Lessons Learned About Transparency, Fairness, and Explainability from Two Automated Scoring Challenges

Anonymous Submission[1]

Abstract

This paper describes the results of two automated scoring challenges that were conducted as research studies to evaluate the feasibility of using automated scoring for fourth and eighth grade reading and math short constructed responses. These challenges demonstrated that these responses could be scored almost as accurately as human raters, with the math items being even more accurately scored than reading items. Challenge review criteria included a required technical report that made the approach used explainable, interpretable, and transparent. In addition, both challenges required the participants to demonstrate that their innovation was fair and did not contribute additional bias in scoring in order for submissions to be considered valid entries. For both the reading and math challenge, no bias was discovered for major demographic groups of race/ethnicity or gender. For both challenges, the participants described their feature engineering process as well as their process of designing and testing their model of interest; however, they did not provide interpretable models due to the use of Large Language Models that have thousands or millions or parameters to represent the student text. This paper describes the fairness and transparency/interpretability results as well as some suggested future directions for the field.

Introduction

Automated scoring of open-ended assessment items is one of the most widely deployed uses of Artificial Intelligence (AI) in Education, and a recent literature review found it to be the most researched area of AI in Education Measurement (Zheng et al., 2023). The National Assessment of Educational Progress (NAEP) is the largest nationally representative and continuing assessment of what students in public and private schools in the United States know and can do across various subjects (U.S. Department of Education, 2023). NAEP makes extensive use of open-ended assessment items and in larger administrations has millions of student responses that are currently scored by human raters. [ANONYMIZED AGENCY] conducted studies using an open data challenge in 2021 and 2023 to evaluate the accuracy, feasibility and cost of automated scoring through a data challenge using released Reading and Math assessment items. Fairness analysis was required for both challenges, as well as transparency and interpretability, in order to build credibility, identify the best performing results, and ensure that they accurately represented all student populations. This paper describes the differences in the approaches taken in crafting the challenge requirements and the differences in results that were achieved across these challenges.

Context and Accuracy Results

In Fall 2021 and Spring 2023, [ANONYMIZED AGENCY] conducted a data challenge using Challenge.gov to evaluate the potential of using automated scoring techniques to score open-ended responses to 20 released NAEP reading assessment items and 10 released NAEP mathematics assessment items. These items provide the opportunity for students to demonstrate their reading comprehension and to explain the answers given to forced choice (e.g., multiple choice) math items. The datasets were relatively large, with over 450,000 student responses included in the reading challenge and over 275,000 student responses included in the math challenge. The datasets also included the item itself, a detailed scoring guide, sample responses, and student demographic and educational preparedness information. The math challenge also included student process and response data from earlier parts of the same item (e.g., forced choice responses).

While automated scoring of open-ended student writing has been conducted accurately for over a decade (Hamner et al., 2012), scoring math responses is a more challenging problem (Baral et al., 2021). The likely reason for this challenge is likely due to the response itself: it combines specific calculations with conceptual information. Humans, however, can score these items very accurately. The purpose of the challenge was to determine the existing capabilities, accuracy metrics, underlying validity evidence of assigned scores, and efficiencies of using automated scoring for mathematics responses. However, challenge winners were able to accurately score 9 out of 10 math items and 18 out of 20 reading items using commonly accepted measures (Williamson et al., 2012).

[1] Disclaimer: the views in this paper reflect those of the author and are not a statement of policy or intended practice by the [ANONYMIZED AGENCY].

Transparency and Interpretability

Automated scoring is an area of substantial commercial activity in the education measurement services industry and several companies have invested decades of time and significant financial resources in developing their algorithmic solution. However, in order to build trust from education stakeholders, for the reading challenge, the [ANONYMIZED AGENCY] team required that all submitted solutions include a technical report that was transparent in describing the algorithmic approaches used and the modeling results that were taken. Further, for the reading challenge the team required that responses include "interpretability" analyses in that technical report. That report was reviewed by subject matter experts and only submissions with approved reports were included in the accuracy scoring challenge.

