Committee on Standards, Design and Methodology

November 16, 2023

4:00 pm – 6:00 pm EST

Hemingway Salon 2

AGENDA

4:00 – 4:05 pm	**Welcome and Updates** *Suzanne Lane, Chair*	
4:05 – 4:55 pm	**Automated Scoring: Math Challenge and Shadow Scoring** *Eunice Greer, Senior Research Scientist, NCES* *Edward Wolfe, Director of Automated Scoring, Pearson*	Attachment A
4:55 – 6:00 pm	**Continued Discussion: Plans for Device Agnostic Administration (CLOSED)** *Enis Dogan, Chief Psychometrician, NCES* *Ranu Palta-Upreti, Project Director - NAEP Platform Development, Educational Testing Service*	Attachment B

Automated Scoring Updates: Math Challenge and Shadow Scoring

November 16, 2023

The National Center for Education Statistics (NCES) has conducted ongoing work exploring the use of automated scoring for constructed response items for NAEP Reading and Math. In 2021 NCES held an automated scoring contest to understand the feasibility of automated scoring of reading constructed response items. NCES briefed the Committee on Standards, Design and Methodology (COSDAM) on the findings of this contest in March of 2022. Information about this challenge is available in appendix A-2. In 2023 NCES continued to progress in studying automated scoring of reading items, including a recent shadow scoring study that examined the comparability between AI and human scores, and held an automated scoring challenge for math constructed response items (see Attachment A-3 for information about the math challenge). The purpose of the November 2023 session is to provide an overview of the progress towards automated scoring that has occurred since the March 2022 COSDAM meeting.

Background

Selected-response items are those for which students respond to an assessment question by selecting from a set of options from which they choose one or more correct answers. This includes traditional multiple-choice like questions, and digitally enhanced items such as drag-and-drop. Selected-response items are straight-forward to score through automation involving high-speed scanners for paper-pencil administrations, and through digital methods for digital assessments. NAEP has been using this method of scoring for selected response items for decades.

Constructed-response items require students to generate their own response (e.g., by inputting information in a text box) and can vary in length from a phrase to several sentences. These items are scored by trained scorers using a rigorous and standardized process to ensure ongoing accuracy. This process requires many people and significant time. More information on current NAEP scoring practices is available here.

Advancements in artificial intelligence (AI), and more specifically natural language processing (NLP)[1], show promise for potentially increasing the efficiency of the scoring of constructed-response items by automating some of the process. NCES is currently exploring the accuracy of different AI models for scoring NAEP reading and math items,

[1] Natural language processing (NLP) is described by IBM as “the branch of artificial intelligence or AI concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.”

and identifying the limitations (e.g., bias and sensitivity concerns, item response types that are difficult to score) and how to best address them.

Attachment A-2

National Assessment of Educational Progress (NAEP) Automated Scoring Challenge

This past fall, NCES held its first automated scoring challenge to score constructed response items for the National Assessment of Educational Progress's reading assessment. The purpose of the challenge was to help NCES determine the existing capabilities, accuracy metrics, the underlying validity evidence, and costs and efficiencies of using automated scoring with the NAEP reading assessment items. The Challenge required that submissions demonstrate interpretability of models, provide score predictions using these models, analyze models for potential bias based on student demographic characteristics, and provide cost information for putting an automated scoring system into operational use.

The challenge was announced and posted on Challenge.gov.

Start data: 9/16/2021
End date: 11/28/2021

A Request for Information Webinar was held 10/4/2021. Approximately 50 persons attended.

25 teams registered for the challenge, submitted the requires non-disclosure agreements and requested data. Teams included commercial entrants, university teams, and independent teams. 17 teams were domestic, 8 were international. While the majority of the teams were comprised of graduate-level data scientists and statisticians, one local team included a high school student.

Description

Automated Scoring using natural language processing is a well-developed application of artificial intelligence in education. Results are consistently demonstrated to be on-par with the inter-rater reliability of human scorers for well-developed items (Shermis, 2014). Currently, the National Assessment of Educational Progress (NAEP) makes extensive use of constructed response items. Annually, contractors assemble teams of human scorers who score millions of student responses to NAEP's assessments.

This challenge sought to ascertain whether a wide array of automated scoring modes could perform well with a representative subset of NAEP Reading, constructed response items administered in 2017 to students in grades 4 and 8. The ultimate goal was to produce reliable and valid score assignments, provide additional information about responses (e.g. length, cohesion, linguistic complexity), and generate scores more quickly while saving money on scoring costs.

