

NAEP Math Automated Scoring Challenge: Request for Information Webinar



March 21, 2023

John Whitmer, Sr. Fellow - Institute of Education Sciences

Eunice Greer, Sr. Research Scientist - National Center for Education Statistics

Maggie Beiting-Parrish, Fellow - Institute of Education Sciences

Agenda

1. Introduction (John)
2. NAEP Context & Math Item Description (Eunice)
3. Dataset Description (Maggie)
4. Selection Criteria & Submission Instructions (John)
5. Q & A

Introduction

Challenge Overview

1. Core challenge: predict human scores of open-ended math items. Consider accuracy compared to human inter-rater reliability AND no bias (systematic variance) based on student characteristics (Prize up to \$40k for winning team)
 - Accuracy evaluated at item level (QWK degradation < 0.05 , SMD increase < 0.10)
 - Detailed technical report required that describes algorithms used and training results
2. Innovation challenge: provide interpretable results (beyond required transparency), given that accurate models likely use methods that are difficult to interpret.

Goals for Automated Scoring (AS) for NAEP

This challenge seeks participants to use natural language processing to predict human scores of open-ended math assessment items.

There are three leading goals for NAEP to implement AS:

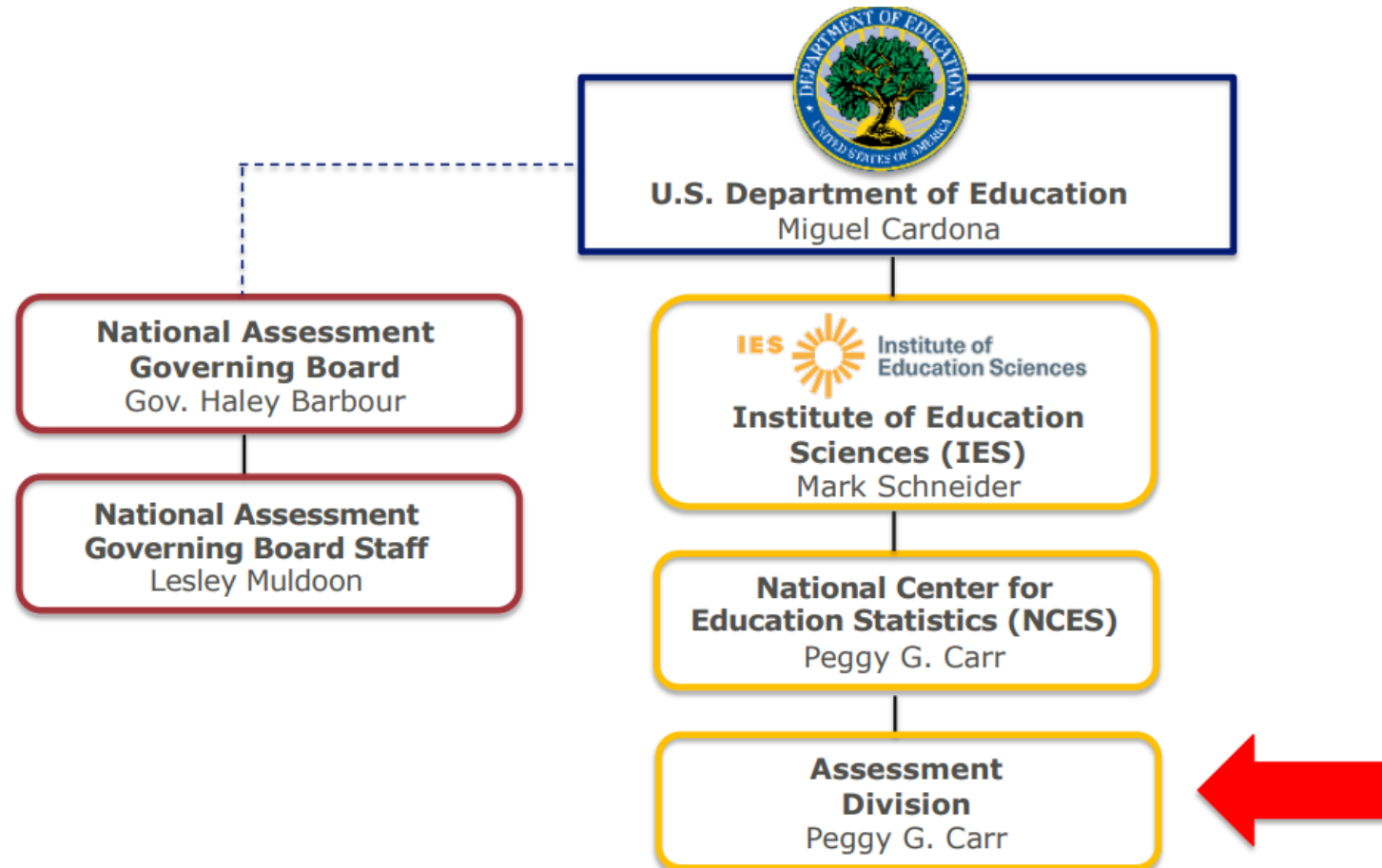
1. Reduce speed to return scores
2. Cost reduction compared to human scoring
3. Increased consistency in scoring

