# Strategies for Deploying Unreliable AI Graders in High-Transparency High-Stakes Exams

Sushmita Azad(✉), Binglin Chen, Maxwell Fowler, Matthew West, and Craig Zilles

University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{sazad2,chen386,mfowler5,mwest,zilles}@illinois.edu

**Abstract.** We describe the deployment of an imperfect NLP-based automatic short answer grading system on an exam in a large-enrollment introductory college course. We characterize this deployment as both high stakes (the questions were on an mid-term exam worth 10% of students' final grade) and high transparency (the question was graded interactively during the computer-based exam and correct solutions were shown to students that could be compared to their answer). We study two techniques designed to mitigate the potential student dissatisfaction resulting from students incorrectly not granted credit by the imperfect AI grader. We find (1) that providing multiple attempts can eliminate first-attempt false negatives at the cost of additional false positives, and (2) that students not granted credit from the algorithm cannot reliably determine if their answer was mis-scored.

**Keywords:** Automatic short answer grading · Computer-based exams · Transparency · Code reading · CS1 · EiPE

## 1 Introduction

Workplace demand for computing skills [19] has led to large enrollments in introductory programming classes [6]. These courses, however, have had historically large failure rates [2,29]. Some evidence suggests that this is due to a premature emphasis on code writing instead of reading-oriented activities [4,14,32]. One important reading skill is the ability to describe the high-level behavior of code [14,17,18,31]. Questions to assess this skill—"Explain in Plain English" (EiPE) questions—aren't widely utilized due to the workload of manually grading natural language responses. Figure 1(A) shows an example prompt of one of our EipE questions.

In this work, we describe our initial efforts in deploying an NLP-based AI grader for EiPE questions and our transition from low-stakes to high-stakes environments. Initially, simple NLP-based AI graders were trained using a small amount of survey data collected from course teaching assistants and upper-level undergraduate computer science students. These simple AI graders were

© Springer Nature Switzerland AG 2020
I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12163, pp. 16–28, 2020.
https://doi.org/10.1007/978-3-030-52237-7_2

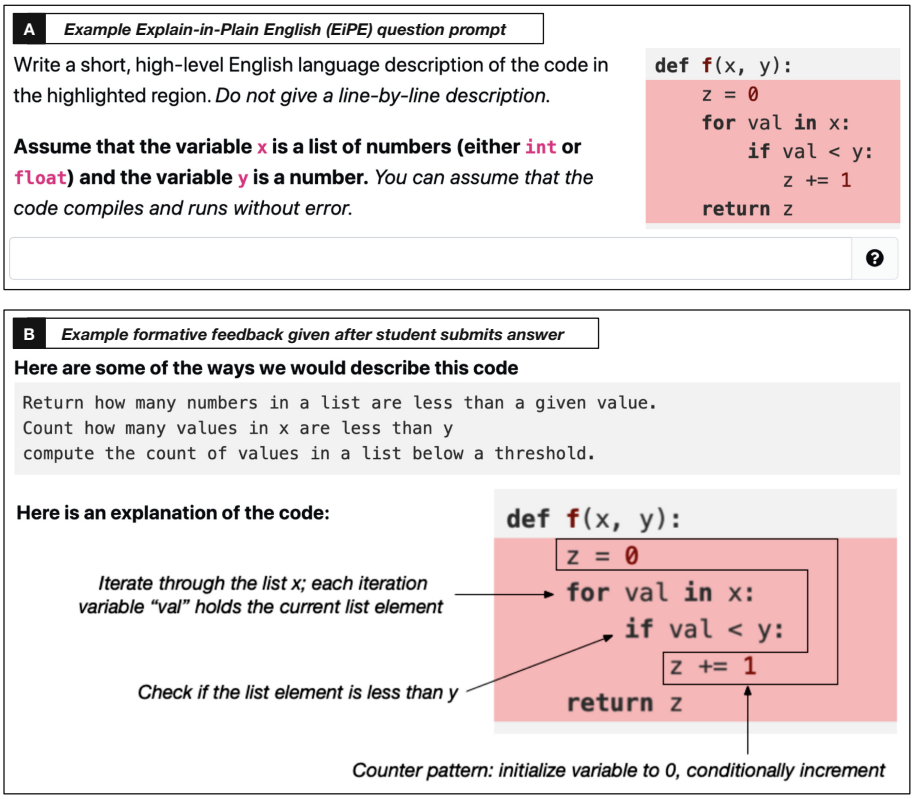


**Fig. 1.** An example mid-semester automated EiPE exercise (A) in a Python-based intro CS course. After a student submits their answer, they are shown example solutions (B) so that they can learn. Non-trivial code fragments are deconstructed so as to show the correspondence between the code and the natural language description.

