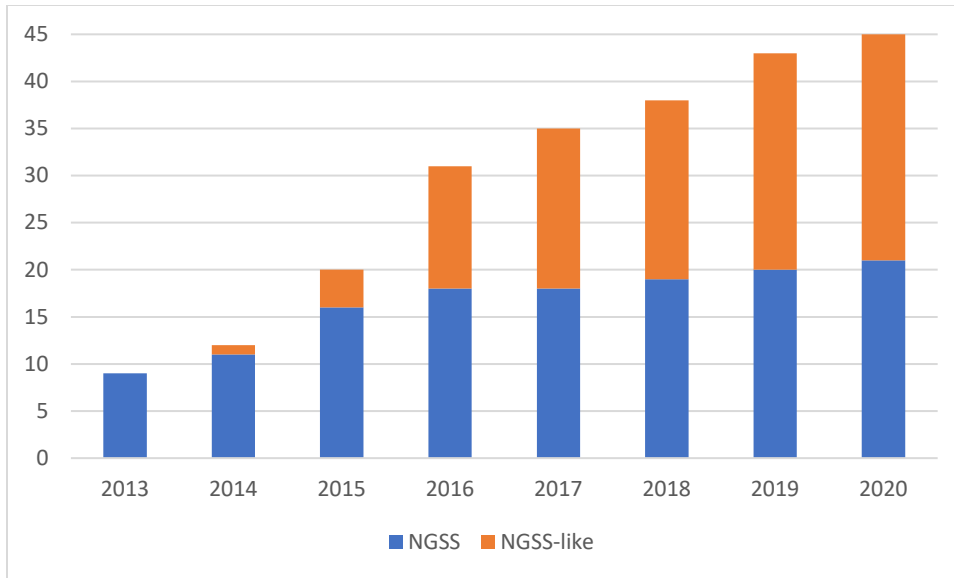
In the extreme, if alignment with current (and potential alignment with future) standards is prioritized, development of a new assessment framework would begin as soon as a “disruption” (Rosenberg, 2022, p. 7), e.g., introduction of new standards, occurs. The resulting assessment would reflect new goals for learning (what is expected of students in the future) and—over time—could show the extent to which students are reaching those expectations.² This attention to ambitious goals for student learning can be seen in the [PISA 2006 Scientific Literacy Framework](#) (Organisation for Economic Co-operation and Development [OECD], 2006). The knowledge and skills assessed in PISA are defined “in terms of what skills are deemed to be essential for future life.” This forward-looking perspective is responsive to shifting conceptualizations of scientific literacy, such that the assessment “is not constrained by the common denominator of what has been specifically taught in the schools of participating countries” (p. 11).

However, practical considerations may require “some lag following a major disruption in a field” (Rosenberg, 2022, p. 7). First, although standards such as the *NSES* and *NGSS* are developed at a national level, decisions about standards adoption are made through political processes within states, and it may not be immediately clear whether new standards will be widely adopted. For example, a snapshot in 2015 would have reflected only 19 states and the District of Columbia with NGSS or NGSS-like standards, and the last states did not adopt standards influenced by the NGSS for five more years. Table 1 shows the changing landscape of NGSS adoption from 2013 to 2020 and illustrates the challenge of knowing, shortly after standards are released, how new standards will be taken up.

Table 1: Number of States (and the District of Columbia) with Adoption of NGSS or NGSS-Like Standards by Year³

² This goes against current NAEP practice, in which current curriculum and instruction are also considered; this is intended as one extreme of a continuum.

³ This chart is constructed using information from Citizens for Objective Public Education (COPE; 2020) and National Science Teaching Association (NSTA; n. d.) The former contains the most detailed information I could find for the years for adoption of state standards aligned with the NGSS; however, due to the policy advocacy of this organization, I checked the information in this document with information from the NSTA site.