There were two components to this challenge; entrants could submit responses to one or both of these components:

1. Component A - Item-Specific Models: Respondents were asked to build a predictive model for each item that could be scored, using current state-of-the-art practices in operational automated scoring deployments. Extensive training data from prior human scoring administrations was provided. The first-place prize for this challenge is $15,000, with up to 4 runner-up prizes of $1,250 each.

2. Component B - Generic Models: Respondents were asked to build a generic scoring model that could score items that were not included in the training dataset, but were from the same administration, subject, and grade level. The prize for this challenge is $5,000, with up to 4 runner-up prizes of $1,250 each.

Participants were provided access to digital files that contain information related the results of human-scored constructed responses to reading assessment items that were administered in 2017, including item text, passages, scoring rubrics, student responses, and human assigned scores (both single and double scored). The responses correspond to items that accompany two genres of 4th and 8th grade reading passages, literary and informational. Items for this challenge are of two response formats, short and extended.

The data set included 20 items for the item-specific models, and 2 items for the generic models. There was an average of 1,181 double-scored responses per dataset. These were divided into a training dataset (50%), a validation dataset (10%), and a test dataset (40%). The validation dataset was augmented with a much larger number of single-scored responses (average 23,000 per item).

In addition to model accuracy compared to human scorers, successful respondents to this Challenge had to provide documentation of model interpretability through a technical report that was evaluated by NCES's team of scorers for transparency, explanability, and fairness. The documentation was evaluated before respondents' scored submissions were evaluated. Only documentation that met acceptance criteria were considered as valid submissions and evaluated for accuracy of the predicted scores compared to the hold-out test dataset. The Federal Government is particularly interested in submissions that provided accurate results and met these objectives, as they have been absent from a good deal of recent research in automated scoring, particularly for solutions using artificial intelligence (e.g. neural networks, transformer networks) and other complex approaches (Kumar & Boulanger, 2020).

This process is consistent with the operational processes that the Department intends to use as part of the approval process for scoring and reporting; only models that can provide substantive validity evidence would be approved for operational use.

Of the 25 teams that registered, 15 completed the challenge and submitted the required work to be judged. Three submissions did not meet the acceptance criteria and were eliminated from competition.

In January 2022 NCES announced that four teams had won top honors in the Challenge. They are *Measurement Incorporated, the University of Massachusetts-Amherst, Cambium Assessment, and the University of Duisburg-Essen. In addition to* awarding the four grand prizes, NCES recognized four runner-up teams, as well. The winners used advanced natural language processing methods that promise to reduce scoring costs while maintaining accuracy similar to human scoring.

Natural language processing uses computer algorithms to identify patterns in language; automated scoring applies these patterns to analyze student responses and assign scores. Those scores are then compared to the scores for each response given by human graders. The most accurate submissions used advanced machine learning approaches based in what are called "transformer network architectures" such as BERT (or "Bidirectional Encoder Representations from Transformers"). These models used NAEP data to fine tune pre-trained language models that were created by analyzing language consistencies and patterns among billions of student writing examples.

This challenge is a key component in modernization efforts to incorporate data science and machine learning into operational activities at NCES. It is the first in a series of challenges that use NAEP data.

Winners

Grand Prizes
Arianto Wibowo, Measurement Incorporated (Item-Specific Model)
Andrew Lan, UMass-Amherst (Item-Specific Model)
Susan Lottridge, Cambium Assessment (Item-Specific Model)
Torsten Zesch, University of Duisburg-Essen (Generic Model)

Runners-up
Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education, Centre for Technology-Based Assessment (Item-Specific Model)
Scott Crossley, Georgia State University (Item-Specific Model)
Prathic Sundararajan, Georgia Institute of Technology and Suraj Rajendran, Weill Cornell Medical College (Item-Specific Model)
Susan Lottridge, Cambium Assessment (Generic Model)

NAEP Math Item Automated Scoring Data Challenge Results: High Accuracy and Potential for Additional Insights

Challenge Overview

In Spring 2023, NCES hosted a data challenge to see how automated scoring techniques compared to humans when scoring open-ended responses to NAEP mathematics test questions.

Open-ended math items can tell us how students approach math problems, not just whether they can answer correctly. However, scoring math responses is difficult for artificial intelligence methods like natural language processing because it combines specific calculations with conceptual information.