deployed in a low-stakes homework context for which we had two goals: 1) we wanted students to improve their ability to provide natural language descriptions of code, so we provided both immediate correct/incorrect feedback and example correct answers as shown in Fig. 1(B) and 2) we wanted to collect additional training data which could be used to train improved NLP-based AI graders.

Positive results with the homework deployment emboldened us to deploy our AI grader on an exam. To our knowledge, this deployment is unique in the research literature for (imperfect) AI-based graders because it was both high stakes—this question was on one of three midterm exams each worth 10% of students' final grades—and high transparency—the question was graded interactively and students are shown correct answers in a way that permits them to evaluate their submitted answer in light of the correct answers.

A high-stakes, high-visibility deployment of an imperfect AI grader, if not well managed, has the potential for student dissatisfaction on a large scale. As such,

we wanted to understand what precautions can be taken to prevent students from feeling that they were harmed by such an imperfect grader. To this end, we were willing to tolerate some number of false positives in order to minimize false negatives, and we were willing to employ some manual labor. All things being equal, however, we sought to minimize false positives and the amount of manual labor required.

We brain-stormed two strategies to minimize false negatives and, hence, student unrest. First, because our exam was graded interactively on a computer, we could permit students to attempt the question multiple times if the AI grader didn't award them credit on their first attempt. This would hopefully permit students to re-word their answers into a form that could receive credit automatically from the algorithm. Second, we could provide students an *appeal system* where they could, after they are shown the correct answer, request a manual re-grade for an EiPE question, if they believed the AI grader had scored them incorrectly.

These two strategies led to two corresponding research questions:

**RQ1:** Does providing students with multiple attempts enable false negatives to earn credit without manual intervention?

**RQ2:** Can students correctly recognize when the AI grader has failed and appropriately appeal for a manual re-grade?

Our findings can be summarized as follows:

1. The two techniques were effective at avoiding large-scale student dissatisfaction.
2. Re-training the AI grader using student responses drawn from the homework deployment improved the accuracy from 83.4% to 88.8%.
3. Providing three attempts (at full credit) enabled all first-attempt false negatives to automatically earn credit from the algorithm. It did, however, have the consequence of yielding additional false positives.
4. Appeals were useful for morale, but were not effective for distinguishing false negatives from true negatives.
5. Students' perception of the grading accuracy of our NLP-based AI grader was lower than that of deterministically-correct auto-graders for true/false, multiple-choice, and programming questions, but only to a modest degree.

This paper is structured as follows. Section 2 briefly reviews related work. Section 3 describes our data collection and AI grader training, while Sect. 4 reviews the AI grader's performance and results. We conclude in Sect. 5.

## 2 Related Work

Automatic grading of free response questions is largely split into two areas of focus: automatic short answer grading (ASAG) and automatic essay scoring (AES). We review briefly the recent work in both areas below.

A review of recent, competitive machine learning ASAG shows only 11% of ASAG papers were focused on computer science [11]. Most of the recent studies are laboratory studies or model evaluations on public or sample data sets [11,16,20,22,25,26,33]. The closest to a high-stakes course exam featured automatic grading for *very short answer*—defined as four or less words—questions, but in a not-for-credit exam-like context rather than on a for-credit exam [23]. The Educational Testing Services (ETS) C-rater is deployed for some ETS standardized exams, but is not high-transparency and focuses on concept mapping [13,24]. ASAG feature selection includes lexical, semantic, syntactic, morphological, and surface features [3,11,26]. Most recently, dialog based systems and intelligent tutoring systems [20,22,25] and end-to-end models have been used for ASAG [16,33]. To our knowledge, no ASAG work has reported on the deployment of AI graders in a high-stakes, high-transparency environment like ours.