Second, major shifts in expectations for student learning are likely to be accompanied by shifts in the way that learning is assessed. For example, in science, the three-part structure of the *NGSS*, in which disciplinary core ideas, scientific and engineering practices, and crosscutting concepts are combined to form performance expectations, a “primary challenge for assessment design” is “to find a way to capture... students’ developing proficiency along the intertwined dimensions” ([NRC, 2014](#), p. 28). Challenges such as this take time to resolve, and moving too quickly to adapt an assessment to new standards may result in a narrowed framework because technologies to assess the new standards are not yet available. By waiting for some time after new standards have been introduced, the burden of NAEP responding to these challenges would be lessened, as others’ solutions might be available. For example, there are now resources available from the [Science Assessment Item Collaborative](#), which includes an assessment framework ([Council of Chief State School Officers, 2022](#)). At the same time, if NAEP is to remain the “gold standard” of assessment in the nation (NCES, 2010, p. 2), perhaps there is an argument to be made for leading rather than following state assessments.

What should trigger large shifts in assessment frameworks?

In the absence of a clear demarcation of shifts in the field (i.e., introduction of new standards), decisions must be made as to when is the “right time” for a big shift to be incorporated into NAEP. Is there a “tipping point” for the number of states (or percentage of students living in states) that have adopted and/or implemented new standards? Does it matter the kinds of difference between state standards and the *NGSS*?

How can large framework shifts be implemented more quickly?

If the balance between current (and future) standards and current practices tilts towards the former, framework revision would occur as quickly as possible. *Reducing the lag between the introduction of standards and the start of a framework revision process* would reduce length of

time that a NAEP assessment is seemingly out-of-step with current standard—while also increasing the risk that anticipated changes either do not occur or occur differently from anticipated. However, while policies (such as the NGSS) may appear to reflect sudden shifts, they reflect consensus that has been building in the field over time, typically through processes that represent similar stakeholder input as is sought in NAEP (NGSS Lead States, 2013). While such consensus does not include state-level decision-makers in all states, involvement of many states and a wide range of stakeholders both reduces the probability of new standards being quickly replaced by new ideas and offers possibilities for a *slightly reduced framework development process by relying more on the consensus processes used to develop standards*.

Time savings may be small, however: When the 2009 NAEP Science Assessment Framework was developed, there was broad consensus about the NSES, yet there was still significant work to move from standards to an assessment framework, which—among other factors—reflected decisions about what content should be included or excluded. This is further complicated by variation in the extent to which states have adopted new standards. For example, 45 states and the District of Columbia (NSTA, n.d.) have adopted standards aligned with the NGSS. Careful attention needs to be paid to the state standards in the six states with standards not aligned with the NGSS (e.g., [Dickinson et al., 2021](#)), especially with respect to differences between the NGSS and the current assessment framework (e.g., [Neidorf et al., 2016](#)), as potential targets for revision. Even if there is broad consensus about new standards, if all states have not adopted them, significant negotiations may be needed to reflect both “greater convergence among state standards” and “science education in states that diverge from” the NGSS (NAGB, 2008b, p. 2).

Further efficiencies might be possible if revisions were limited to those needed to reflect changes in the new standards. When large changes are needed, there may be a tendency to make lots of smaller changes as well—i.e., to completely overhaul the framework, rather than making only needed revisions to the old one. If debate about issues that have already been settled in a NAEP framework, and that have not been changed by shifting consensus in the field, are avoided, this could save time within the process of framework development.

More frequent, smaller shifts in assessment frameworks

What kinds of changes in the field suggest more frequent, smaller shifts in assessment frameworks?

As noted above, seismic shifts in a field, such as new standards, are outward manifestations of more gradual changes that have occurred over time through consensus-building and may start a process of gradual change in practice as standards and other policies are adopted and implemented over time. Especially given the complexity of new expectations for student learning, implementation may take a significant amount of time to occur, such that changes in response to new standards may take place across several NAEP administrations. For example, the [NRC \(2015\)](#) estimates that district and school implementation of the NGSS will take 5-10 years. More frequent, smaller shifts in assessment frameworks would place NAEP assessments

in a better position to capture current practice (and shifts in current practice) as stakeholders adjust to expectations for student learning that have been introduced more suddenly. Smaller shifts in assessment frameworks would also permit advances in assessment technology to influence assessment frameworks more rapidly and, thus, to ensure that NAEP assessments are best positioned to assess valued outcomes (even if those outcomes do not change significantly between frameworks). In addition, with small shifts in the framework between administrations trend could be maintained even as current practice changes significantly on a longer time scale (Rosenberg, 2022).

