





Exploring Fairness in Automated Grading and Feedback Generation of Open-Response Math Problems

Ashish Gurung^(✉)  and Neil T. Heffernan 

Worcester Polytechnic Institute, Worcester, MA 01609, USA
{agurung,nth}@wpi.edu

Abstract. The rapid growth and development of NLP techniques have resulted in Computer-Based Learning Platforms (CBLPs) leveraging innovative approaches toward automated grading and feedback generation of open-ended problems. Researchers have explored these techniques in driving a varying range of interventions that range from assessing the quality of the work and recommending changes to the answers that can enhance the quality of the responses for students to automated grading and feedback generation of responses for teachers. A crucial aspect of the automated assessment of student response is identifying and addressing fairness and equity issues in an educational context, as academic performance can impact the types of opportunities available to the students. While prior works have conducted posthoc analysis exploring aspects of algorithmic fairness of various models, the assessment of open-ended answers is often subjective. Teachers leverage contextual knowledge such as the perception of the student effort or students' prior knowledge. While such factors exist, it is not obvious how data from the teacher can introduce biases or introduce measurable risks to the fairness and equity of the NLP models. In this paper, we build on our prior analysis of the grading behavior of teachers on open-ended math problems for middle school students and explore possible next steps we can take to expand on our work. First, we propose a simulation study to explore the various risks associated with Human-AI interaction in the automated grading of open-ended problems. Second, we propose an extensive study expanding on our work to generate grades for open responses when a student is anonymized vs. not anonymized.

Keywords: Open-ended problems · Fairness · Bias · Grading

1 Introduction

The integration of CBLPs into classrooms and the willingness of teachers to utilize them in various capacities in their classrooms has enabled researchers to explore the effectiveness of CBLPs through data-driven methods. Consequently, researchers have focused on developing their CBLPs in alleviating difficult or tedious tasks faced by teachers in their everyday classroom activities.

Researchers, however, have faced challenges in supporting open-ended problems due to the variance in the answers. While the recent advancement in NLP and machine learning have made progress towards automating the assessment of open-ended questions in various domains, the evaluation of open-ended responses remains a predominantly manual task for teachers. Writing is a critically important skill, and it facilitates students with an avenue to exhibit their thought processes and ability in formulating arguments and providing justifications for their work [11, 26]. In mathematics, it enables teachers to gauge whether students have a strong comprehension of mathematical concepts. Furthermore, teachers can leverage open-ended problems to identify situations where students may be able to answer close-ended problems correctly by shallowly learning and applying procedural rules [17, 23].

The assessment of open response problems is a largely subjective task. Giving concise responses to open-ended maths problems further underscores the subjective nature of grading open response problems. While grading of open-ended responses often relies on rubrics or other standardized procedures to help optimize the evaluation procedure, teachers often account for contextual factors. The students' past academic performance, persistence exhibited during lessons, or other qualities may affect the teachers' grades. It is important to emphasize that this does not necessarily mean that the grading is unfair. The subjective nature accounting for student ability can positively impact students through personalized feedback [13, 14]. However, teachers usage of contextual information in the assessment of students' performance presents a unique challenge in automating the grading of open-ended responses and raises concerns about ensuring the fairness.

Our goal in this work is to build on our prior work and explore teacher grading behavior of open-ended problems and the role of student identity on the grades. Explore the effects of anonymized vs. non anonymized data in the automated grading and feedback generation of open response problems. As such, this paper aims to address the following research questions:

1. Does using anonymized grades in NLP models mitigate possible biases introduced by student identity?
2. What factors affect the teacher's perception of AI agents in automated grading of open-ended math problems?
3. How does teacher perception of AI agents influence their behavior?

2 Background

Growth and innovation in Education Technology (Ed-Tech) have influenced the adaption and regular usage of CBLPs in classrooms. Through ease of logging data, the adaption of CBLPs has motivated researchers to explore the effectiveness of various design paradigms, from traditional teacher-driven designs to self-paced learning, peer learning, discussion-oriented learning, demonstration-focused learning, and flipped classrooms. Several platforms often provide a selection of these features for teachers and students to leverage instead of simply

focusing on a single one. Similar to the different design paradigms, researchers have also taken a varied approach in prioritizing the focus of their platform. Some provide a generic platform to host content and leverage crowdsourcing to address learner needs, such as generating problems and solutions [2, 7], collecting hints and explanations [4, 27]. Other platforms focus on specific domains such as writing skills [3, 22], mathematics [5, 12], programming [21] to facilitate learning by providing content that addresses the specific needs of learners. It is important to note that these two approaches of prioritizing focus are not mutually exclusive. Platforms often leverage a combination of designs that focus on a specific domain while also facilitating crowdsourcing features that address learner needs.