Respondents overall met the criteria for transparency in describing their data cleaning, preprocessing, and modeling approaches, but provided very limited information in the transparency area. Because most teams used a variety of transformer-based models, creating transparent results at the model level is a difficult technical problem. As a result, for the math challenge a new prize area was created and a $20,000 prize was made available for transparency in addition to the prize for accuracy in predicted scores. A specific rubric and criteria for interpretability were created (e.g., post-hoc alternative measures); no team met those criteria, despite highly accurate results, and that prize was not awarded. Feedback from respondents and discussions suggested several potential reasons for this result, including time restrictions, effort required for accurate modeling, and potential differences of expertise for prediction accuracy compared to explainability. Future challenge teams may need to include subject matter experts and/or measurement experts who can understand the relationship of the variables to the predicted scores. More needs to be understood in this area to make the scoring algorithms more interpretable to external stakeholders to further foster trust in these models.

Fairness Analyses

Another required component for both challenges was fairness analysis. In the Reading Challenge, a specific method was not required for the technical report and was left open to the discretion of the respondent, although for results evaluation the standardized mean difference (SMD) in accuracy between subpopulations was used to calculate accuracy in results. In this study, no additional bias was contributed by the automated scoring engines for the major demographic groups of race/ethnicity and gender which the participants knew as part of the training and test set; by contrast, the participants did not receive demographic information about English Language Learner status (LEP) or Individualized Education Plan status (IEP), but subsequent post hoc studies by the authors found that the scoring engines contributed additional bias for both of these demographic subgroups, especially LEP status.

Based on the variability in approaches used in the technical reports, for the Math Challenge the SMD calculation was required for the technical report, although alternative and potentially more insightful approaches were suggested. No team implemented any of these innovative approaches. The threshold for maximum SMD was reduced between the challenges from 0.15 for the Reading Challenge to 0.10 for the Math Challenge. The participants in the Math Challenge also received additional demographic information as part of the challenges, including LEP and IEP status. For this study, none of the top three winning teams demonstrated additional bias for any of the major demographic subgroups and all SMDs were substantially smaller than 0.10; for other participating teams, the most common categories to demonstrate bias were LEP and IEP status. None of the winning math participants submitted additional fairness analyses outside of SMD.

Based on the lack of innovation in responses, a focused analysis on innovative approaches to fairness seems well-justified and important to the field. Additionally, there are currently no formalized best practices for how to ensure that an automated scoring engine is fair for all subpopulations; this is an additional area for future study. Finally, because the writing of students with Limited English Proficiency seems to be particularly difficult to score fairly with current automated scoring approaches, more needs to be understood about this population's writing.

Implications

The results achieved demonstrate clearly that accurate automated scoring is possible for NAEP Reading and Math items. The responses also demonstrate that commercial assessment services companies and academics are willing and able to transparently describe their approaches used and that no "magic box" solutions need to be accepted for the sake of accuracy.

The results in fairness and interpretability, however, indicate that these areas need further investment and dedicated attention to change from the status quo approaches. For machine learning-informed methods to be thoroughly useful to the field, a set of best practices and standard methodologies must be developed to ensure that these algorithms are fair for all demographic subgroups and they must be sufficiently transparent/interpretable for stakeholders to understand how these scores are being predicted for full confidence in these systems.

References

S. Baral, A. F. Botelho, J. A. Erickson, P. Benachamardi, and N. T. Heffernan, "Improving Automated Scoring of Student Open Responses in Mathematics," International Educational Data Mining Society, 2021. Accessed: Nov. 26, 2023. [Online]. Available: https://eric.ed.gov/?id=ED615565

Hamner, B. Morgan, J., lynnvandev, Shermis, M., & Vander Ark, T. 2012. The Hewlett Foundation: Automated Essay Scoring. *Kaggle*. kaggle.com/competitions/asap-aes.

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2023 NAEP Data Explorer. www.nationsreportcard.gov/ndecore/landing.

Williamson, D.M., Xi, X. and Breyer, F.J. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice* 31: 2-13. doi.org/10.1111/j.1745-3992.2011.00223.x .

Zheng, Y., Nydick, S., Huang, S., & Zhang, S. 2023. MxML (Exploring the paradigmatic relationship between measurement and machine learning in the history, current time, and future): Current state of the field. https://doi.org/10.35542/osf.io/n9reh .