For the reasons just described, a more frequent approach with smaller shifts in assessment frameworks is well-suited to changes in practice. It seems less appropriate for changes in the field's view of learning. New standards are developed when the field coalesces around a consensus view. Prior to that, smaller shifts occur as ideas are debated within the field. Thus, changes to assessment frameworks in such debate periods may reflect transitory ideas that are yet to be vetted by the community and could be subject to significant variation depending on who is involved in the development work and/or require larger committees to ensure that multiple perspective are included. When a framework is reviewed periodically (e.g., every two years for math and language arts, every four years for science), there is a danger that changes will reflect ideas still being wrestled with and not yet representative of consensus. As such, there is a danger that—over time—the assessment framework may not capture progress towards more contemporary expectations for student learning but rather reflect a meandering pathway indicative of the field's lack of consensus. Although practice may also change in complex ways, the longer timescale for shifts in practice reduces the risk of random variation from one test administration to the next.

What challenges might arise with more frequent, smaller shifts in assessment frameworks?

Despite potential affordances for capturing changes in practice over time, this solution has some practical and conceptual issues centered around how the need for framework revision will be determined and implications for framework revision when change is slow.

When new expectations for student learning are identified (e.g., in standards), there are clear dates when standards are released, which might help to determine when large framework revisions might be needed. Although more detailed descriptions will still be needed to determine the appropriate lag between a “disruption” and the state of framework revision, standards can serve as a clearly identifiable triggering event (even if revisions do not occur immediately) and signal the significance of changes to be made. Similar events do not necessarily occur to prompt revisions to capture gradual shifts in practice. Are there detectable shifts in practice (nationwide) that might suggest when smaller revisions should be considered? How big of a shift is required to warrant small revisions to the framework?

The nature of shifts warranting framework revision is especially salient given that research on the implementation of standards indicate the slow pace at which change occurs. Results from the

National Survey of Science and Mathematics Education (NSSME) found “few differences in science instruction [in 2018] compared to 2012” and that “teachers in [NGSS-]adopting states vary little from those in non-adopting states” (Smith, 2020, p. 608). At the same time, the data suggested small changes that—coupled with the wider availability of NGSS-aligned curriculum materials (e.g., [OpenSciEd](#))—may portend larger shifts. In the meantime, if shifts are hard to detect, how can the need for framework revision be determined? If current practice does not change significantly in response to new expectations for student learning, should standards continue to reflect old standards?

Unless the criteria to trigger a small framework revision are clear, revisions may need to occur on a regular schedule (e.g., with each administration of the assessment or with every other administration for the frequently tested subjects of math and reading). However, if frameworks are reviewed every time the assessment is administered, there could be a tendency to make unnecessary changes because the opportunity to do so has been presented, not necessarily because the changes are truly needed. Therefore, any process of framework revision must favor a conservative approach (i.e., making as few changes as possible). For example, revisions to the TIMSS assessment framework are considered for each administration. However, suggestions for revision are closely bounded—e.g., consideration of the appropriateness (but not the importance or relevance) of content in the framework, not changing the item blueprint between administrations—and focus on updating the framework (in its current form).

Small revisions with each administration would be impractical if the current timelines for framework and item development remain. The proposals for standing subject-matter panels to “have a role in incremental updates to frameworks” (Rosenberg, 2022, p. 4) and to “provide content expertise to the NCES item development contractor” (Rosenberg, 2022, pp. 8-9) seem promising. I would argue that standing committees already play both roles, to a limited degree, as standing committee members’ evolving interpretations of key terms may lead to small shifts in the way the framework is operationalized and Standing Committee feedback on items includes feedback on the item content. However, the composition of the Standing Committees becomes crucial because a single group of people would be expected to provide a wider range of expertise and have tremendous influence on the assessment in a given subject area. This is particularly true if the process is further expedited by reducing stakeholder input, such that additional stakeholder groups may need to be represented on the Standing Committees.

