IES Institute of Education Sciences

November 2023 U.S. DEPARTMENT OF EDUCATION

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

Institute of Education Sciences

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 choose the correct answer. However, scoring math responses is difficult for artificial intelligence methods like natural language processing because it combines specific calculations (including mathematical notation) with conceptual information which uses normal text.

Humans can score most of these items very accurately. The purpose of the Challenge was to tell us whether automated scoring for mathematics responses could be equally accurate and what would be required for NAEP to 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.

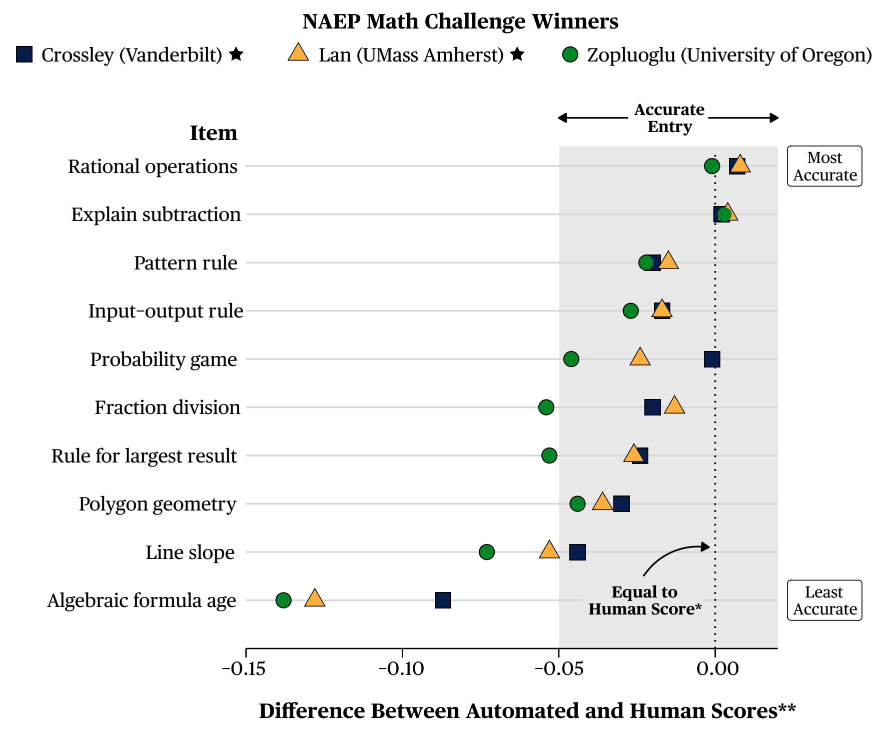
Judges first evaluated technical reports, which described the methods used for scoring. If reports met transparency and fairness analysis requirements, then teams' entries were analyzed for accuracy and for whether bias was observed in the teams' predictions.

https://doi.org/10.35542/osf.io/eyzgd

Implications for NAEP

- **Automated scoring can accurately score open-ended math items.** Use of automated scoring should be determined at the item level, ensure accuracy before being used, and include fairness analyses. Automated scoring methods can save time and money and they can allow for deeper analysis of the data.
- **Automated scoring has the potential to expand the usefulness of NAEP.** 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 accurate, 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 pre-processing of student responses was required. This included 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 with winning submissions 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 were less clear. Item content or presentation could be a problem to examine in such items in the future. 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.

Summary of Methods Used by Winning Teams

Team	Summary of Approach
S. Crossley & LEAR Lab (Vanderbilt)	This team's 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.
A. Lan (UMASS-Amherst)	This team first corrected students' 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.
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).

Authors

John Whitmer
Magdalen Beiting-Parrish
(Institute of Education Sciences)

Charles Blankenship
Amy Fowler-Dawson
McCall Pitcher
(American Insitutes for Research)

Challenge Contributors

Scott Crossley, Joon Suh Choi, Langdon Holmes, Wesley Morris (Vanderbilt University); Andrew Lan, Mengxue Zhang, Jaewook Lee, Hunter McNichols, Wanyong Feng, Alex Scarlatos (University of Massachusetts Amherst); Cengiz Zopluoglu (University of Oregon)