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 (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

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 construct[1] 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

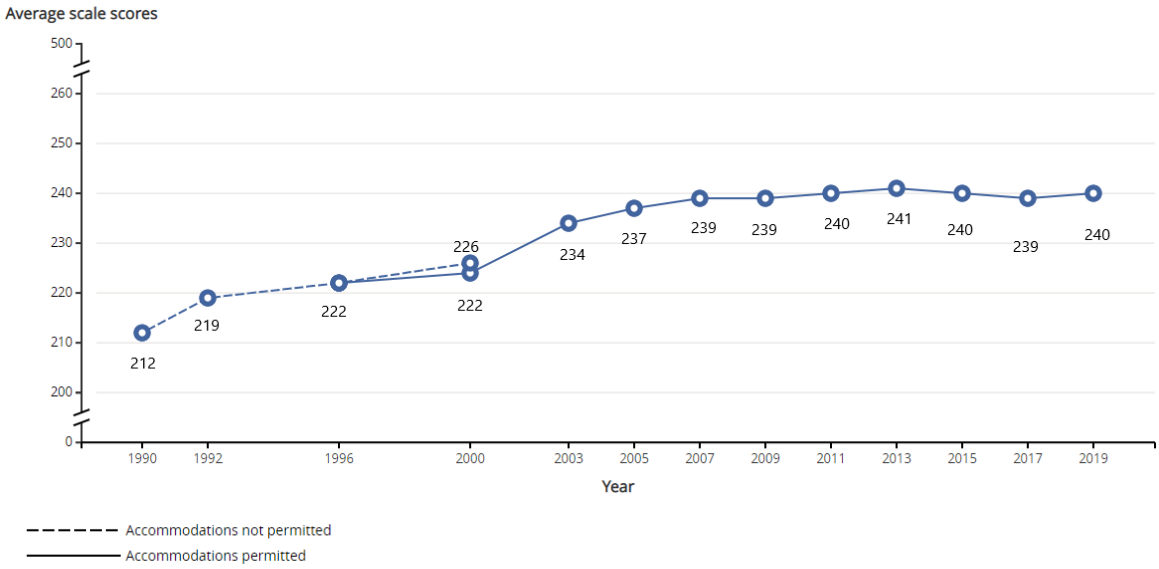
Because they define assessment content, NAEP frameworks also affect the reporting of trend data. As suggested by the words "educational progress" in its name, NAEP's primary purpose is to serve as an independent monitor—reporting on the status of achievement in the nation at the time of each assessment, 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).

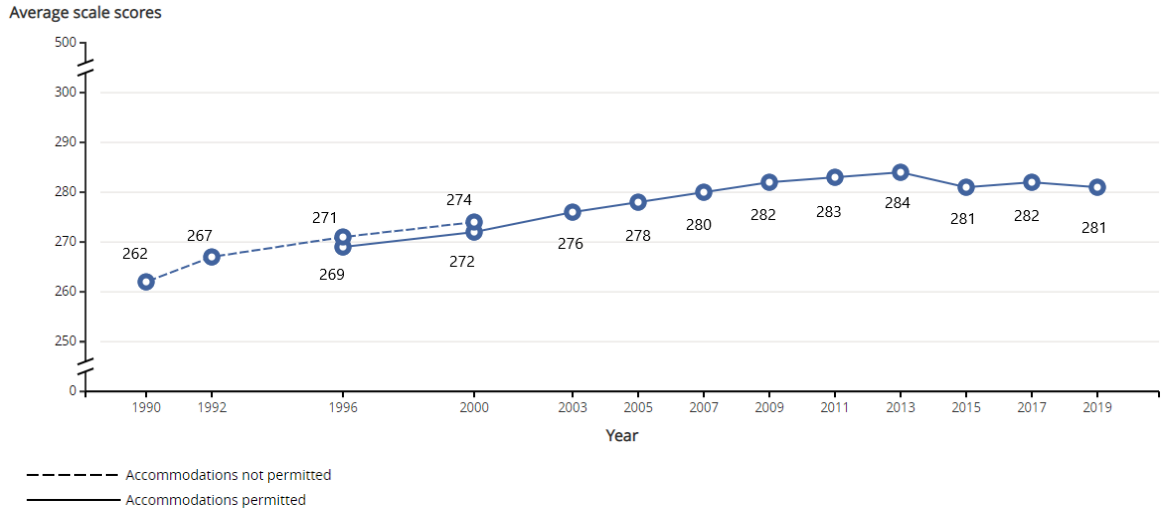
Exhibit 1. NAEP score trends for math and reading grades 4 and 8, national public