Humans score these items very accurately. The purpose of the challenge was to tell us whether automated scoring for mathematics responses could be equally accurate and whether NAEP could use these methods in the future.

Over a dozen teams participated in the Challenge, and three teams earned awards.

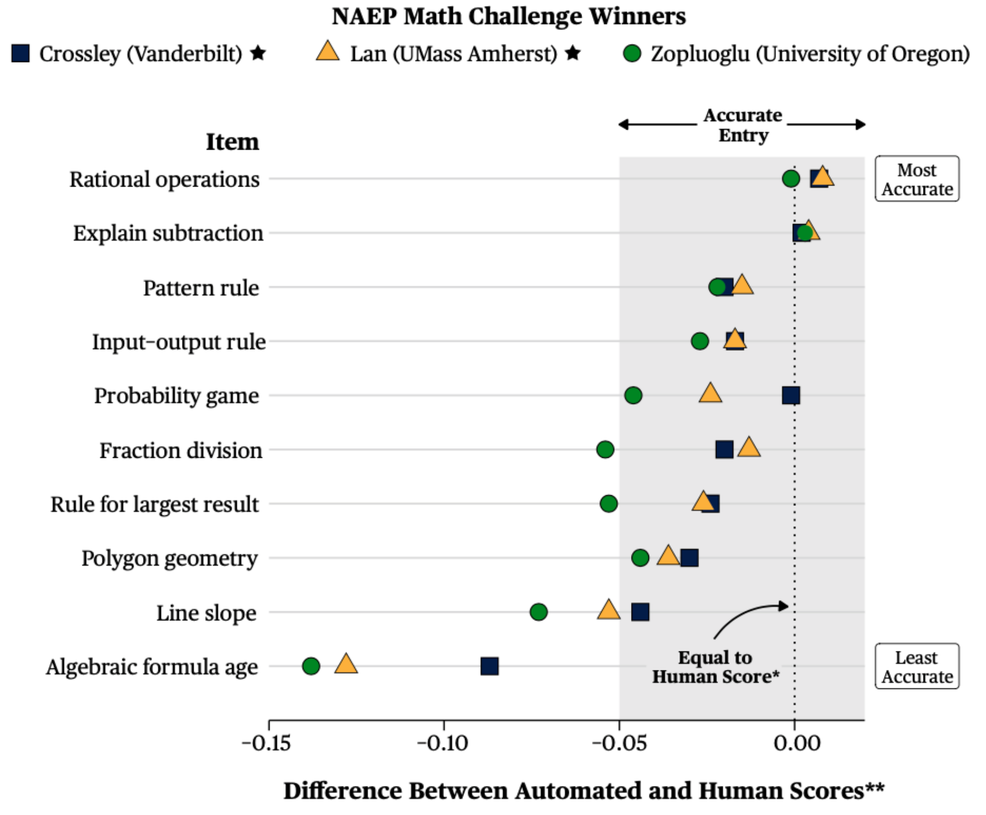
Two teams earned grand prizes: UMASS Amherst, led by Dr. Andrew Lan; and Vanderbilt University, led by Dr. Scott Crossley. One team earned a runner-up prize: University of Oregon, led by Dr. Cengiz Zopluoglu.

Judges first evaluated technical reports, which described the methods used for scoring. If they met transparency and fairness analysis requirements, their entries were analyzed for accuracy and for whether bias was observed in the teams' predictions.

Implications for NAEP

- **Automated scoring can accurately score open-ended math items.** The use of automated scoring should be determined at the item level, ensure accuracy before using, and include fairness analyses. Automated scoring methods can save time and money, allowing for deeper analysis of the data.
- **Automated scoring has the potential to expand the usefulness of NAEP testing.** It can provide additional insights about item-level performance and increased diagnostic information about respondents. These insights can help districts better understand student performance and help NAEP improve the design of future tests.
- **Automated scoring can be fair and unbiased when properly implemented.** Advanced fairness analysis can ensure results do not exhibit bias in scoring. This issue is required for all NAEP results and can be achieved.

Accuracy Results by Item by Team



★Grand prize winners
*The machine and the human scorer agreed to the same extent as two human scorers.
**Scores were measured in terms of average Quadratic Weighted Kappa (QWK).

