

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

Andrew Ho, *Harvard Graduate School of Education*

October 7, 2022

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

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

The challenge and importance of measuring educational progress

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Recommendation 3: Adopt different perspectives on trend validation and reporting

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

a) A “moving window” perspective on trend validation

As content and contexts shift over time, the change is gradual, not static and then suddenly different. Accordingly, I consider the goal of scale maintenance to be the goal of enabling scale meaning in any “moving window” of 4-8 years along a timeline, rather than ensuring rigid comparability of the most recent year of results to an originating baseline year. One metaphor that illustrates this comes from the

proverb, “the bamboo that bends is stronger than the oak that resists.” The bamboo tree’s flexibility is akin to an equating design like NAEP’s, where developers deploy common items and tasks through a window and then retire them to allow new items to become common items. The oak tree’s inflexibility is akin to a reference set of items that stays constant over time, where new items are regularly linked back to this reference set. As the metaphor suggests, the former design enables flexible trend estimation within moving windows, whereas the latter design is destined to break as the content and context of learning changes.

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



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

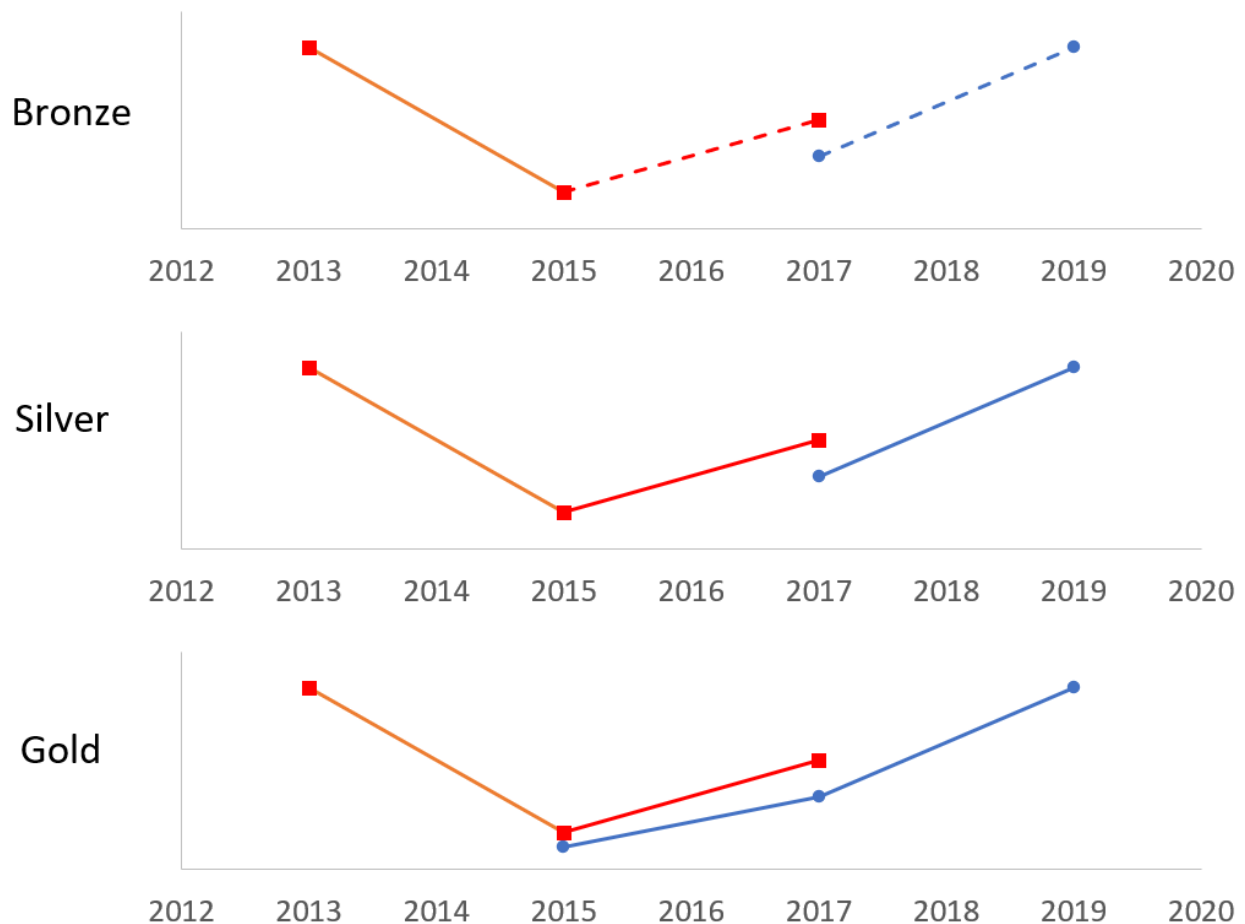
b) Three different levels of “bridge studies”

Bridge studies provide valuable evidence for trend validation. In cases where the Governing Board and NCES anticipate that trend maintenance will be difficult due to shifting content and contexts, high-quality bridge studies can act as a safeguard against a broken trend. Bridge studies can support proposed transitions from existing to newer assessment content and context. The Governing Board can deploy higher quality bridge studies when maintaining trend is particularly challenging. I distinguish these bridge studies here as bronze, silver, and gold.

