IES Automated Scoring Challenge for NAEP Data: Request for Information Webinar

October 4, 2021

Eunice Greer, Sr. Research Scientist
National Center for Education Statistics
Mark Shermis, Sr. Consultant
John Whitmer, Sr. Fellow
Institute of Education Sciences

NAEP
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

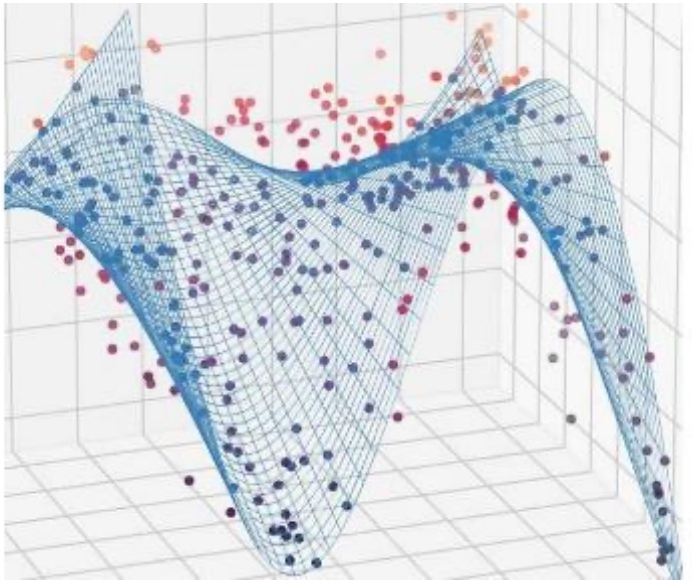
Agenda

1. Introduction - John Whitmer
2. NAEP Reading Assessment - Eunice Greer
3. Challenge Design & Transparency Requirements - John Whitmer
4. Dataset Structure and Scoring - Mark Shermis
5. Q&A