Mathematics Grade 4



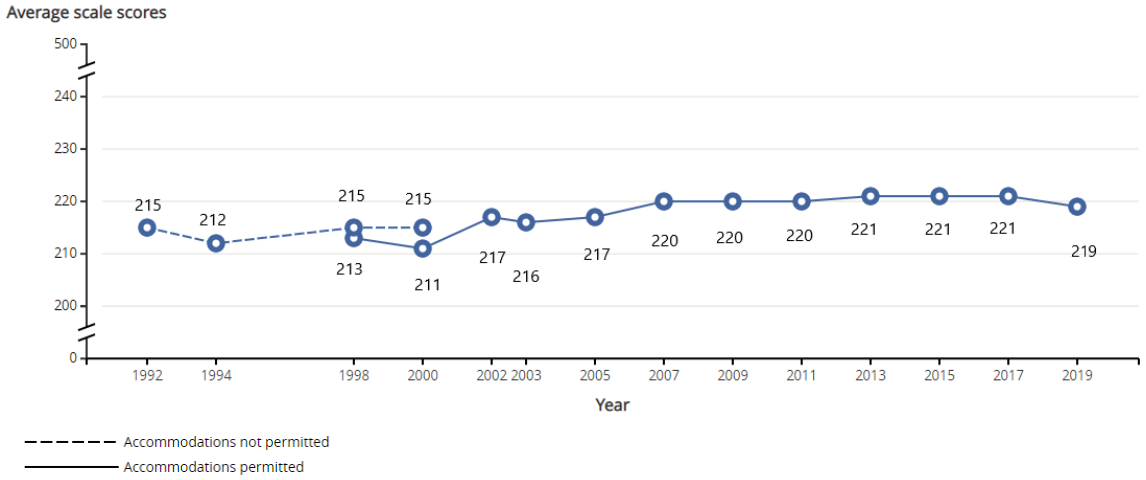
NOTE: The NAEP Mathematics scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 1990, 1992, 1996, 2000, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, and 2019 mathematics assessments.

Mathematics Grade 8



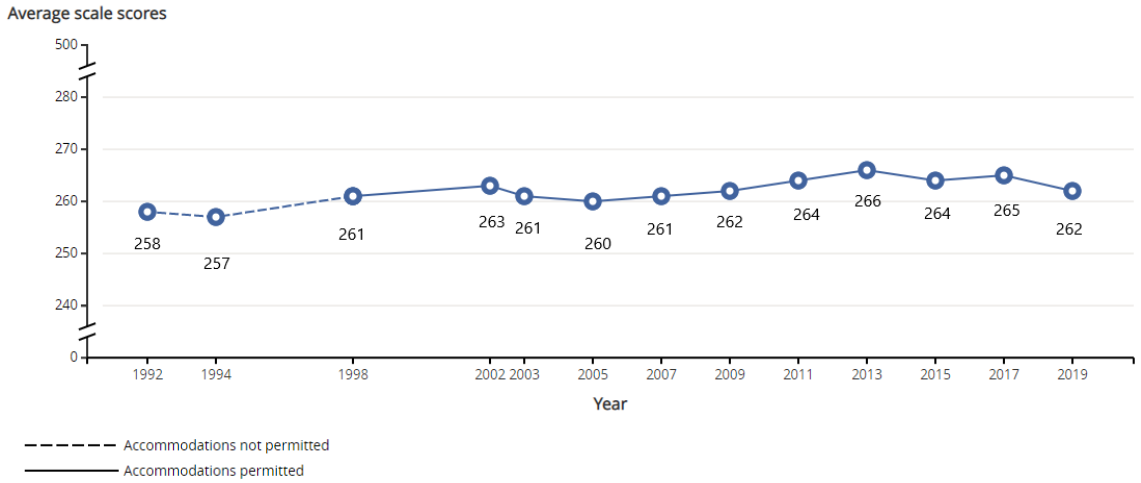
NOTE: The NAEP Mathematics scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 1990, 1992, 1996, 2000, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, and 2019 mathematics assessments.

Reading Grade 4



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



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).