Participation Timeline & Requirements

- ❖ Eligibility: Institutions and individuals that have the ability and capacity to conduct research are eligible to apply. Eligible applicants include, but are not limited to, non-profit and for-profit organizations and public and private agencies and institutions, such as colleges and universities. Participants must be located in the United States.
- ❖ NCES Confidential Data Security application required to participate: **DEADLINE 04/17/2023** (docs available online)
- ❖ Response deadline (technical report & predictions): 05/25/2023
- ❖ Winners Announced: June 2023

NAEP Context & Math Item Description

Governance



Main NAEP: Subjects Assessed

Mathematics



Reading

Writing



Science

Civics



Geography

U.S. History



Economics

Vocabulary














Music

Visual Arts



Main NAEP: Subjects Assessed

Mathematics			Reading
Writing			Science
Civics			Geography
U.S. History			Economics
Vocabulary			Music
Visual Arts			

Mathematics Framework That Guides NAEP Math Assessment

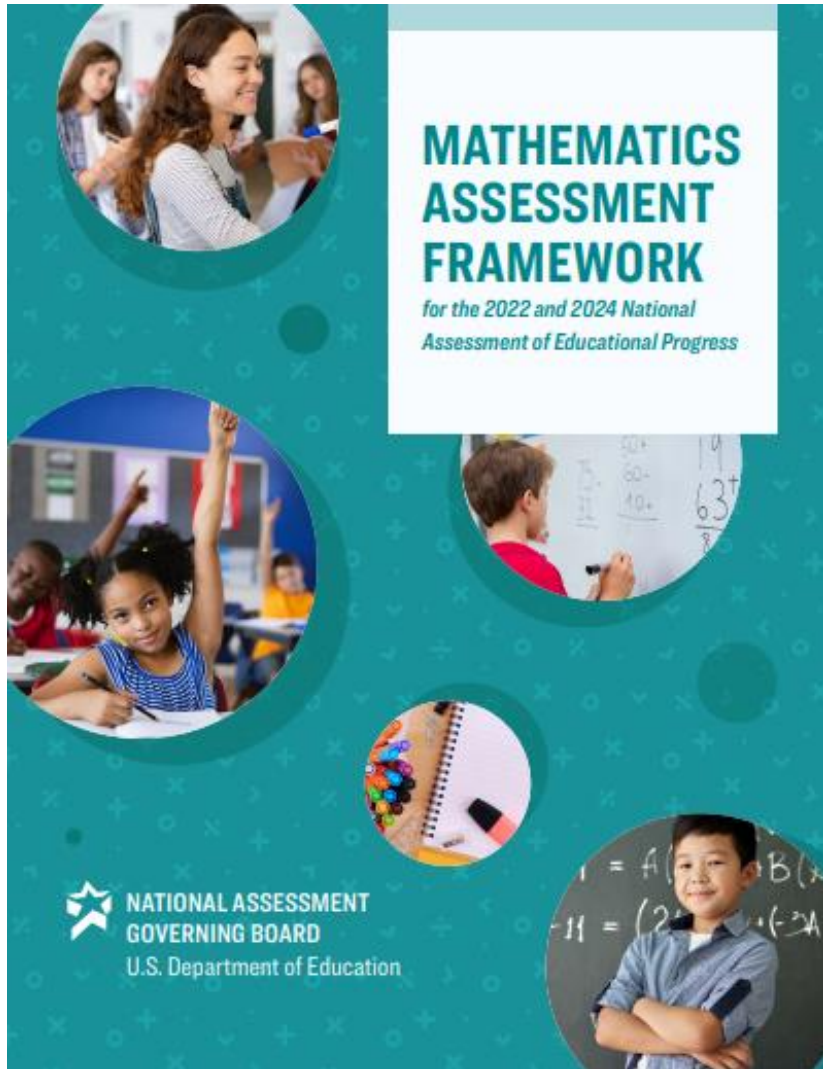
NAEP MATH IS COMPRISED OF TWO COMPONENTS:

❖ Content

- ❖ Number Properties & Operations
- ❖ Measurement
- ❖ Geometry
- ❖ Data Analysis, Statistics, & Probability
- ❖ Algebra

❖ Complexity

- ❖ Low Complexity (25% of testing time)
- ❖ Moderate Complexity (50% of testing time)
- ❖ High Complexity (25% of testing time)



Detailed Information About Each Domain

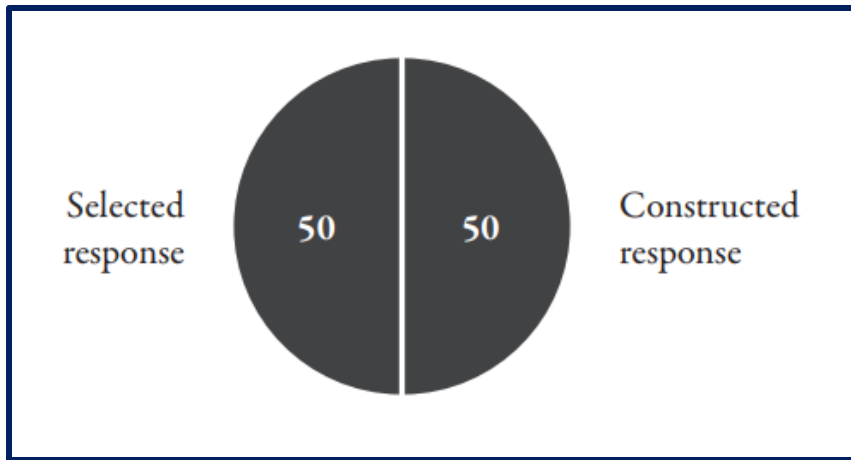
[MATHEMATICS ASSESSMENT FRAMEWORK FOR
THE 2022 AND 2024 NATIONAL ASSESSMENT OF
EDUCATIONAL PROGRESS \(NAGB.GOV\)](https://nagb.gov)

Detailed Information About Mathematical Complexity

Level	Description
Low Complexity	Questions typically specify what a student is to do, which usually involves carrying out a routine mathematical procedure.
Moderate Complexity	Questions involve more flexibility of thinking and often require a response with multiple steps.
High Complexity	Questions make heavier demands on students' thinking and often require abstract reasoning or analysis in a novel situation.

Item Formats

- ❖ Selected Response
 - ❖ (Multiple Choice, Card Sort, Yes/No)
- ❖ Constructed Response
 - ❖ (Short and Extended)



Dataset Description

Detailed Information About Challenge Items

Item Number	Grade	Year(s)	Response Type	Complexity	Content Area
VH134067	4	2017 & 2019	Short Constructed Response	Moderate	Algebra
VH139380	4	2019	Short Constructed Response	Low	Algebra
VH266015	8	2019	Extended Constructed Response	High	Number Properties & Operations
VH266510	8	2017 & 2019	Short Constructed Response	Moderate	Algebra
VH269384	4	2019	Extended Constructed Response	Moderate	Data Analysis
VH271613	4	2017 & 2019	Extended Constructed Response	High	Algebra
VH302907	8	2017 & 2019	Extended Constructed Response	High	Geometry
VH304954	4	2017	Short Constructed Response	Moderate	Number Properties & Operations
VH507804	4	2019	Extended Constructed Response	High	Number Properties & Operations
VH525628	8	2019	Short Constructed Response	High	Number Properties & Operations