Key Takeaways

- **Accurate scoring required responses beyond the text.** Significant data pre-processing was required. This included things like correcting spelling mistakes and using information students provided in other parts of the question to evaluate their response. This process is also used by human raters.
- **Items were consistently easy/hard to score for all teams and approaches.** Despite using many different types and approaches to modeling, teams had relatively consistent accuracy across items. While some items had a clear cause for inaccurate results (e.g., 94% incorrect responses), the reasons other items were difficult to score was less clear. Item content or presentation could be a problem to examine in such items. However, only one item could not be scored accurately.
- **Large language models (LLMs) performed better than other approaches.** LLMs consider the context beyond isolated words, which helps extract greater meaning from student writing. All but one entry used an LLM. The team that did not use an LLM did not score a single item within accuracy thresholds. None of the teams used the more popular LLMs (e.g., ChatGPT) due to privacy restrictions.

- **Results did not exhibit bias, unlike reading predictions.** Predicted scores were extremely accurate overall and analyses for subpopulations did not find substantive differences by subpopulations identified in NAEP (e.g., English Learners, Race/Ethic groups, Sex, IEP status). In the Reading Challenge, there were some items in which significant bias was observed for English Learners, which would be identified prior to the use of any model in an operational administration.

Methods Used Summary

#	Team	Summary of Approach
1	**S. Crossley & LEAR Lab (Vanderbilt)**	This approach first recognized that the data were imbalanced in favor of scores of 1 (incorrect), so the authors decided to use a Stochastic Gradient Descent classifier to filter out many of the responses with a score of 1. Additionally, to increase samples of writing receiving 2s and 3s, the authors included augmented high-scoring paraphrases as well as data from additional columns to augment the written responses. The authors used the DeBERTa V3 Large model to carry out their predictions.
2	**A. Lan (UMASS-Amherst)**	This research group first corrected the spelling and then represented the additional variables within questions as part of the scored item. The authors concluded that input text with a mixture of structural aspects and some textual representation led to the highest Kappa scores. The authors also used several LLMs but found that the Flan-T5 system worked best for these data.
3	**C. Zopluoglu (University of Oregon)**	This method used spelling correction and other preprocessing steps to prepare the data. The author also created exemplary written responses for each item and then used cosine similarity to measure how close each student response was to these exemplars alongside sentence embeddings. The author also investigated 18 different transformer-based LLMs for each item for a total of 180 models explored. Different models worked best for different items, but Math-RoBerta was the most accurate for the most items (4/10 were scored using Math RoBerta).

NAEP Modernization: Plans for Device Agnostic Administration

November 16, 2023

The Governing Board has received high-level updates on modernization plans in recent months, including plans to move towards a device agnostic administration where NAEP would be administered on school-owned devices. There are many technical considerations before moving to a device agnostic administration to ensure scores remain valid and reliable across devices, and that mitigate risks to trend. COSDAM received a presentation by the National Center for Education Statistics (NCES) regarding steps towards a successful transition in August 2023. COSDAM members requested a second presentation in November 2023 to continue the discussion and receive additional information on item display comparability across devices.

Background

The NAEP program is considering multiple changes in administration practices to move towards a more efficient and modern assessment program. One such change is to move towards a device agnostic administration. Since NAEP first became a digital assessment, NCES has provided all necessary equipment, currently Microsoft Surface Pro tablets with styluses. The assessment administrators carry large cases, called pelican cases, that hold all equipment required to administer the assessment and to transmit the assessment data without accessing a school's Wi-Fi or using their computers. In 2024 NAEP assessments will be administered on NAEP-provided Microsoft Surface Pro tablets and Google Chromebooks. Bridge studies are planned to evaluate score comparability between Surface Pro and Chromebook devices and the feasibility of linking results collected from this hybrid device mode to the existing trendlines. Chromebooks were chosen because of their widespread use in classrooms and their lower cost. In 2026 NCES plans to administer NAEP on school-based equipment that meet NAEP's technology requirements, with NAEP-provided Chromebooks as backups where needed. NAEP is also moving towards a fully online-based assessment in 2024, which is a necessary step towards administering on school devices.

NCES recently conducted a field test to examine the logistic impacts of moving to an online assessment administered on Chromebooks. Additional proof-of-concept studies and field test are planned to examine the impacts of moving towards a hybrid administration mode of school-based equipment and NAEP-provided Chromebooks. The November 2023 COSDAM session will be a continuation of an August 2023 session and will provide an opportunity for COSDAM members to see examples of item display comparability across device types.