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	8																	■●			□	

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 NAEd 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

<table>
<tr><th></th><th colspan="2">1992–2007 NAEP
Reading Framework</th><th colspan="3">2009–2019 NAEP
Reading Framework</th></tr>
<tr><td>Content</td><td>Content of assessment:
• Literary
• Informational
• Document</td><td>Contexts for reading:
• For literary experience
• For information
• To perform task</td><td colspan="2">• Literary text
• Fiction
• Literary nonfiction
• Poetry</td><td>• Informational text
• Exposition
• Argumentation and persuasive text
• Procedural text and documents</td></tr>
<tr><td>Cognitive Processes</td><td colspan="2">Stances/aspects of reading:
• Forming general understanding.
• Developing interpretation.
• Making reader/text connections.
• Examining content and structure.</td><td colspan="3">Cognitive targets distinguished by text type
Locate/ recall | Integrate/ interpret | Critique/ evaluate</td></tr>
<tr><td>Vocabulary</td><td colspan="2">Vocabulary as a target of item development, with no information reported on students' use of vocabulary knowledge in comprehending what they read.</td><td colspan="3">Systematic approach to vocabulary assessment with potential for a vocabulary subscore.</td></tr>
<tr><td>Poetry</td><td colspan="2">Poetry included as stimulus material at grades 8 and 12.</td><td colspan="3">Poetry included as stimulus material at all grades.</td></tr>
<tr><td>Passage Source</td><td colspan="2">Use of intact, authentic stimulus material.</td><td colspan="3">Use of authentic stimulus material plus some flexibility in excerpting stimulus material.</td></tr>
<tr><td>Passage Length</td><td colspan="2">Grade 4: 250–800 words
Grade 8: 400–1,000 words
Grade 12: 500–1,500 words</td><td colspan="3">Grade 4: 200–800 words
Grade 8: 400–1,000 words
Grade 12: 500–1,500 words</td></tr>
<tr><td>Passage Selection</td><td colspan="2">Expert judgment as criterion for passage selection.</td><td colspan="3">Expert judgment and use of at least two research-based readability formulas for passage selection.</td></tr>
<tr><td>Item Type</td><td colspan="2">Selected-response and constructed-response items included at all grades.</td><td colspan="3">Selected-response and constructed-response items included at all grades.</td></tr>
</table>

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

	Categorical Concurrence		Range of Knowledge		Balance of Knowledge	
Content Area	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

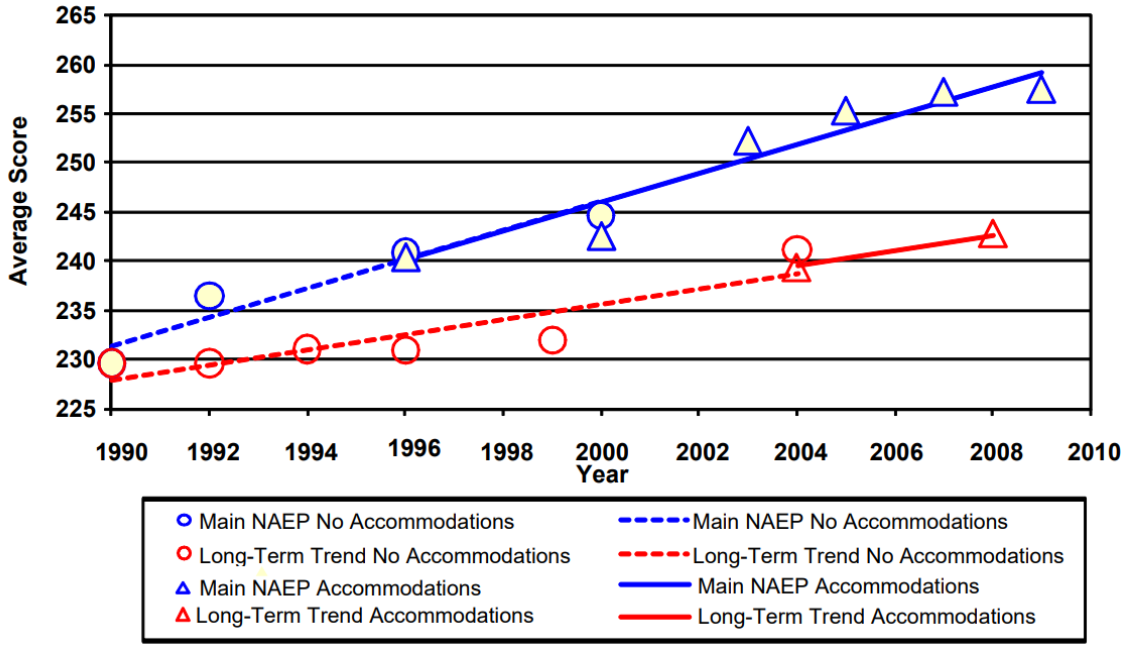
	2003 NAEP		LTT	
Depth of Knowledge Level	Mean % of Items Rated This Level		Mean % of Items Rated This Level	
	Mean	SD	Mean	SD
1	42.01	17.17	69.74	17.84
2	47.97	15.26	29.59	17.73
3	9.51	6.92	1.5	1.15
4	0.95	0.37	X	X
	Mode: 1	Range: 2.44	Mode: 1	Range: 1.44