AES work is more familiar with high-stakes environments. The ETS E-rater receives yearly updates and is used in both high-stakes settings like the GRE and low-stakes such as the SAT online test [21]. However, these systems are not high-transparency as students are provided no means to judge the validity of their scores and there is no process to contest scores. Further, AES' major impact is reduction of human labor, with the evaluation of essays focusing broadly on how essay features correlate to human-grader provided marks rather than specific content grading [12]. Recent AES approaches include GLMMs [8], autoencoders [7], statistical classifiers [15], and various deep-learning neural network approaches [1,9,10,27].

## 3 Methods

In Fall 2019, we developed and deployed automated EiPE questions in an introductory CS course for non-technical majors at a large U.S. university. This 600-student course introduces basic principles of programming in both Python and Excel to a population largely without any prior programming experience. The course was approaching gender balance with 246 women and 355 men.

We constructed our EiPE AI graders using logistic regression on bigram features. These graders were initially trained with minimal data from a series of surveys. Each survey asked participants to provide two correct responses and two plausible incorrect responses for each of the EiPE questions. These surveys were completed by the course's instructor and TAs and a collection of upper-level CS students who were compensated with an Amazon gift card for each survey. These surveys resulted in approximately 100–200 responses per question. Survey data was manually reviewed by a research team member to perform any necessary re-categorization of the responses.

This survey-data-trained AI grader was deployed on four homework assignments during the first half of the semester. The questions were deployed using the PrairieLearn [30] online learning platform, the course's primary assessment system. Each assignment included a pool of 10–12 EiPE questions, and each

time a student attempted a question they were given a random draw from the pool. To tolerate the AI grader's inaccuracy in this low-stakes, formative context, students could attempt the activity as many times as they wanted; points were granted for any correct answers with no penalty for incorrect answers. As such, any false negatives would only delay (rather than prevent) students from getting points. Furthermore, the AI graded EiPE questions were one of many activities on the students' weekly assignment, and they could ignore the activity completely and earn the week's homework points through answering other questions instead.

We next deployed the auto-graded EiPE questions as one of 24 questions on a proctored, computer-based mid-term exam in the 12th week of the course (also run using PrairieLearn). We selected the pool of EiPE questions deployed on the homework during the 5th week of the course. Prior to deployment, two members of the research team manually labeled the students' homework responses to these questions and used as additional training data to improve the grader. The AI graders deployed on the exam were trained with 500–600 labeled responses per question.

Four of the problems in the pool were not included on the exam because they exhibited a noticeable difference in difficulty from the rest. Students were randomly assigned one of the remaining eight problems. Students were given three attempts to submit a correct answer, receiving correct/incorrect feedback on each submission and were shown correct answers (as shown in Fig. 1(B)) once all attempts had been used or their answer was scored as correct.

The students submitted a total of 1,140 responses. After the exam was completed, for the purpose of this research, two members of the research team familiar with the course content independently scored each response without knowing the AI grader's score. For any responses where these two scores matched, the score was considered the final ground truth. Final ground truth for the remaining responses was established by a process of discussion and reconciliation between both scorers and a third research team member until consensus was reached. Necessary grade corrections were made for all students who had incorrectly been denied credit. All further analysis in this paper has been done on this set of 1,140 auto-graded exam responses.

To understand how students perceived the accuracy of auto-graded EiPE questions as compared to other types of auto-graded questions, we asked students to fill out a survey in the week after the exam with the EiPE question. Using a 1–5 Likert scale, students were asked: "For each type of question, rate it based on how reliably accurate you feel the grading for that kind of question is".

## 4 Results

**Comparing AI Grader and Human Performance.** 51% of students had their EiPE question scored as correct by the reconciled human graders, and the AI grader achieved an accuracy of 89%, with a 12% False Positive (FP) rate and a 9% False Negative (FN) rate. We used Cohen's kappa to compare the

inter-rater reliability of humans and the AI grader. Cohen's kappa between the two experienced human graders was 0.83 ("almost perfect" agreement [28]) and between the AI grader and the ground truth (reconciled human graders) was 0.74 ("substantial" agreement [28]).