1. Introduction

Goals for Automated Scoring in NAEP

1. Increased information about respondents
2. Speed to return scores
3. Cost reduction compared to human scoring



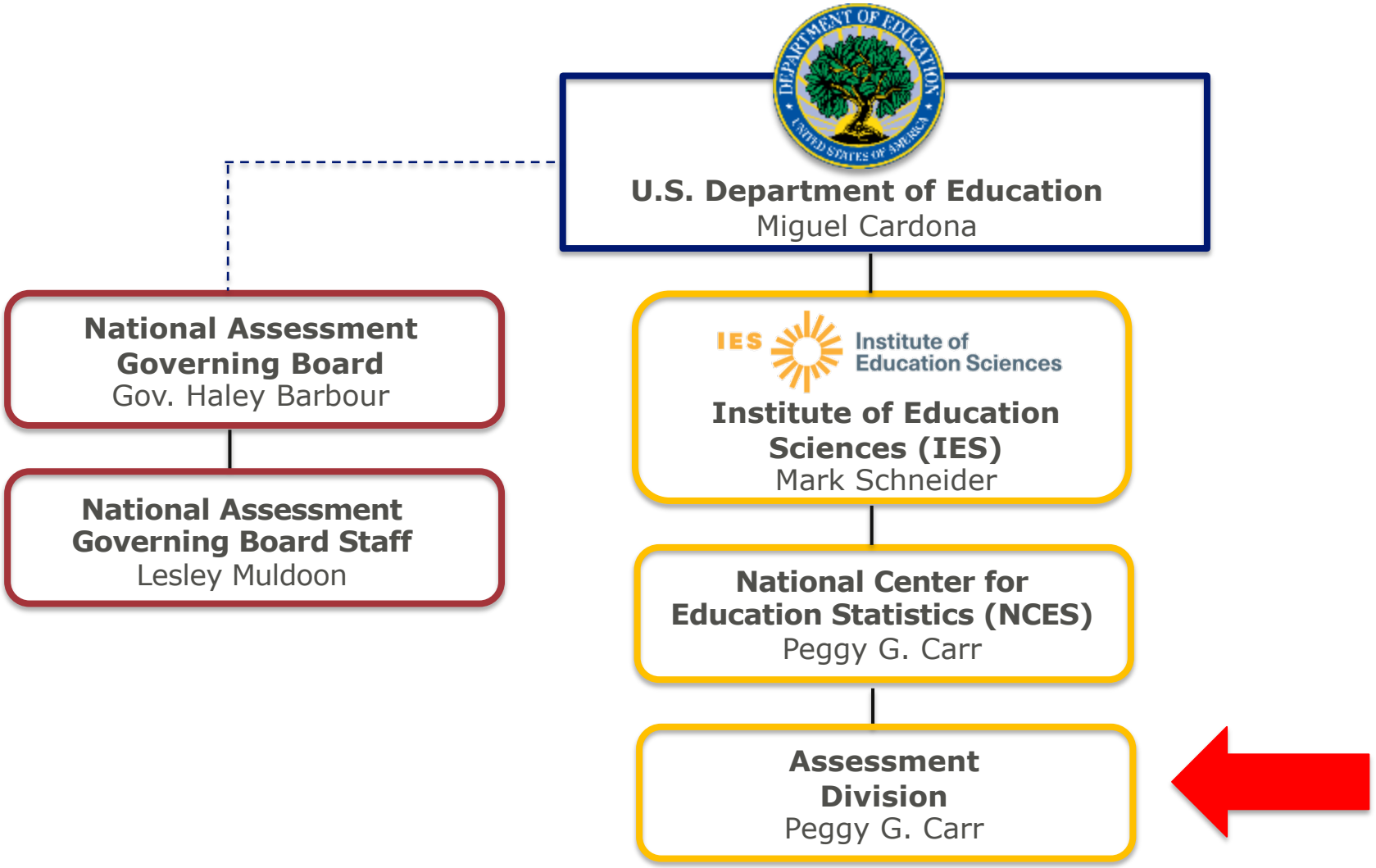
Source: Tech Xplor

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.
- NCES Confidential Data Security application required to participate: **DEADLINE 10/20/2021** (docs available online)
- Response deadline (technical report & predictions): 11/28/2021
- Winner Announced: 12/16/2021

2. NAEP Reading Assessment

Governance


U.S. Department of Education
Miguel Cardona
National Assessment Governing Board
Gov. Haley Barbour
National Assessment Governing Board Staff
Lesley Muldoon
IES Institute of Education Sciences
Institute of Education Sciences (IES)
Mark Schneider
National Center for Education Statistics (NCES)
Peggy G. Carr
Assessment Division
Peggy G. Carr