The automated grading and feedback generation of open-ended problems has been particularly challenging. Researchers have explored various approaches to provide real-time feedback and assessment of open responses to support students. Similar efforts have also been made to support teachers by automating the assessment of open-ended responses. Various approaches such as hand-crafted boutique pattern matching [24], and deconstructing grading rubrics into knowledge components [25]. The rapid growth and innovation of NLP have provided a significant advantage to automating the assessment of open-ended student responses. Researchers have explored NLP in evaluating a diverse range of responses from short-answer responses in mathematics [1, 9] to long-form responses such as essays [3, 16]. Neural network models such as Word2Vec [19], Glove [20], and BERT [8] have enabled the ability to capture semantic and contextual information from responses. While using deep learning models has improved the NLP models' performance, they require a large corpus of data that often are not readily available or easy to compile.

Researchers have explored the effectiveness of NLP models in the automatic grading and feedback generation of responses; there is a requirement for examining the effectiveness of the automation while accounting for fairness. Most examinations of fairness revolve around the algorithm's performance [10, 15] and model generalizability across target groups to identify possible biases [6, 18]. However, post hoc analysis of models can be rather challenging. These biases can only be mitigated if we are conscious of their existence beforehand or by detecting the existence of biases across certain aspects, such as genders or biases across ethnicity. We propose exploring the utility and effectiveness of NLP models when trained on anonymized data vs. when trained on non-anonymized data.

3 Teacher Grading Behavior

In prior work, we reported on a pilot study where we asked 14 teachers to grade anonymized open-ended responses of students who worked on three open response problems in the month prior to the study. Of the 14 teachers, only 9 completed the study. The data corpus only included the students of the 14 teachers in the pilot study. A random sample of 25 responses was generated per teacher, where we checked to ensure that at least 10 of the 25 responses were responses from their students. If the random sample had less than 10 open

responses from their students, then additional open responses were selected for the teacher to grade by randomly selecting additional responses from their students. If a teacher did not have any of their students in the random sample, they were assigned an additional 10 responses, making the total number of problems they graded 35. Table 1 reports on the 9 teachers who completed the study along with the total number of problems they graded (N) non-anonymized beforehand and anonymized during the pilot study. Some teachers had less than 10 problems to grade because we had to remove duplicate answers (e.g., empty responses or answers of “I do not know”) to ensure that the teacher graded a unique set of responses.

As shown in Table 1, we explored the teacher’s grading behavior by applying Cohen’s Kappa to measure the variation in their grading of student responses when anonymized vs. non-anonymized. We found the agreement coefficient to be as low as $k = 0.163$ and as high as $k = 0.67$, which was concerning as it indicated that the teacher disagreed with themselves when it came to scoring their students when their students across conditions. The grading behavior was lower than anticipated, indicating significant differences in teacher grading behavior when students were anonymized. Given that the grades are given on a 5 point scale, and the teacher’s assessment may reasonably vary by a small degree, we also explored a relaxed calculation of Kappa. We computed the intra-rater reliability of each teacher with an off-by-one adjustment; if the absolute difference in score across conditions was one or less, then we treated it as equivalent. The adjustment resulted in notably higher kappas indicating that teachers have consistent general grading behavior. We also computed the average difference in the grades across conditions. While most teachers were more lenient graders when they knew the student’s identity, some of the teachers were more lenient when the student was anonymized.

Table 1. Exploring the grading behavior of teachers when they had access to students’ identity vs. when students were anonymized.

Teacher	N	Intra-rater reliability (Cohen’s kappa)	Intra-rater reliability (Relaxed Cohen’s Kappa)	Avg grade diff (initial - anonymized)
Teacher1	10	0.2857	0.8550	-0.2
Teacher2	10	0.6774	1.0000	0.2
Teacher3	10	0.2307	0.6666	-0.2
Teacher4	10	0.5161	0.8387	0.3
Teacher5	11	0.1630	0.5268	0.27
Teacher6	19	0.4264	0.7816	0.57
Teacher7	9	0.3793	0.3793	0.44
Teacher8	10	0.4366	0.5522	0.3
Teacher9	9	0.5344	0.8301	-0.66

3.1 Analysis Plan

Currently, we are designing a larger study expanding our pilot study to explore teacher grading behavior and investigate if the proportion of the behavior where some teachers are more lenient grader than others repeats itself across teachers. The more extensive study also provides the data to train the NLP models to compare the model performance when trained on anonymized grades versus non-anonymized grades.