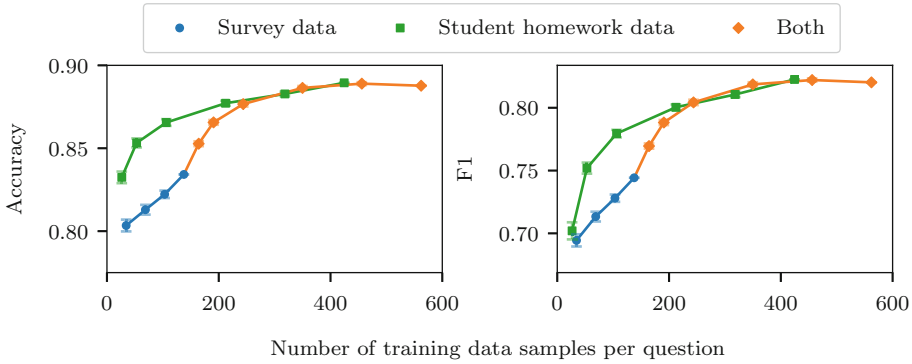


**Fig. 2.** The performance of the AI grader on the 1,140 exam responses when trained on different combinations of data with different sample sizes.

**AI Grader Accuracy Versus Amount of Training Data.** To understand how much training data is needed for obtaining a reasonable AI grader and whether there is a qualitative difference between survey data and student homework data, we trained graders with different subsamples of data and show the mean of the grader's performance in Fig. 2. There are three main sources of training data: (1) a subset of the survey data, (2) a subset of the student homework data, and (3) both, meaning all of the survey data and a subset of the student homework data. Although more data consistently lead to better performance, the student homework data seems qualitatively better than survey data, suggesting that the course staff and senior students creating the survey data were only somewhat able to generate realistic training data.

**Student Perceptions of Accuracy.** Students perceived the grading of AI graded EiPE questions as being less accurate than that of other kinds of questions to a statistically significant degree ($p < 0.001$). Compared to the next-lowest question type (programming), code-reading questions were $d = 0.48$ standard deviations lower, a "medium" effect size [5]. Mean Likert scores for each type of question are shown in Fig. 3 with 95% confidence intervals. We failed to find any correlation between students' perception of the EiPE AI grader and whether it mis-graded their answers on the exam. Instead, a student's perception of accuracy for all kinds of questions is weakly correlated with the student's performance on that kind of question (mean $r = 0.22$).

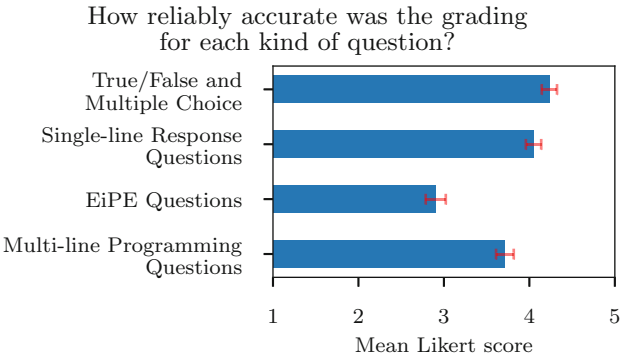


**Fig. 3.** Responses to a survey question auto-grader accuracy by question type. Choices were from 1 = "Very Unreliable" to 5 = "Very Accurate".

**Multiple-Attempt Accuracy.** We need to differentiate between the AI grader's performance on a single student submission versus the net performance over all student submissions to a question. To describe the latter, we define the *Multi-Attempt-k* outcomes as shown in Table 1. Whenever we use terms like *False Positive (FP)* without the prefix of "Multi-Attempt", we are referring to the performance on a single-submission level.