Incremental implementation of assessment frameworks

Given the relative advantages and disadvantages of the two approaches—and the need for balance between current standards and current practice—I take up Rosenberg’s (2022) suggestion of “an alternative to updating frameworks on an incremental basis”: “implement[ing] framework changes incrementally” (p. 10)

What are some advantages of incremental implementation of assessment frameworks?

In addition to the advantages Rosenberg describes (distributing the cost of new item development across the years of implementation and maximizing opportunities for trend), I see several affordances of this approach:

- Framework revision could be responsive to new consensus in the field, while acknowledging that changes in practice occur incrementally. This could allow for a more natural balance between current standards and current practice. A relatively quick reaction to new consensus views could be followed by incremental implementation of the new assessment framework in parallel with the relatively slow process of moving from theoretical principles in a standards document to widespread curriculum and instruction aligned with those standards.
- NAEP assessment frameworks could serve as a model (i.e., “gold standard”) for other assessments throughout the country and the world, without needing to be at the forefront of all aspects of assessment technology.
- As compared to major framework changes that are rolled out all at once, NAEP assessments would be better positioned to respond to new knowledge about assessment and to the relative speed with which new standards are being reflected in practice. Even if draft plans were in place to describe the incremental changes to be made in each successive implementation of an assessment framework, these could be revisited prior to each test administration. Small changes might be made to the implementation plan based on
 - “lessons learned” from the previous test administration,
 - recent developments in the innovative assessment techniques available for use in NAEP, and
 - uptake of new expectations for student learning (e.g., standards being adopted faster or slower than anticipated).

Especially when shifts in consensus views of student learning require novel approaches to assessment, smaller shifts in items (and item types) might be more feasible than larger shifts, as the field works to overcome challenges with assessing new expectations. By being responsive to innovation as it occurs—rather than waiting for the next large framework revision—the incremental approach to implementation of a new assessment framework could position NAEP assessments to continue to represent state-of-the-art item design, even as demands for novel assessment techniques increase.

- Framework revision could reflect gradual progress towards ambitious goals (i.e., looking to the future). This may be especially important, as assessments communicate to stakeholders what should be prioritized in students’ learning (NAGB, 2008). If framework revision is too closely tied to status quo curriculum and instruction, this could slow the pace of reform, by signaling that old approaches are satisfactory when shifts in practice stall.

What might incremental implementation of an assessment framework look like?

With this approach, a major framework revision could be “triggered” by a major disruption, such as the introduction of standards reflecting substantially different expectations for student

learning. By relying on a major disruption to signal the need for framework revision, there is less ambiguity as to whether a revision is needed (as well as a greater probability that a break in trend may be needed). As noted above, there would need to be some lag (e.g., for standards adoption and implementation, for further development of views put forth in the standards, for the availability of appropriate assessment techniques), but the process could begin much earlier than has occurred over the past two revisions of the science framework.

As part of the new framework, a plan could be devised for incremental movement from the current assessment framework to the new one. The process of articulating these incremental changes might also be helpful in fleshing out exactly what is changing with the new framework and the implications of those changes for items design. Thus, a plan for incremental implementation of an assessment framework could be developed iteratively with the framework itself.

Under such a plan, the first assessment administered with the new assessment framework would better reflect the old framework than the new one, while still including small changes towards the new framework. Incremental changes might occur over years of test administration by addressing changes in one new component of the framework in each iteration, by making incremental changes in all components of the framework simultaneously, or some combination of the two. The more intertwined the changes in the new framework, the more incremental changes to the whole framework (as opposed to concentrating on a single component) would be needed.

This approach is, of course, not a panacea. Decisions would still need to be made about when to start a major framework revision, the period of time over which it should be implemented, and the best way to introduce change. The process of adjusting a planned shift in implementation would need to be streamlined (and focused only on shifts in the framework identified in the plan) to avoid unnecessary revision and, thus, cost. However, this approach provides a way to reduce some of the tensions identified for each of the other approaches and has some clear advantages in terms of relevance, trend, and lessons learned and could have cost benefits, depending on how the incremental implementation takes place.