Multiple Choice Example

Example 1: Multiple Choice

Grade 4

Number Properties and Operations: Number operations

Source: 2005 NAEP 4M12 #2

Percent correct: 53%

No calculator

$$\frac{4}{6} - \frac{1}{6} =$$

- A. 3
- B. $\frac{3}{6}$
- C. $\frac{3}{0}$
- D. $\frac{5}{6}$

Correct answer: B

Short Constructed Response Example

Example 3: Short Constructed Response

Grade 8

Data Analysis, Statistics, and Probability: Characteristics of data sets

Source: 2003 NAEP 8M7 #13

Percent correct: 19%

Calculator available

Score	Number of Students
90	1
80	3
70	4
60	0
50	3

The table above shows the scores of a group of 11 students on a history test. What is the average (mean) score of the group to the nearest whole number?

Answer: _____

Short Constructed Response Example

Example 3: Short Constructed Response

Grade 8

Data Analysis, Statistics, and Probability: Characteristics of data sets

Source: 2003 NAEP 8M7 #13

Percent correct: 19%

Calculator available

Score	Number of Students
90	1
80	3
70	4
60	0
50	3

The table above shows the scores of a group of 11 students on a history test. What is the average (mean) score of the group to the nearest whole number?

Answer: _____

Scoring Guide
1 – Correct response: 69
0 – Incorrect

Short Constructed Response Example 2

Example 4: Short Constructed Response

Grade 4

Algebra: Patterns, relations, and functions

Source: 2003 NAEP 4M7 #6

Percent correct: 29%, 51%

(incorrect), 17% (partial)

Calculator available

A schoolyard contains only bicycles and wagons like those in the figure below.



On Tuesday, the total number of wheels in the schoolyard was 24. There are several ways this could happen.

- a. How many bicycles and how many wagons could there be for this to happen?

Number of bicycles _____

Number of wagons _____

- b. Find another way that this could happen.

Number of bicycles _____


Number of wagons _____

Short Constructed Response Example 2

Example 4: Short Constructed Response
Grade 4
Algebra: Patterns, relations, and functions

Source: 2003 NAEP 4M7 #6
Percent correct: 29%, 51%
(incorrect), 17% (partial)
Calculator available

A schoolyard contains only bicycles and wagons like those in the figure below.



On Tuesday, the total number of wheels in the schoolyard was 24. There are several ways this could happen.

a. How many bicycles and how many wagons could there be for this to happen?
Number of bicycles _____
Number of wagons _____

b. Find another way that this could happen.
Number of bicycles _____
Number of wagons _____

Scoring Guide
<p>Solution:</p> <p>Any <u>two</u> of the following correct responses:</p> <p>0 bicycles, 6 wagons</p> <p>2 bicycles, 5 wagons</p> <p>4 bicycles, 4 wagons</p> <p>6 bicycles, 3 wagons</p> <p>8 bicycles, 2 wagons</p> <p>10 bicycles, 1 wagon</p> <p>12 bicycles, 0 wagons</p>
<p>2 – Correct: Two correct responses</p>
<p>1 – Partial</p> <p>One correct response, for either part a or part b</p> <p>OR</p> <p>Same correct response in both parts</p>
<p>0 – Incorrect: Any incorrect or incomplete response</p>

Extended Constructed-Response Example

Example 5: Extended Constructed Response

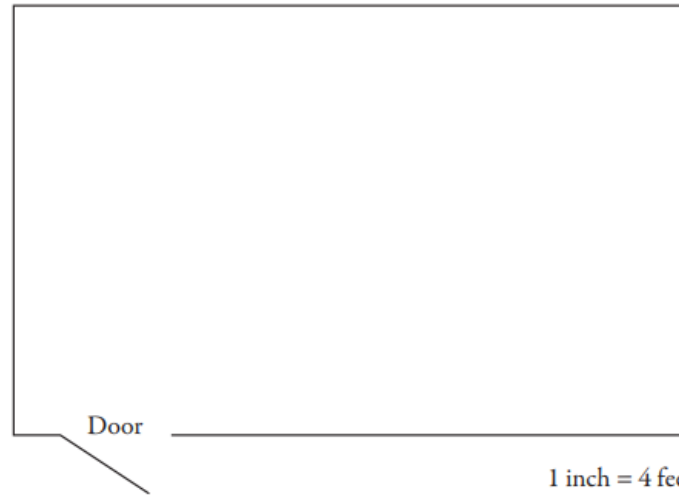
Grade 8

Measurement: Measuring physical attribute

Source: 2005 NAEP 8M3 #18

Percent correct: 9% (extended credit), 5% (satisfactory), 4% (partial), 5% (minimal)