**Table 1.** Definitions of "multi-attempt" terminology.

| Term | Definition |
|---|---|
| Multi-Attempt-$k$ True Positive | Within the first $k$ attempts, student submits at least one correct answer and AI grader awards points for some submission |
| Multi-Attempt-$k$ False Positive | Within the first $k$ attempts, student submits no correct answer but the AI grader awards points for some submission |

We visualized how multiple attempts impact the performance metrics in Fig. 4. We see that as students attempted the question more times (moving from MA-1 to MA-3), the true positive rate increased somewhat (93.2% to 97.7%), but at the expense of a substantially higher multi-attempt false positive rate (14.9% to 32.9%). The reference ROC curve is for the AI grader evaluated on only the first-attempt responses, and we see that the multi-attempt performance is always worse than this.

**Trajectories with Multiple Attempts.** Figure 5 shows the trajectories students took through the multiple attempts at the EiPE questions. This reveals several features. First, all students who were falsely graded as incorrect (FN) on

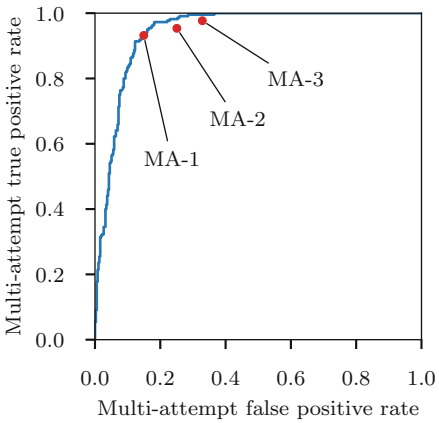


**Fig. 4.** Multi-attempt AI grader performance (MA-$k$) using only the first $k$ attempts (see Table 1). The blue ROC curve is for the AI grader on the first-attempt data only. (Color figure online)

the first attempt were able to use the multiple attempts to eventually be graded as correct (as TP or FP). A majority (73%) of these students needed a second attempt to be graded correct, and only 27% needed three attempts. Second, students who were falsely graded as incorrect (FN) re-attempted the question at a higher rate than students who were truly graded as incorrect (TN) (100% versus 96%, $p = 0.013$). Third, the ratio of falsely-graded incorrects (FN) to truly-graded incorrects (TN) decreased as students used more attempts (4.7% to 3.2%, $p = 0.015$).

**Strategies with Multiple Attempts.** Students marked as incorrect by the AI grader on either first or second attempt deployed two correction strategies: (1) *reword*, where students rephrased their previous answer, and (2) *change*, where students submitted a response different in meaning from their previous answer. Figure 6 plots the paths through these strategies taken by the student population. From a standpoint of strategy *selection*, we see that students who had an actually-correct answer (FN) used the reword strategy at a higher rate than students who did not (TN) (57% vs 42%, $p = 0.022$). Considering strategy *effectiveness*, we observe that for FN students the reword strategy was more successful for receiving points than the change strategy, but not significantly so (75% versus 25%, $p = 0.22$), whereas for TN students the change strategy was significantly more effective (81% vs 19%, $p = 0.036$).

**Appeals to Human Graders.** Out of the 203 students who were graded as incorrect by the AI grader, 69 appealed for a human re-grade and 4 of these were truly correct (rate of 5.8%). Among those that did not appeal, 3 were truly