Conclusion

While the incremental implementation of standards (previous section) seems logical to me, the question of which approach is best seems to be somewhat of a values question. The balance between current standards and current practice seems unavoidable in considerations of how NAEP assessment frameworks should respond to shifts in the field. The lag between standards and practice depends on the nature and magnitude of the shift being introduced. For seismic shifts in a field, this lag creates significant tension between current standards and current practice, and this tension could persist for years. While the “Goldilocks” solution of implementing less frequent, larger shifts incrementally provides one way to balance between the two; a slight weighting of the balance towards one or the other may render one of the other

approaches to be more reasonable. From this perspective, clearer guidance about the nature of this “delicate balance” (NAGB, 2008a, p. 2) could be an important part of guidelines for framework revision. Such guidance could include tensions that arise when current practice may vary widely between states that have and have not adopted national standards. Clarity about the extent to which NAEP should be forward-looking (reflecting more ambitious goals for student learning that may not yet be well-supported) or grounded in the present (reflecting the opportunities that are currently available for student learning) would allow for clearer decisions about framework revision.

References

- American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. Oxford University Press.
- Council of Chief State School Officers. (2015, September). *Science Assessment Item Collaborative Assessment framework for the Next Generation Science Standards*. https://ccsso.org/sites/default/files/2017-12/SAICAssessmentFramework_FINAL.pdf
- Citizens for Objective Public Education. (2020, July). *State adoptions of science standards*. <https://www.copeinc.org/docs/State-Adoptions.pdf>
- Dickinson, E. R., Gribben, M., Schultz, S. R., Spratto, E., & Woods, A. (2021, February). *Comparative analysis of the NAEP science framework and state science standards: Technical report*. <https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/science/NAEP-Science-Standards-Review-Final-Report-508.pdf>
- Mullis, I. V. S., & Martin, M. O. (Eds.) (2013). *TIMSS 2015 assessment frameworks*. TIMSS & PIRLS International Study Center. <https://timssandpirls.bc.edu/timss2015/frameworks.html>
- Mullis, I. V. S., & Martin, M. O. (Eds.) (2017). *TIMSS 2019 assessment frameworks*. TIMSS & PIRLS International Study Center. <https://timss2019.org/wp-content/uploads/frameworks/T19-Assessment-Frameworks-Chapter-2.pdf>
- Mullis, I. V. S., Martin, M. O., & von Davier, M. (Eds.) (2021). *TIMSS 2023 assessment frameworks*. TIMSS & PIRLS International Study Center. <https://timssandpirls.bc.edu/timss2023/frameworks/index.html>
- National Assessment Governing Board. (2005). *Science Framework for the 2005 National Assessment of Educational Progress*.
- National Assessment Governing Board. (2022a, March). *National Assessment Governing Board assessment framework development policy statement*.
- National Assessment Governing Board. (2022b, May). *The National Assessment Governing Board charge to the Steering and Development Panels for the 2028 National Assessment of Educational Progress (NAEP)*.
- National Assessment Governing Board. (2008, September). *Science Framework for the 2009 National Assessment of Educational Progress*.