Acknowledgements. We would like to thank the NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), and Schmidt Futures.

References

1. Baral, S., Botelho, A.F., Erickson, J.A., Benachamardi, P., Heffernan, N.T.: Improving automated scoring of student open responses in mathematics. International Educational Data Mining Society (2021)
2. Bhatnagar, S., Lasry, N., Desmarais, M., Charles, E.: DALITE: asynchronous peer instruction for MOOCs. In: Verbert, K., Sharples, M., Klobučar, T. (eds.) EC-TEL 2016. LNCS, vol. 9891, pp. 505–508. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45153-4_50
3. Burstein, J., Tetreault, J., Madnani, N.: The e-rater® automated essay scoring system. In: Handbook of Automated Essay Evaluation, pp. 77–89. Routledge (2013)
4. Cambre, J., Klemmer, S., Kulkarni, C.: Juxtapeer: comparative peer review yields higher quality feedback and promotes deeper reflection. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2018)
5. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adap. Inter.* **4**(4), 253–278 (1994)
6. Crawford, K.: The trouble with bias. In: Conference on Neural Information Processing Systems, invited speaker (2017)
7. Denny, P., Hamer, J., Luxton-Reilly, A., Purchase, H.: PeerWise: students sharing their multiple choice questions. In: Proceedings of the Fourth International Workshop on Computing Education Research, pp. 51–58 (2008)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
9. Erickson, J.A., Botelho, A.F., McAteer, S., Varatharaj, A., Heffernan, N.T.: The automated grading of student open responses in mathematics. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pp. 615–624 (2020)
10. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 329–338 (2019)

11. Graham, S., Perin, D.: Writing next: effective strategies to improve writing of adolescents in middle and high schools. A report to Carnegie Corporation of New York. Alliance for Excellent Education (2007)
12. Heffernan, N.T., Heffernan, C.L.: The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int. J. Artif. Intell. Educ.* **24**(4), 470–497 (2014). <https://doi.org/10.1007/s40593-014-0024-x>
13. Hill, H.C., Schilling, S.G., Ball, D.L.: Developing measures of teachers' mathematics knowledge for teaching. *Elem. Sch. J.* **105**(1), 11–30 (2004)
14. Jacob, R., Hill, H., Corey, D.: The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *J. Res. Educ. Effect.* **10**(2), 379–407 (2017)
15. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 35–50. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33486-3_3
16. Kim, Y.S.G., Schatschneider, C., Wanzek, J., Gatlin, B., Al Otaiba, S.: Writing evaluation: rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Read. Writ.* **30**(6), 1287–1310 (2017)
17. Livne, N.L., Livne, O.E., Wight, C.A.: Enhancing mathematical creativity through multiple solution to open-ended problems online (2008). http://www.iste.org/Content/NavigationMenu/Research/NECC_Research_Paper_Archives/NECC2008/Livne.pdf
18. Mayfield, E., et al.: Equity beyond bias in language technologies for education. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 444–460 (2019)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
20. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
21. Price, T., Zhi, R., Barnes, T.: Evaluation of a data-driven feedback algorithm for open-ended programming. International Educational Data Mining Society (2017)
22. Roscoe, R.D., Allen, L.K., McNamara, D.S.: Contrasting writing practice formats in a writing strategy tutoring system. *J. Educ. Comput. Res.* **57**(3), 723–754 (2019)
23. Silver, E.A.: The nature and use of open problems in mathematics education: mathematical and pedagogical perspectives. *Zentralblatt für Didaktik der Mathematik/Int. Rev. Math. Educ.* **27**(2), 67–72 (1995)
24. Sukkarieh, J.Z., Pulman, S.G., Raikes, N.: Automarking: using computational linguistics to score short, free- text responses (2003)
25. Sukkarieh, J.Z., Blackmore, J.: C-rater: automatic content scoring for short constructed responses. In: Twenty-Second International FLAIRS Conference (2009)
26. Walton, D.N.: Plausible Argument in Everyday Conversation. SUNY Press (1992)
27. Williams, J.J., et al.: AXIS: generating explanations at scale with learnersourcing and machine learning. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale, pp. 379–388 (2016)