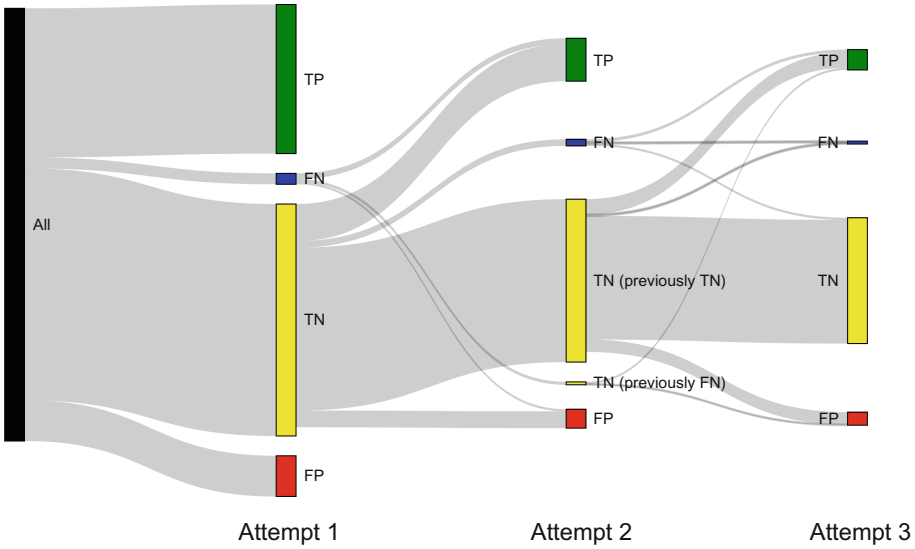


**Fig. 5.** Trajectories of all students through multiple attempts of the AI graded questions. Students who were scored as correct by the AI grader, either truly (TP) or falsely (FP), do not attempt further.

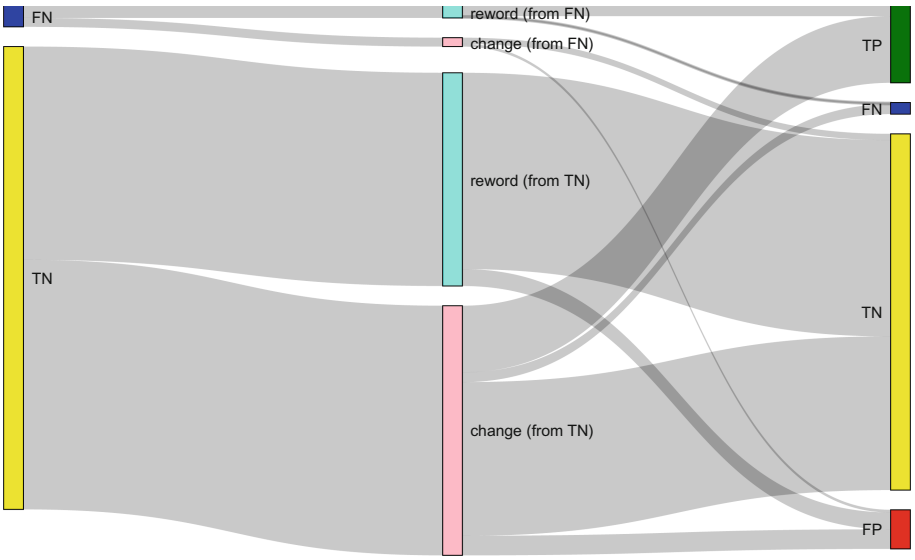


**Fig. 6.** Strategy selection and effectiveness after a submission was graded as incorrect. There was no significant dependence on attempt number, so this figure collapses all attempts together.

correct (2.2%). The difference in rates of true-correctness was not statistically significant between students who appealed and those that did not ($p = 0.20$).

## 5 Discussion and Conclusion

These initial results suggest automatically grading "Explain in plain English" (EiPE) questions may be a simpler task than other ASAG contexts. Even using just bigrams, our results (accuracy of 88.78%) are competitive with other ASAG results using much more sophisticated algorithms. We believe that this high accuracy is the result of specific elements of disciplinary vocabulary (e.g., "count", "even") being effective markers of when students have correct answers.

It is not surprising that the student homework responses were more effective than survey data for training the algorithm to predict student exam responses. The surveys did enable us to deploy the algorithm in the low stakes homework context to collect that homework training data, but our conclusion is that we could get by with fewer survey responses, especially if we were to quickly score early homework responses and re-train the model.

While students' perception of accuracy of our NLP model was statistically significantly below their perceptions of accuracy for the other question types, we were surprised by how small the difference in perceptions was. In our minds, the deterministic autograders and our NLP model are categorically different things. The students rated the deterministic autograders much lower than we anticipated (means near 4 out of 5) and the NLP model only d = 0.48 standard deviations below the deterministic autograders.

While the answer to RQ1—does providing students with multiple attempts enable false negatives to earn credit without manual intervention?—is yes, there are a number of caveats. First, while all first attempt FN students automatically earned credit on subsequent attempts, a few did so through submitting FP answers, which will potentially hinder those students' learning. Second, rather than merely reword their answer, many students used the multiple attempts to submit conceptually different answers. That is, while FN students primarily used the multiple-attempt feature to rephrase their answer for clarity (as intended by us), TN students appear to be aware that they don't know the answer, and used the multiple-attempt feature as a way to take more "shots in the dark", changing their answer in the hope that they'd strike the correct response and gain credit. Because some of these "shots" resulted in FP, giving students multiple attempts negatively impacted the FP rate.