No calculator



The floor of a room in the figure above is to be covered with tiles. One box of floor tiles will cover 25 square feet. Use your ruler to determine how many boxes of these tiles must be bought to cover the entire floor.

_____ boxes of tiles

Extended Constructed-Response Example

Scoring Guide
<p>Solution: 7 boxes Correct process for solution includes evidence of each of the following (may be implied or explicit):</p> <ol style="list-style-type: none">measuring dimensions correctly (getting 2.5 inches and 4 inches) (if centimeters are used the measure must be exact)converting to feet correctly (getting 10 feet and 16 feet)finding the area (160 square feet)dividing by 25 to find the number of boxes (6.4) <p>(Note: Steps b and c may be interchanged; if done this would yield 10 square inches and 60 square feet, respectively)</p>
<p>4 – Extended: Correct responses</p>
<p>3 – Satisfactory Response contains correct complete process as outlined above (a through d) but has a minor error (such as dimensions in inches are measured incorrectly OR the answer to the scale conversion is incorrect OR one other minor computational error OR does not round)</p>
<p>2 – Partial 7 with no explanation OR response contains correct complete process as outlined above (a through d) but has a major conceptual error (such as use of incorrect conversion factor OR use of perimeter [52] instead of area OR perceives the floor as a square and performs all 4 steps)</p>
<p>1 – Minimal 6.4 with no explanation OR Measures 2.5 inches and 4 inches correctly and gets 10 square inches for area OR Measures 2.5 inches correctly and converts correctly to 10 feet and 16 feet (may also indicate area is 160 square feet)</p>
<p>0 – Incorrect Any incorrect response</p>

Data Fields

Variable	Description	Type
student_id	pseudonymous student ID -- not linkable across item-years	string
accession	Item number	string
score_to_predict	Outcome to predict	integer
predict_from	Text related to "score_to_predict"	string
year	Year assessment was administered	integer
srace10	Student's race reported by the school	string
dsex	Student's sex	integer
accom2	Student accommodations. Note: Item VH304954 did not have accom2 so for this item accom2 is entirely NA.	integer
iep	IEP	integer
lep	English learner status	integer
rater_1	Score given by human rater (Type II items only)	string
pta_rtr1	Part A human rater score	string
ptb_rtr1	Part B human rater score	string
ptc_rtr1	Part C human rater score	string
composite	Composite score	integer
score	Score (containing partial credit codes)	string
assigned_score	Simplified numeric score total for item (1, 2, 3...) from either "rater_1" or "composite"	integer
ee_use	Item used equation editor	integer

Selection Criteria & Submission Instructions

Two Challenges

- ❖ Challenge 1: Score Prediction Challenge
 - ❖ First Prize is \$40,000
 - ❖ Two runner-up prizes of \$15,000 each
- ❖ Challenge 2: Interpretability Analysis Challenge
 - ❖ First prize is \$20,000
 - ❖ Two runner-up prizes of \$5,000 each

Prediction Challenge

- ❖ Detailed technical report that meets transparency and fairness requirements reviewed before accuracy evaluated
- ❖ Predict the human raters' scores as accurately as possible using a computer-based algorithm
- ❖ Accuracy determined at item-level (count), then average accuracy within items (in case of tie).
- ❖ Ensure that there is no bias against any demographic group

Interpretability Challenge

- ❖ If the challenger submits a sufficiently accurate model (in top 10 entries), they are also included in the interpretability challenge
- ❖ This challenge is looking for simulatability and clearly described post-hoc measures

Submission Requirements

Valid submissions will include:

- ❖ A Technical Report with a thorough description of the modeling process and comprehensive fairness analysis
- ❖ Predicted scores (in CSV format), using the test dataset