The bronze bridge study takes a relatively small number of common or calibrated items on the time 1 scale and embeds them into a time 2 administration. This is a standard test equating procedure where there is no anticipated substantial shift in construct, mode, population, or method over the intervening

time period. Standard psychometric methods can evaluate whether items are functioning differently over time. If the old and new items are not comparable on the same scale, then developers may have to salvage an imprecise trend from the time 1 common items alone. It is also possible that the trend breaks altogether because items change inconsistently in their location or format. This trend may be more imprecise. It only guarantees a tentative measure of progress from time 1 to time 2 and only on the time 1 scale. The time 2 items establish a new baseline that is not comparable to the time 1 scale. The time 2 items may also form an imprecise baseline due to limited deployment. Reports sometimes show results from a bronze bridge study as staggered points shown in Figure 2 below. This is misleading because it implies that a simple shift can equate the trends across the break.

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



The silver bridge study administers two different modes, frameworks, or methods at timepoint 2. For example, if NAEP aimed to bridge from a paper-based test mode to a digitally based test mode, it could administer both at timepoint 2. Comparison of measurement properties, both within timepoint 2 and

for the common administration from timepoint 1 to timepoint 2, provides additional reassurance that mode A is stable from timepoint 1 to timepoint 2 and that mode B is comparable to mode A in timepoint 2. If the bridge fails, then there is a common pretrend for mode A that ends at timepoint 2, and a new trend baseline that begins at timepoint 2 in mode B. Thus, if the bridge fails, there is a measure of progress from timepoint 1 to timepoint 2, but only for mode A.

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

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

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

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

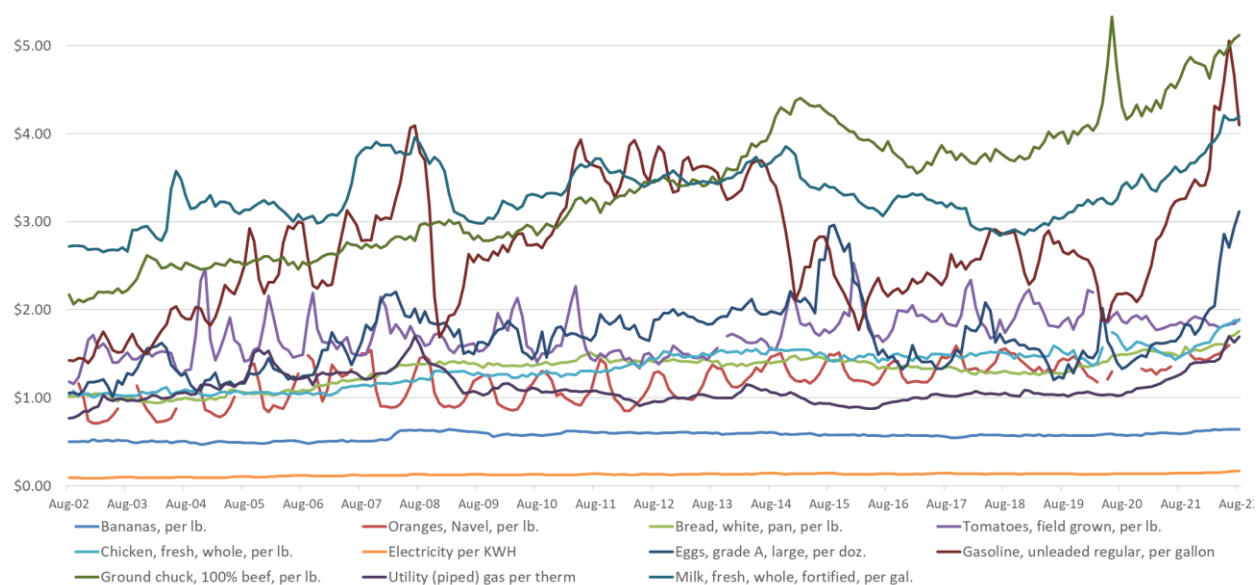
The Future of NAEP Report ([2012](#)) made the apt comparison to the Consumer Price Index (CPI) that monitors inflation. The CPI rotates goods through a “market basket” as goods become antiquated; for example, a rotary phone is eventually replaced by mobile phone. Validation of this “linkage” is not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