Main NAEP: Subjects Assessed

Mathematics

Reading

Writing

Science

Civics

Geography

U.S. History

Economics

Vocabulary

Music

Visual Arts

Technology and Engineering Literacy

Main NAEP: Subjects Assessed

Mathematics

Reading

Writing

Science

Civics

Geography

U.S. History

Economics

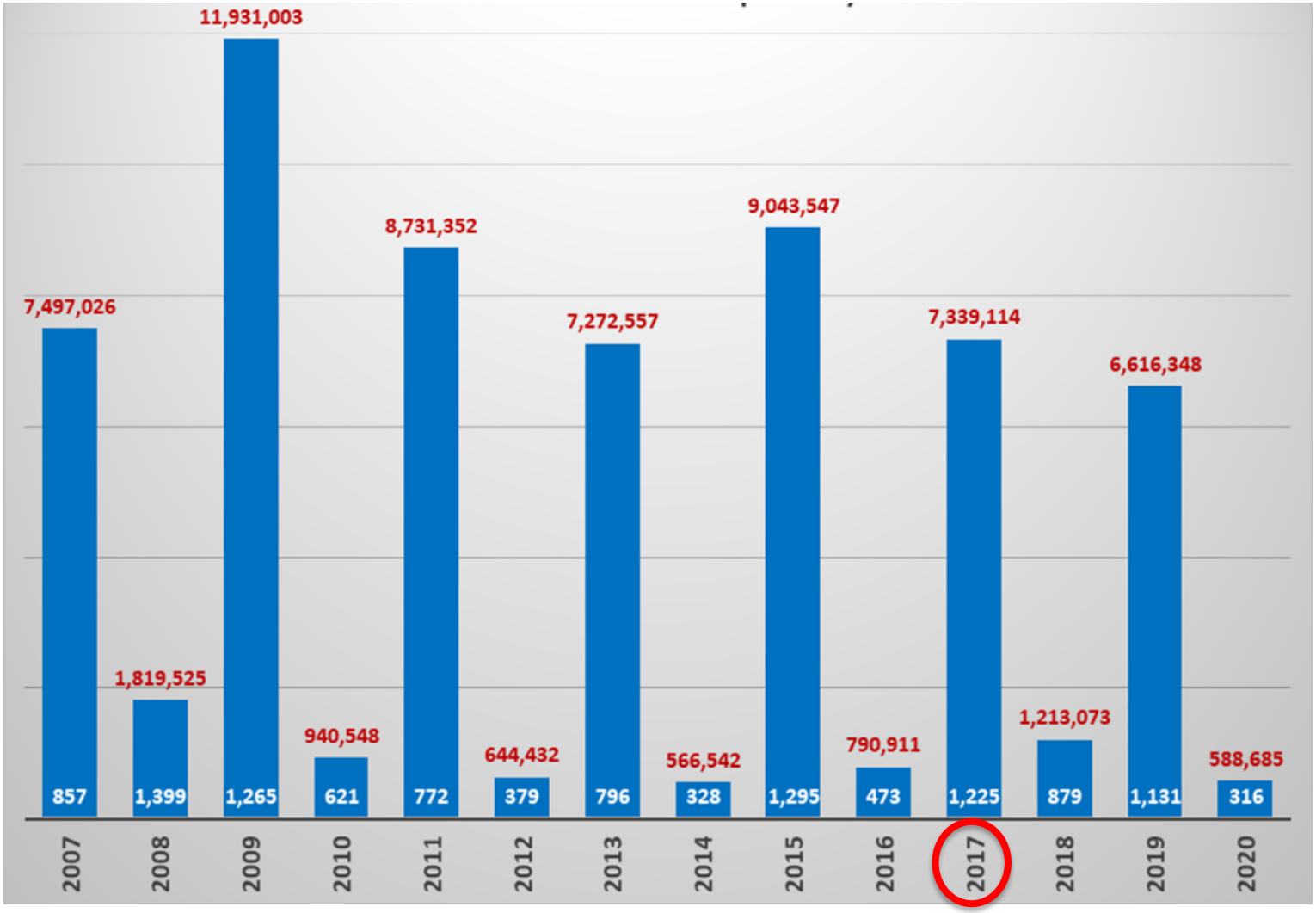
Vocabulary

Music

Visual Arts

Technology and Engineering Literacy

Number of Constructed Response Items, by Year



Definition of Reading that guides NAEP Reading Assessment

Conceptualizes reading as a dynamic cognitive process

Reading is an active and complex process that involves:

- Understanding written text;
- Developing and interpreting meaning;
- Using meaning as appropriate to type of text, purpose, and situation.

This definition applies to the assessment of reading achievement on NAEP and is not intended to be an inclusive definition of reading or of reading instruction.

Structure of NAEP's Reading Assessment

Each student completes two blocks of the assessment. Each block includes:

One to two passages, and

9 – 13 questions about the passage content.

Characteristics of Passages

- Taken from authentic texts
- Reflective of cultural diversity
- Set in contexts familiar to students
- Usually presented in entirety
- Vary in length by grade:

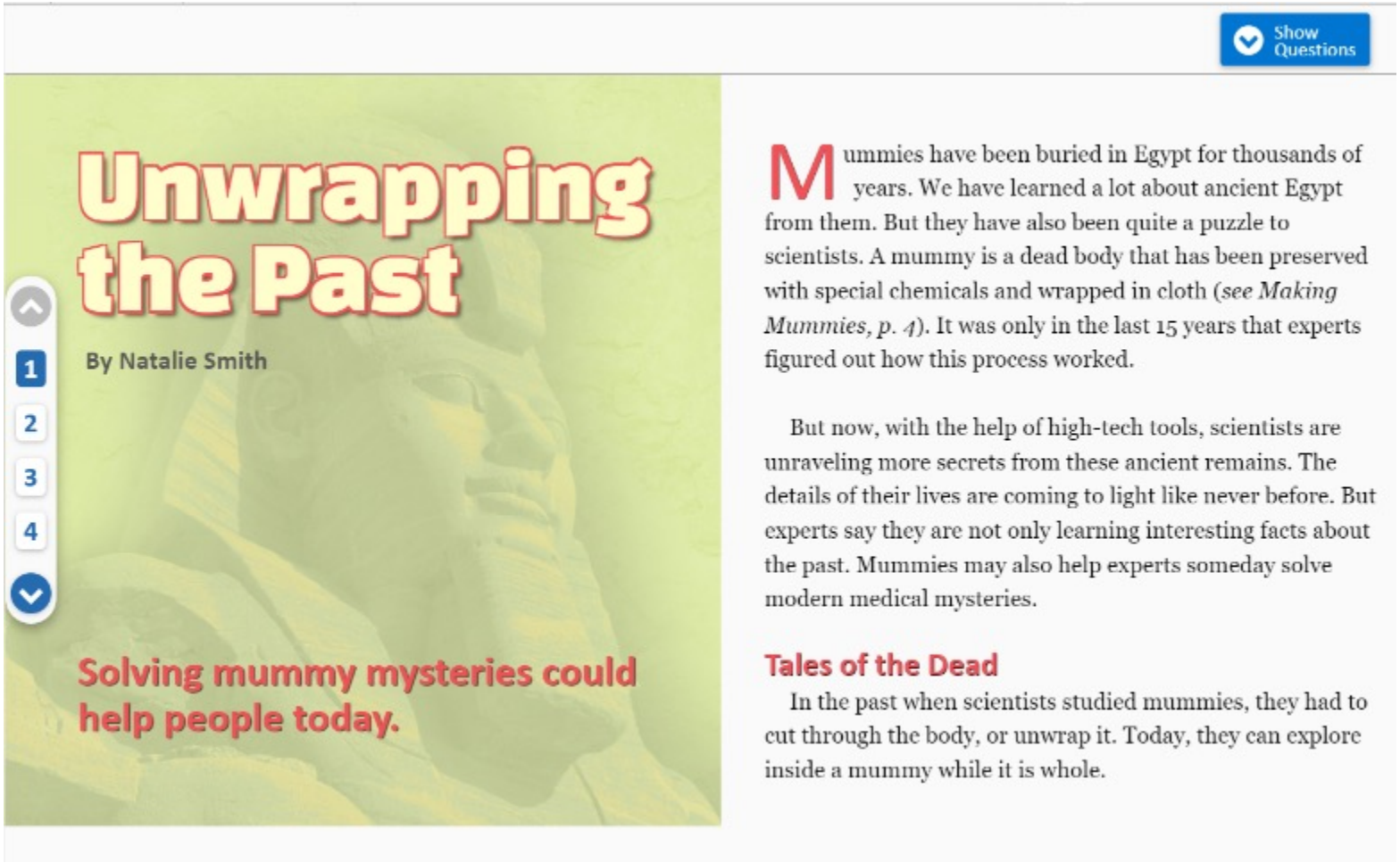
Grades	Range in Words
4th grade	200 – 800 words
8th grade	400 – 1,000 words
12th grade	500 – 1,500 words

Types of Passages

- **Literary**
  - Fiction
  - Literary Nonfiction
  - Poetry

- **Informational**
  - Exposition
  - Argumentation and Persuasive Text
  - Procedural Texts and Documents

Sample 8th Grade Informational Text



Show Questions

Unwrapping the Past

By Natalie Smith

1 2 3 4

Solving mummy mysteries could help people today.

Mummies have been buried in Egypt for thousands of years. We have learned a lot about ancient Egypt from them. But they have also been quite a puzzle to scientists. A mummy is a dead body that has been preserved with special chemicals and wrapped in cloth (*see Making Mummies, p. 4*). It was only in the last 15 years that experts figured out how this process worked.

But now, with the help of high-tech tools, scientists are unraveling more secrets from these ancient remains. The details of their lives are coming to light like never before. But experts say they are not only learning interesting facts about the past. Mummies may also help experts someday solve modern medical mysteries.

Tales of the Dead

In the past when scientists studied mummies, they had to cut through the body, or unwrap it. Today, they can explore inside a mummy while it is whole.

Sample 8th Grade Informational Text

To Everything There Is a SEASON

Fresh-picked food is just plain good.

By Melinda Hemmelgarn

Strawberries in January, peaches in March, tomatoes in December. Unless you live in an area with a very long growing season, all of the above violate the laws of eating naturally—in other words, eating in season.

When we eat in rhythm with the seasons, we can appreciate Earth's natural cycles. Let's consider the peach. That fuzzy fruit defines summer. Fruits taste best and reach their nutritional peak when picked ripe and eaten shortly after harvest. We can buy imports from Chile all winter long, but out-of-season peaches lack fragrance and the sweet juice that drips down our chins.