Evaluation Criteria: Prediction Challenge

- ❖ Technical Report

 - ❖ Transparency

 - ❖ Fairness

 - ❖ Insights

- ❖ Scoring Model Accuracy

 - ❖ Number of items meeting accuracy thresholds

 - ❖ Prediction compared to human agreement (Quadratic-Weighted Kappa)

 - ❖ Results do not demonstrate bias (Standardized Mean Difference less than 0.10)

Evaluation Criteria: Interpretability Challenge

- ❖ Construct Coverage
 - ❖ Complete coverage of constructs, evidence of trustworthy model
- ❖ Subpopulation Analyses
 - ❖ Complex methods for identifying intersectional identity, diversity of backgrounds included
- ❖ Clarity
 - ❖ Intuitive to reader, inclusion of visualizations, minimal time and effort needed to understand

Data Security Application Requirements

- ❖ Will need to fill out a data security packet
- ❖ This includes an NDA as well as a plan for how to keep NAEP data secure
- ❖ The NDA needs to be notarized and mailed, we can take a digital copy for the initial data submission but do need the paper copy to fully release the data
 - ❖ Can get free notary service at most major banks
- ❖ Also need data destruction form returned within 30 days of the end of the challenge

Timeline

**	Item	Duration (in Days)	Start	Finish
2.1	Challenge Posting Period	41	03/07/2023	04/17/2023
2.2	Request for Information Webinar	1	03/21/2023	
2.3	Application Deadline	1	04/17/2023	
2.4	Training Data Available	78	03/08/2023	05/25/2023
2.5	Test Data Available	4	05/22/2023	05/25/2023
2.6	Submission Deadline	1	05/26/2023 @ 11:59 EST	
2.7	Winners Announced	1	June 2023	

Dataset Structure & Scoring

Training Dataset Included Files

- ❖ PDF containing all of the item descriptions and scoring material
- ❖ Longer PDFs for each item that show examples of different work products and their scores used to train the human raters
- ❖ Two training files containing all items in .txt format and .rda format
- ❖ A ReadMe that explains all of the variables unique to each item
- ❖ Two files called “Example_test_data.txt” and “Example_test_data.rda” which show which variables will be converted to “NA” in the test set
- ❖ A script to support import of these files into Python

Included Demographic Variables

- ❖ Sex
- ❖ Race
- ❖ Student Accommodation Status (Binary)
- ❖ IEP Status (Binary)
- ❖ English Learner Status (Binary)

Scoring

- ❖ We are looking for the submission with the highest Quadratic-weighted Kappa (QWK) between the predictions and the existing human scores for the most items
- ❖ THEN the most interpretable report
- ❖ We are also using Standardized Mean Difference (SMD) to look for evidence of bias between demographic groups

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

$$\frac{\mu_A - \mu_B}{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

Q & A

Participate!

- ❖ Challenge URL: Challenge.Gov
- ❖ Deadline for Security Applications: **04/17/2023**
- ❖ Deadline for Submissions: **05/26/2023 @ 11:59 EST**
- ❖ Questions: automated-scoring-challenge@ed.gov

Useful Links

NAEP Math 2019 Framework

[MATHEMATICS FRAMEWORK for the 2019 NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS \(nagb.gov\)](https://nagb.gov/math/framework/2019)

NAEP Item Maps

[Item Maps \(nationsreportcard.gov\)](https://nationsreportcard.gov/item-maps)

NAEP Questions Tool

[The NAEP Questions Tool \(nationsreportcard.gov\)](https://nationsreportcard.gov/questions-tool)

NAEP Tutorials on the Web

[Digitally Based Assessment | NAEP](https://nationsreportcard.gov/digital-assessment)

2017 Sample Questions for 4th and 8th Grade

[NAEP Mathematics: Sample Questions \(nationsreportcard.gov\)](https://nationsreportcard.gov/sample-questions)

General Information on the NAEP Math Assessment

[NAEP Mathematics: Framework \(nationsreportcard.gov\)](https://nationsreportcard.gov/framework)

2025 NAEP Framework Update

[Home | NAEPFrameworkUpdate](https://naepframeworkupdate.gov)