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

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

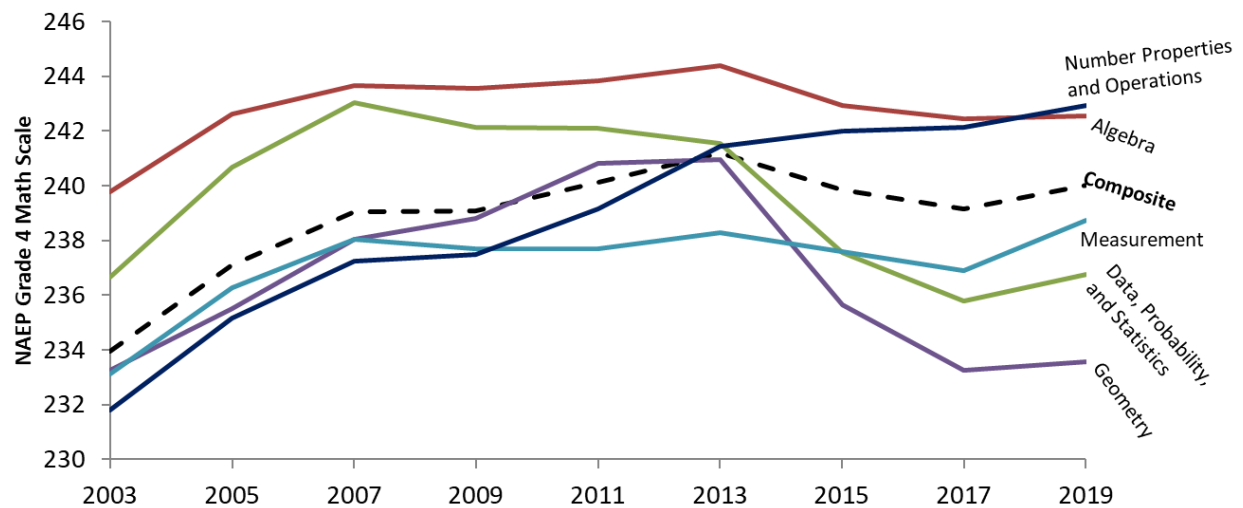
From the perspective of indexing, such incoherence may not be consequential. Consider the goal of measuring the change in average consumer prices, as Figure 3 illustrates below. If gasoline becomes much more expensive but bananas do not, there may in fact be some incoherence, because people who eat more bananas and take public transportation may not be experiencing the same price increases as people who drive and eat beef. However, the CPI’s target inference is at an aggregate level, and inflation has important economic implications. Coherence is a small sacrifice to answer an important economic question.

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



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

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



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

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

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

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

References

- Haertel, E., Beaugard, R., Confrey, J., Gomez, L., Gong, B., Ho, A., et al. (2012). NAEP: Looking Ahead – Leading assessment into the future. Recommendations to the Commissioner. National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/future_of_naep_panel_white_paper.pdf
- Ho, A. D. (2020). Departing remarks to the National Assessment Governing Board. Retrieved from <https://scholar.harvard.edu/files/andrewho/files/andrewhodepartingnagb.pdf>
- National Research Council. (2001). NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting. Washington, DC: The National Academies Press. <https://doi.org/10.17226/10049>
- National Academies of Sciences, Engineering, and Medicine. (2017). Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress. Washington, DC: The National Academies Press. <https://doi.org/10.17226/23409>
- National Academies of Sciences, Engineering, and Medicine. (2022). A Pragmatic Future for NAEP: Containing Costs and Updating Technologies. Washington, DC: The National Academies Press. <https://doi.org/10.17226/26427>
- National Assessment Governing Board (2017). Mathematics Framework for the 2017 National Assessment of Educational Progress. Retrieved from <https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/mathematics/2017-math-framework.pdf>
- National Assessment Governing Board (2022). Assessment Framework Development. Policy Statement. Retrieved from <https://www.nagb.gov/content/dam/nagb/en/documents/policies/assessment-framework-development.pdf>
- National Research Council. (2001). NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting. Washington, DC: The National Academies Press. <https://doi.org/10.17226/10049>
- Orr, C. S. (2021). History, Policy, and Decision Points for Developing NAEP Frameworks. Retrieved from <https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/History-of-NAEP-Frameworks-Report-Final.pdf>
- Rosenberg, S. (2022). Considerations for smaller, more frequent changes to NAEP assessment frameworks. Retrieved from https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/quarterly-board-meeting-materials/2022-05/6-Assessment_Development_Committee.pdf