Feasting on Fossil Fuel

Our global food system allows us to eat just about anything we want, any time of year. However, choosing foods grown and harvested thousands of miles away takes its toll on our

Sample 8th Grade Informational Text

planet. For example, long-distance trucking to transport food from faraway places requires fossil fuel, adding hidden costs, such as global warming. "Seasonal eating is environmental eating," explains David Bruce, a Wisconsin organic farmer.

"We are the only species that can protect our planet," says Kathy Cobb, a consultant to the Centers for Disease Control and Prevention's National Fruit & Vegetable Program. Cobb knows fruits and vegetables help us stay fit and healthy, and you probably do too. But, she says, there are environmental benefits of eating local seasonal produce.

"When we eat food that is planted and grown locally during each of the four seasons, we allow the earth and soil to replenish itself, and reduce harmful effects on the environment caused by transporting food long distances," Cobb says.

Stashing fruits and vegetables in a refrigerator may help reduce nutrient losses. But it's better to get the produce from the plant to your plate pronto.

Healthy for Earth—and for You

Nourishing ourselves "goes beyond just filling our bellies," according to registered dietitian Amanda Archibald. She favors seasonal foods because of their overall quality. "If you use the season as your guide, you will always get the best flavor and nutrient content."

Items Assess Three Cognitive Targets

"Cognitive targets" refer to the types of thinking required to answer the question:

- **Locate and recall** information from text
- **Integrate and interpret** information and ideas presented in text
- **Critique and evaluate** information and ideas presented in text and the ways in which authors present text

All passages are "mapped" to identify purpose, organization and key content. These maps guide item writing.

Locate/Recall

Locate/Recall items may ask students to:

- Identify explicitly stated information, including main ideas or supporting details
- Find essential elements of a story, such as characters, time, or setting
- Make very simple inferences within or across a few sentences

Locate/Recall (cont.)

Locate/Recall items often:

- Involve matching information in the item to literal or synonymous information in text in order to choose the correct answer or construct a response
- Focus on information contained in relatively small amounts of text

Sample Locate/Recall Item

According to the article, why is it important to get produce "from the plant to your plate" as quickly as possible?

Sample Locate/Recall Item: Scoring Guide

Acceptable

Responses at this level provide a reason from the article that explains why it is important to get food from plant to plate as quickly as possible.

- *It's important to get produce "from the plant to your plate" quickly because fruits and vegetables picked too early can't develop their full flavor and the extra time needed to ship food from the farm cuts nutrients.*
- *When you eat locally grown fruits and vegetables, it makes less pollution going into the atmosphere than it would by shipping the food from a different country.*
- *Because fresh tastes better and is more healthy for you.*

Unacceptable

Responses at this level provide irrelevant details or personal opinions, or may simply repeat the question.

- *Because if not, the produce will spoil.*
- *I really like to eat fresh fruits and vegetables.*
- *It is important because it's best to get fresh food quickly.*

Integrate/Interpret

Integrate/Interpret items may ask students to:

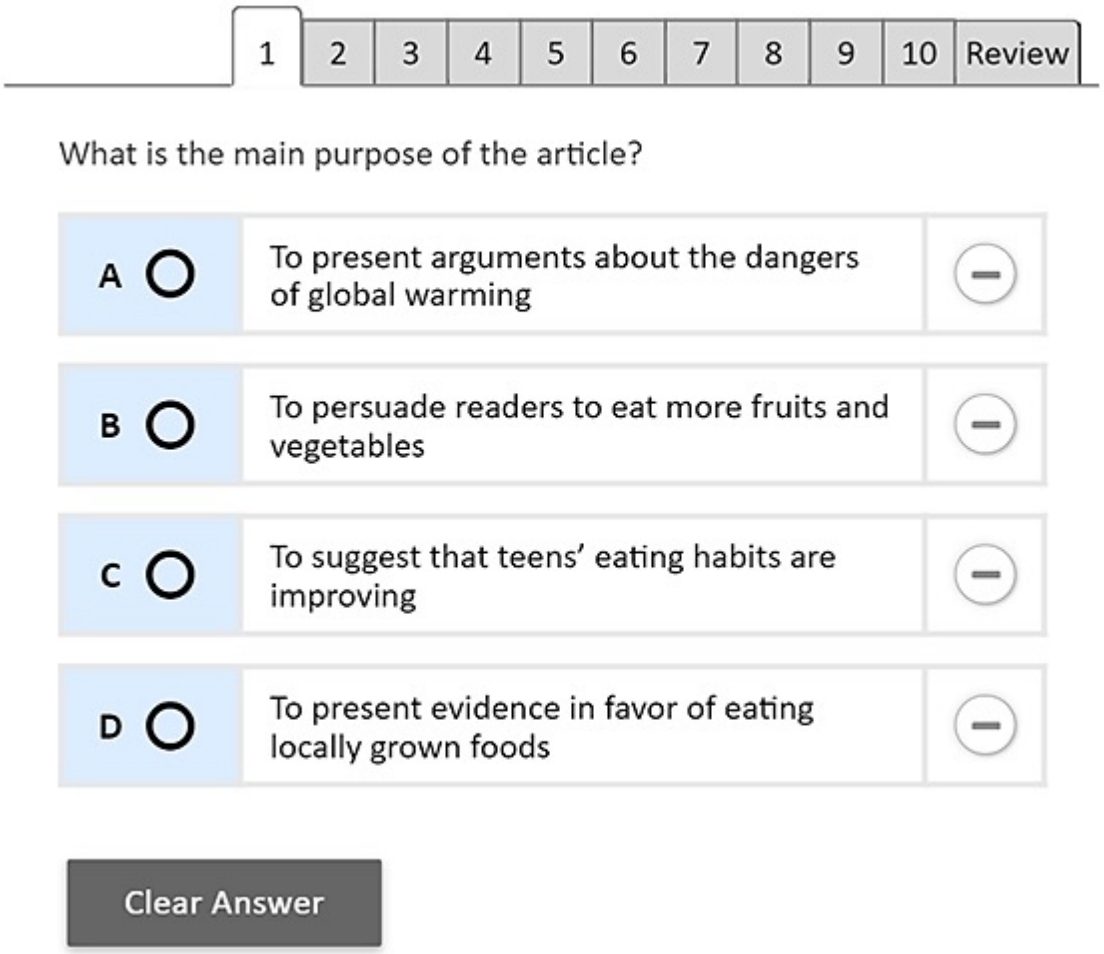
- Compare/contrast information or character actions
- Examine relations across aspects of text
- Make causal connections
- Consider alternatives to what is presented in the text

Integrate/Interpret (cont.)

Integrate/Interpret items often:

- Involve thinking across large portions of text, across the text as a whole, or even across multiple texts
- Require students to apply what they know to what they are reading

Sample Integrate/Interpret Item



Critique/Evaluate

Critique/Evaluate items may ask students to:

- View the text critically by examining it from numerous perspectives
- Evaluate overall text quality or the effectiveness of particular aspects of the text
- Judge the author’s craft
- Evaluate the author’s perspective or point of view

Critique/Evaluate (cont.)

Important attributes of Critique/Evaluate items:

- Items go beyond the text
- Require that the student evaluate the effectiveness of some aspect of the text
- Involve processes of:
  - Analysis
  - Critique
  - Judgment
  - Support

Sample Critique/Evaluate Item

What are two types of evidence that the author uses in the article to support her argument? Explain why using both types of evidence helps to strengthen the author's argument.

Sample Critique/Evaluate Item: Scoring Guide

Extensive

Responses at this level identify two types of evidence that the author uses in the article to support her argument and explain why using both strengthens her argument.

- *The author used teenagers and facts to defend her argument. Using teen taste testers helps because she has opinions of others, not just her own, as support that locally grown produce tastes better. Facts help support her argument because she is proving that not only do the fruits taste better but they also have more nutrients than canned and processed fruits.*
- *One type of evidence that the author uses is quotes from a registered dietitian. This helps with her argument by giving the sense of a respected practice reassuring her argument. The second is the use of teenage kids. The author used their opinions so the reader can relate to the people in the story.*

Essential

Responses at this level identify one or two types of evidence that the author uses in the article but only explain why using one helps strengthen her argument.