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.

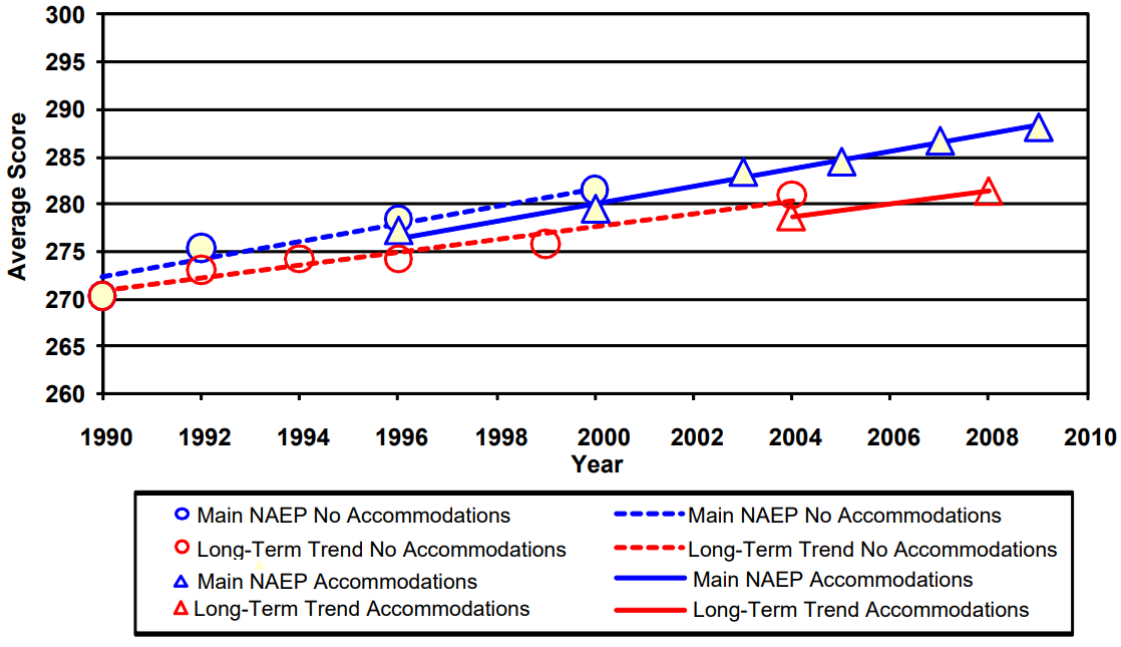
Beaton and Chromy (2010) investigated numerous differences between the two assessment programs, such as the introduction of accommodations and greater inclusion and private school coverage. They also provided in-depth analyses of differential trends by racial and ethnic groups, of trend components attributable 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