- National Assessment Governing Board. (2014, November). *Science Framework for the 2015 National Assessment of Educational Progress*.
- National Assessment Governing Board. (2019, February). *Science Framework for the 2019 National Assessment of Educational Progress*.
- National Center for Educational Statistics. (n.d.). *Participating Countries*.
<https://nces.ed.gov/timss/participation.asp>
- National Center for Educational Statistics. (2010.) *An introduction to NAEP: National Assessment of Educational Progress* (NCES 2010-468).
<https://nces.ed.gov/nationsreportcard/pdf/parents/2010468.pdf>
- National Center for Education Statistics. (2021, May). *What does the NAEP science assessment measure?* <https://nces.ed.gov/nationsreportcard/science/whatmeasure.aspx>
- National Research Council. (1996). *National science education standards*. The National Academies Press. <https://doi.org/10.17226/4962>
- National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.
<https://doi.org/10.17226/13165>
- National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. The National Academies Press. <https://doi.org/10.17226/18409>.
- National Research Council. (2015). *Guide to implementing the Next Generation Science Standards*. The National Academies Press. <https://doi.org/10.17226/18802>.
- National Science Teaching Association. (n.d.) *About the Next Generation Science Standards*.
<https://ngss.nsta.org/About.aspx>
- Neidorf, T., Stephens, M., Lasseter, A., Gattis, K., Arora, A., Wang, Y., Guile, & Holmes, J. (2016, May). *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*.
https://nces.ed.gov/nationsreportcard/subject/science/pdf/ngss_naep_technical_report.pdf
- NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press. <https://doi.org/10.17226/18290>
- Organisation for Economic Co-operation and Development. (2006). *Assessing scientific, reading, and mathematical literacy: A framework for PISA 2006*.
- Pellegrino, J. W. (2021, October). *NAEP validity studies white paper: Revision of the NAEP science framework and assessment*.
- Rosenberg, S. (2022, April). *Considerations for smaller, more frequent changes to NAEP assessment frameworks*.
- 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.
<https://doi.org/10.1080/1046560X.2020.1786261>



Update on NAEP Item Development for the 2026 Mathematics and Reading Assessments

At the August 2022 meeting of the Assessment Development Committee (ADC), the National Center for Education Statistics (NCES) provided an update on development of mathematics and reading items for the 2026 grades 4 and 8 assessments that are aligned to the new NAEP Mathematics and Reading Assessment Frameworks, including the development of easier items. The November 2022 session will continue that discussion by updating the ADC on mathematics item pre-testing activities that will inform item development for the 2026 assessment, with a focus on items assessing mathematical practices, a new dimension of the mathematics framework. Selected pilot mathematics discrete items and scenario-based tasks (SBTs) were recently administered to students in cognitive interviews and tryouts to collect information about how students understand and respond to the items and SBTs. The pre-testing was designed to examine the following questions:

- Do students demonstrate an understanding of the items or SBTs and what they are being asked to do?
- Do the items elicit the expected responses?
- How do items and scoring guides perform in conditions similar to an operational assessment?

During this session, NCES will present findings from the mathematics pre-testing to the ADC, including showing selected secure items and student pretesting performance, and discuss how the pre-testing results will be used to revise the pilot items.

During this session, NCES also will brief the ADC on plans to introduce multi-stage adaptive testing in the NAEP mathematics and reading assessments. The presentation will describe preliminary approaches to routing students to blocks targeted to ability and plans for the development of routers and targeted blocks, validating routing decisions, and implementation.

**Assessment Development Committee
Item Review Schedule
January – July 2023
As of October 21, 2022**

Review Package to Board	Board Comments to NCES	Survey/ Cognitive	Review Task	Approx. Number Items	Status
2/13/2023	3/10/2023	Cognitive	Reading (4, 8) <i>2024 Operational</i>	Flagged Items Only [30 items, including 1 discrete item (DI) and 3 Scenario- based tasks (SBTs)]	
2/13/2023	3/10/2023	Cognitive	Mathematics (4, 8) <i>2024 Operational</i>	Flagged Items Only (4 items)	
3/15/2023 (Off-cycle)	4/5/2023 (Off-cycle)	Survey	Reading (4, 8) <i>2026 Operational (2024 Pilot)</i>	111 subitems*	
3/15/2023 (Off-cycle)	4/5/2023 (Off-cycle)	Survey	Math (4, 8) <i>2026 Operational (2024 Pilot)</i>	178 subitems*	
5/3/2023	5/26/2023	Cognitive	Mathematics (4, 8) <i>2026 Operational (2024 Pilot)</i>	10 blocks (Approx. 315 DIs and 30 items in 8 SBTs)	
5/3/2023	5/26/2023	Cognitive	Reading** (4, 8) <i>2028 Operational (2024 Pilot)</i>	84 items*	
5/16/2023 (Off-cycle)	6/9/2023 (Off-cycle)	Cognitive	Reading (4, 8) <i>2026 Operational (2024 Pilot)</i>	15 blocks (Approx. 150-162 items)	

*Cross-grade items are included and counted once.

**To support multi-stage testing in 2028.