- *The two types of evidence are statistics and pictures. The pictures strengthen her argument because they make the idea of eating fresh and local food look fun and easy to do.*
- *It was a good idea to include a bunch of quotes from kids. This makes the article easier to relate to and makes you want to agree with the author about eating food that is in season.*

Sample Critique/Evaluate Item: Scoring Guide (cont.)

Partial

a) Responses at this level identify one or two types of evidence that the author uses in the article to support her argument but do not explain why using either helps strengthen her argument.

- *She uses testimonies that give other people's opinions. Another type of evidence is a table that shows when different fruits are in season. The author says on page 3 that fresh-picked food is just plain good.*
- *The author includes a lot of facts to support her argument. On page 4 there is a list of the many ways fresher is better. I think she must have done a lot of research.*
- *The author uses personal testimonies from teenagers. These help strengthen her argument.*

OR

b) Responses provide content from the article and explain why including the content helps strengthen her argument.

- *The types of evidence are that locally grown fruits taste better and they have more nutrients. This evidence helps strengthen her argument by using people's real life experiences to convince her readers.*
- *One type of evidence is that "long-distance trucking requires fossil fuel". This helps strengthen her argument because it gets the attention of people who care about the environment.*

Unsatisfactory

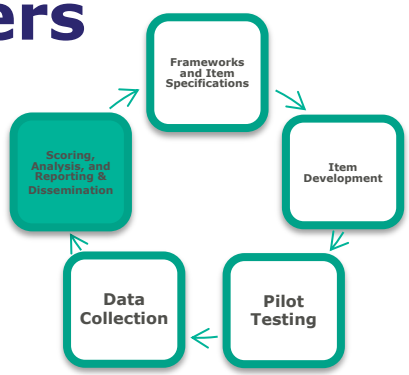
Responses at this level may summarize the main points in the article but do not explain why the author's inclusion of the content helps strengthen her argument. Responses also may provide irrelevant details or unsupported personal opinions, or may simply repeat the question.

- *The author says on page 3 that fresh-picked food is just plain good.*
- *She says when we eat in rhythm with the seasons we can appreciate the Earth's cycles.*
- *I think the author makes a good argument in this article.*

Scoring Process

NCES **scores multiple-choice items electronically** and employs **human scorers for short and extended constructed-response items**

- Develops **scoring guides** that match criteria in assessment frameworks
- Recruits and trains qualified **scorers**
- Monitors **scoring consistency**



3. Challenge Design & Transparency Requirements

Two Components for Challenge

- Component A - Item-Specific Models
  - The first-place prize for this challenge is $15,000, with up to 4 runner-up prizes of $1,250 each.
  - These models are anticipated to reach parity with IRR of human scoring.
- Component B - Generic Models
  - The prize for this challenge is $5,000, with up to 4 runner-up prizes of $1,250 each.
  - These models are not anticipated to reach parity with IRR of human scoring (but could be useful as QA tools or formative feedback).

Item-Specific Models

- Use NLP features & machine learning to predict score to student responses for each of 18 items.
- Most responses are relatively brief, and participants will be provided with reading passage, scoring rubric, and other information used by raters.

Generic Models

- Create a model to score responses to two items (one 4th grade, another 8th grade) without access to only the responses.
- Data from item-specific model component and/or external data or algorithms should be used for model training.
- Answers the question: can you create a model that "generalizes" from similar data to a new item, without seeing that prompt or responses?

Participants may enter submissions for either or both

Submission Requirements

Valid submissions will include:

- A technical report that provides model interpretability.
- Predicted scores (CSV format) from the test data responses.
- For commercial entities, a pricing sheet that includes actual costs related to the production implementation of automated scoring.

Evaluation Criteria

- **Part 1: Model Interpretability**. Submissions must provide a technical report that explains the model development process and results appropriate to a technical audience with educational measurement expertise.

  It is not expected that competitors will reveal confidential information, but will provide evidence that enables an external reviewer to assess the validity and fairness of the automated scoring process and models.

  These reports will be submitted with submissions of predicted scores, but must be approved as providing a sufficient degree of interpretability per the criteria below before the response predictions will be evaluated.

- **Part 2: Model Accuracy**. (explained by Mark Shermis)

Interpretability Criteria

a) **Transparency** -- explanation of the ***process for model training and testing***, the features extracted from the text, and the algorithms used in model building. While these may describe a general workflow, they should also include the specific text features and algorithmic choices used to create the final models that score items in this Challenge.

b) **Explainability** -- explanation of the ***resulting item model and/or individual scores*** that includes the input features considered, the modeling results, and algorithm choices.

c) **Fairness** -- analysis into any **differences based on student demographic background** in automated scoring compared to those found in human-scored results.