This distinction between rewording and changing answers is important, because they have different implications on how much credit a student should receive. A student whose answer was correct, but needed rewording to be accepted by the algorithm, presumably deserves full credit. In contrast, a student that hedges by changing their answer on each submission, probably has a more fragile understanding and may deserve only partial credit. If we were to use multiple attempts again, we would probably: 1) provide only two attempts, since the majority of FNs were able to self correct within by their second try,

and 2) have a small penalty (10–30%) for credit earned on a second attempt. That said, in our current implementation providing a single attempt and just shifting the implementation along its ROC curve may provide a better FN/FP trade-off.

The answer to RQ2—can students correctly recognize when the AI grader has failed and appropriately appeal for a manual re-grade?—appears to be no. Students that appealed had a statistically equivalent rate of being correct as the whole population of students that didn't earn credit from the algorithm. Relying on students to self report appears to be an inequitable strategy that rewards "noisier" students. One important caveat is that appeals were evaluated in a context with multiple attempts; appeals could be more useful in a single-attempt context where more FNs are present.

In short, in this first report on strategies for deploying imperfect AI graders in high stakes, high visibility contexts, we found that our strategies were ultimately successful. There was no obvious student discontent and only 0.5% (3 out of 600) of students would have incorrectly not received credit (FN) had we not manually scored all responses. While our strategy was passable, there remains a lot of opportunity for improvement. Because perfect auto-graders will not be achievable for many important problems, it is important to explore hybrid AI/human systems that can mitigate algorithmic shortcomings with minimal manual effort.

**Acknowledgments.** This work was partially supported by NSF DUE-1347722, NSF CMMI-1150490, NSF DUE-1915257, and the College of Engineering at the University of Illinois at Urbana-Champaign under the Strategic Instructional Initiatives Program (SIIP).

## References

1. Alikaniotis, D., Yannakoudakis, H., Rei, M.: Automatic text scoring using neural networks. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 715–725 (2016)
2. Bennedsen, J., Caspersen, M.E.: Failure rates in introductory programming. SIGCSE Bull. **39**(2), 32–36 (2007). https://doi.org/10.1145/1272848.1272879
3. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. **25**(1), 60–117 (2014). https://doi.org/10.1007/s40593-014-0026-8
4. Clancy, M.J., Linn, M.C.: Patterns and pedagogy. In: The Proceedings of the Thirtieth SIGCSE Technical Symposium on Computer Science Education, SIGCSE 1999, pp. 37–42. ACM, New York (1999). https://doi.org/10.1145/299649.299673
5. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Routledge, Abingdon (1988)
6. Computing Research Association: Generation CS: Computer Science Undergraduate Enrollments Surge Since 2006 (2017). https://cra.org/data/Generation-CS
7. Converse, G., Curi, M., Oliveira, S.: Autoencoders for educational assessment. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11626, pp. 41–45. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23207-8_8