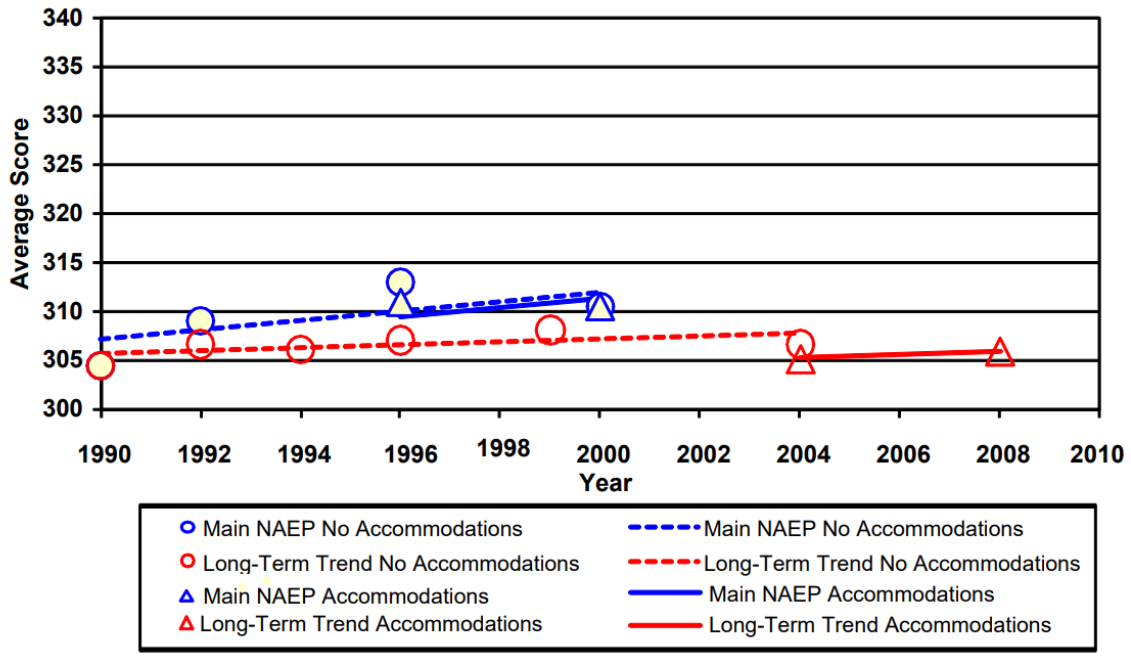
NOTE: The 1990 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.
SOURCE: Beaton & Chromy, 2010, figure 2.1.

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



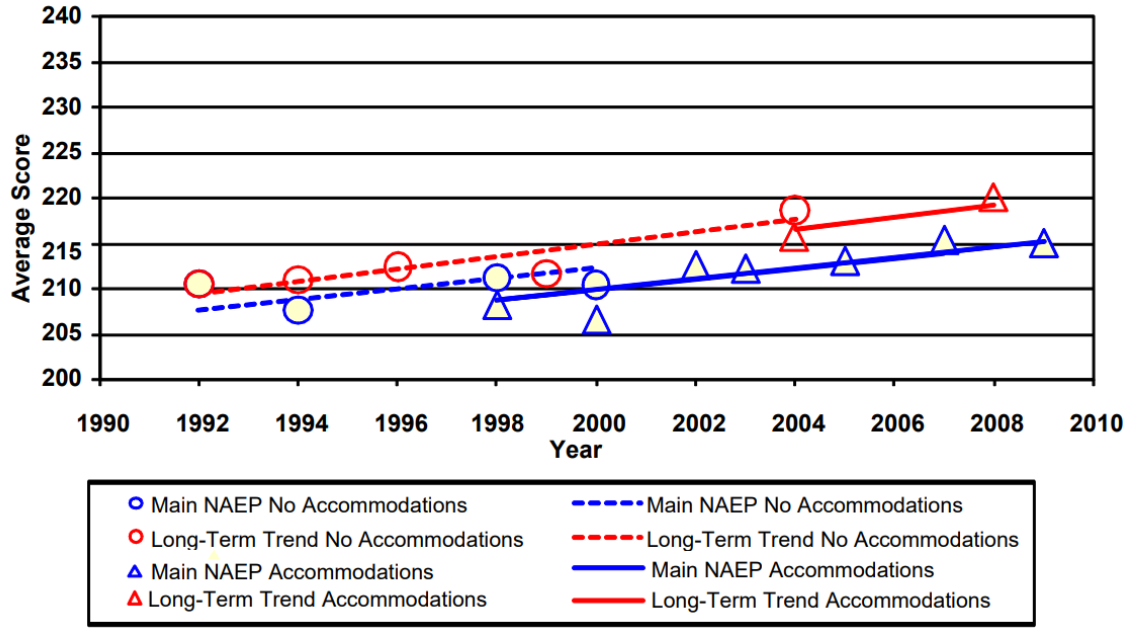
NOTE: The 1990 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.
SOURCE: Beaton & Chromy, 2010, figure 2.2.

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 LLT 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 LLT 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
Weight in NAEP framework	*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 detectible 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 a 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.

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.

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*.

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.