Timeline

**	Item	Duration (D)	Start	Finish
2.1	Challenge posting period	30	16-Sep	20-Oct
2.2	Request for information webinar		4-Oct	4-Oct
2.3	**Application deadline ***			**20-Oct**
2.4	Provide dataset			28-Oct
2.5	Competitors prepare responses*	30	29-Oct	28-Nov
2.6	**Response deadline**			**28 Nov**
2.7	Select winner			16-Dec

4. Dataset Structure and Scoring

Item Specific Competition

Each item has four files associated with it:

1. Pdf of the item itself, scoring guide, and anchor or scored sample items.
2. A training file in csv format
3. A cross-validation file in csv format
4. A test file in csv format

Training, Cross-Validation, and Test Files

- All in CSV format. No corrections applied to any responses.
- The training file contains double-scored items and the text response of the writer. The first score is the score of record. The second score was used for computing human-rater reliability.
- The cross-validation file has a small sample of double-scored responses and large sample of single scored responses along with the text of the item. It is used to help competitors cross-validate their models based on the training file, and for use in calculating a leaderboard. Note that there are non-scored responses contained in this file. These should be treated as missing data.
- The test file has only the text of the item. Competitors are asked to provide a score prediction for at least 99.5% of the responses.

Demographic Information & Word Count

- Information on Gender and Race is provided for each respondent
- This information can be used to determine the mean difference between a reference and focal group divided by their pooled standard deviation. If this result is > .15, we would conclude that there is evidence of differential item functioning (DIF). For example,
- $\frac{M_{Males} - M_{Females}}{SD_{pooled}} > |.15|$ then evidence of DIF
- Competitors can self-check their predictions on the training and cross-validation sets in writing their technical reports
- Information on Word Count is an estimate only. This estimate may be influenced by misspellings and typographical errors.

Scoring

- Each response in the test file has a valid human-rater response. Competitors are expected to score at least 99.5% of the responses for each test file item.
- Winners will be determined by quadratic weighted kappa (qwk) to the third decimal place. QWK is calculated as:
- $\kappa_w = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$
- where *k* = number of codes and w_{ij}, x_{ij}, and m_{ij} are elements in the weight, observed, and expected matrices, respectively.
- see https://www.kaggle.com/aroraaman/quadratic-kappa-metric-explained-in-5-simple-steps for examples

Generic Competition

Each item has two files associated with it:

1. Pdf of the item itself, scoring guide, and anchor or scored sample items.
2. A test file in csv format

Test File, Demographic Information, Word Count, & Scoring

- The test file has only the text of the item. Competitors are asked to provide a score prediction for at least 99.5% of the responses. You are permitted to use any authorized external data sets, databases, or even data from other items in this competition to model your predictions. The technical report should document what information you used to develop your model.
- As with the item specific competition, demographic information is provided so that you can check for potential differential item functioning.
- Information on Word Count is an estimate only. This estimate may be influenced by misspellings and typographical errors.
- Scoring is done via quadratic weighted kappa, as explained in the item specific slides.

5. Q&A

Participate!

- Challenge URL: https://www.challenge.gov/challenge/naep-automated-scoring-challenge/
- Deadline for Security Applications: 10/20/2021
- Deadline for Submissions: 11/28/2021
- Questions: automated-scoring-challenge@ed.gov

Useful Links

NAEP Reading Framework
https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/reading/2017-reading-framework.pdf

NAEP Item Maps
https://www.nationsreportcard.gov/itemmaps/

NAEP Questions Tool
http://nces.ed.gov/nationsreportcard/itmrlsx/

NAEP Tutorials on the Web
https://nces.ed.gov/nationsreportcard/dba/

2017 DBA Sample Questions
https://www.nationsreportcard.gov/reading_2017/sample-questions?grade=4
https://www.nationsreportcard.gov/reading_2017/sample-questions?grade=8

General Information on the NAEP Reading Assessment
https://www.nationsreportcard.gov/reading_2017/about/framework?grade=4

2025 NAEP Framework Update
https://www.naepframeworkupdate.org/


Thank You