8. Crossley, S.A., Kim, M., Allen, L., McNamara, D.: Automated summarization evaluation (ASE) using natural language processing tools. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 84–95. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_8
9. Dasgupta, T., Naskar, A., Dey, L., Saha, R.: Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pp. 93–102. Association for Computational Linguistics, Melbourne (2018)
10. Dong, F., Zhang, Y.: Automatic features for essay scoring - an empirical study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1072–1077. Association for Computational Linguistics, Austin (2016)
11. Galhardi, L.B., Brancher, J.D.: Machine learning approach for automatic short answer grading: a systematic review. In: Simari, G.R., Fermé, E., Gutiérrez Segura, F., Rodríguez Melquiades, J.A. (eds.) IBERAMIA 2018. LNCS (LNAI), vol. 11238, pp. 380–391. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03928-8_31
12. Hussein, M.A., Hassan, H., Nassef, M.: Automated language essay scoring systems: a literature review. PeerJ Comput. Sci. **5**, e208 (2019). https://peerj.com/articles/cs-208
13. Leacock, C., Chodorow, M.: C-rater: automated scoring of short-answer questions. Comput. Humanit. **37**(4), 389–405 (2003). https://doi.org/10.1023/A:1025779619903
14. Lister, R., Fidge, C., Teague, D.: Further evidence of a relationship between explaining, tracing and writing skills in introductory programming. In: Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education, ITiCSE 2009, pp. 161–165. ACM, New York (2009). https://doi.org/10.1145/1562877.1562930
15. Liu, M., Shum, S.B., Mantzourani, E., Lucas, C.: Evaluating machine learning approaches to classify pharmacy students' reflective statements. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 220–230. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_19
16. Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G.Y., Liu, Z.: Automatic Short Answer Grading via Multiway Attention Networks. arXiv:1909.10166 [cs] (2019). http://arxiv.org/abs/1909.10166
17. Lopez, M., Whalley, J., Robbins, P., Lister, R.: Relationships between reading, tracing and writing skills in introductory programming. In: Proceedings of the Fourth International Workshop on Computing Education Research, pp. 101–112. ACM (2008)
18. Murphy, L., McCauley, R., Fitzgerald, S.: 'Explain in Plain English' questions: implications for teaching. In: Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, SIGCSE 2012, pp. 385–390. ACM, New York (2012). https://doi.org/10.1145/2157136.2157249
19. National Academies of Sciences, Engineering, and Medicine: Assessing and Responding to the Growth of Computer Science Undergraduate Enrollments. The National Academies Press, Washington, DC (2018). https://doi.org/10.17226/24926. https://www.nap.edu/catalog/24926/assessing-and-responding-to-the-growth-of-computer-science-undergraduate-enrollments

20. Ndukwe, I.G., Daniel, B.K., Amadi, C.E.: A machine learning grading system using chatbots. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11626, pp. 365–368. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23207-8_67
21. Ramineni, C., Williamson, D.: Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. ETS Res. Report Ser. **2018**(1), 1–31 (2018). https://onlinelibrary.wiley.com/doi/abs/10.1002/ets2.12192
22. Saha, S., Dhamecha, T.I., Marvaniya, S., Sindhgatta, R., Sengupta, B.: Sentence level or token level features for automatic short answer grading?: Use both. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 503–517. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93843-1_37
23. Sam, A.H., et al.: Very-short-answer questions: reliability, discrimination and acceptability. Med. Educ. **52**(4), 447–455 (2018)
24. Sukkarieh, J.Z., Blackmore, J.: C-rater: automatic content scoring for short constructed responses. In: FLAIRS Conference (2009)
25. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 469–481. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_39
26. Suzen, N., Gorban, A., Levesley, J., Mirkes, E.: Automatic Short Answer Grading and Feedback Using Text Mining Methods. CoRR (2019). arXiv: 1807.10543
27. Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1882–1891. Association for Computational Linguistics, Austin (2016)
28. Viera, A.J., Garrett, J.M., et al.: Understanding interobserver agreement: the Kappa statistic. Fam. Med. **37**(5), 360–363 (2005)
29. Watson, C., Li, F.W.: Failure rates in introductory programming revisited. In: Proceedings of the 2014 Conference on Innovation & #38; Technology in Computer Science Education, ITiCSE 2014, pp. 39–44. ACM, New York (2014). https://doi.org/10.1145/2591708.2591749
30. West, M., Herman, G.L., Zilles, C.: PrairieLearn: mastery-based online problem solving with adaptive scoring and recommendations driven by machine learning. In: 2015 ASEE Annual Conference & Exposition. ASEE Conferences, Seattle, Washington (2015)
31. Whalley, J., et al.: An Australasian study of reading and comprehension skills in novice programmers, using the bloom and SOLO taxonomies. In: Eighth Australasian Computing Education Conference, ACE 2006 (2006)
32. Xie, B., et al.: A theory of instruction for introductory programming skills. Comput. Sci. Educ. **29**(2–3), 205–253 (2019)
33. Yang, X., Huang, Y., Zhuang, F., Zhang, L., Yu, S.: Automatic Chinese short answer grading with deep autoencoder. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 399–404. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93846-2_75