Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

# Considerate, Unfair, or Just Fatigued? Examining Factors that Impact Teacher

**Ashish GURUNG[a*], Anthony F. BOTELHO[b], Russell THOMPSON[a], Adam C. SALES[a], Sami BARAL[a] & Neil T. HEFFERNAN[a]**
[a]*Worcester Polytechnic Institute, USA*
[b] *University of Florida, USA*
*agurung@wpi.edu

**Abstract:** It is particularly important to identify and address issues of fairness and equity in educational contexts as academic performance can have large impacts on the types of opportunities that are made available to students. While it is always the hope that educators approach student assessment with these issues in mind, there are a number of factors that likely impact how a teacher approaches the scoring of student work. Particularly in cases where the assessment of student work requires subjective judgment, as in the case of open-ended answers and essays, contextual information such as how the student has performed in the past, general perceptions of the student, and even other external factors such as fatigue may all influence how a teacher approaches assessment. While such factors exist, however, it is not always clear how these may introduce bias, nor is it clear whether such bias poses measurable risks to fairness and equity. In this paper, we examine these factors in the context of the assessment of student answers to open response questions from middle school mathematics learners. We observe how several factors such as context and fatigue correlate with teacher-assigned grades and discuss how learning systems may support fair assessment.

**Keywords:** halo effect, grading biases, fairness, subjective assessment

## 1. Introduction

In the context of education, a significant amount of research has been devoted to identifying, examining, and mitigating risks that particular policies, interventions, and instructional strategies may introduce as to the types of opportunities made available to students. Particularly with the introduction of computer-based learning platforms (CBLP), researchers are able to explore issues of fairness and bias through data-driven methods.

Traditional assessments are conducted in two formats: subjectively and objectively. The rise in the integration of technology in classrooms has facilitated the growth of CBLPs. Various CBLPs have been developed with the goal of alleviating difficult or tedious tasks faced by teachers. The most notable of the functionalities is the automation of objective assessments. Common in numerous contexts, the use of close-ended questions such as multiple choice and fill-in problems can be easily automated by computers where there are traditionally a small finite number of acceptable correct answers. The use of such questions has allowed developers to expand upon assessment processes to enrich the learning experience by offering additional feedback and on-demand help (Aleven et al., 2003; 2004; Ostrow et al.,2015; Patikorn et al., 2020). It is challenging to extend the same type of support to open-ended problems such as short answers and essay problems. While recent advancements in natural language processing (NLP) have made strides in automating assessment, this remains predominantly a manual task.

Writing is a critically important skill that helps teachers understand their students' thought processes and the ability to formulate arguments and justifications for their work (Biancarosa et al., 2004; Walton, 1992). In the domain of mathematics, on which the analyses of this work focus, teachers commonly use open-ended problems to gauge student knowledge when close-ended problems can often be solved by shallow learning and applying procedural rules (Livne et al., 2008; Silver, 1995). While subjective assessments are highly valuable, automating the process is not without risk; they are more dynamic compared to objective assessments. The dynamism is due to the variance in responses and the

incongruity in teacher grades; the incongruity is caused by various intrinsic and extrinsic factors. Prior work has found biases in grading behavior attributed to the "Halo Effect" (Cooper, 1981), characterized by a judgment made based on an attribute or characteristic of an individual; commonly, such attributes include gender (Spear, 1984; Roen, 1992), ethnicity (Wen, 1979; Fajardo, 1985), and name or surname (Lebuda et al., 2013). Furthermore, researchers exploring the Halo Effect found the initial favorable impression of students influenced their later evaluation (Malouff et al., 2013; 2014). While the use of rubrics or other standardized procedures provide a common set of metrics, the ultimate grade is typically based on the teachers' judgement. It is important to emphasize that measured biases do not necessarily equate to unfair assessment or evaluation as a teacher's knowledge of their students can also be very positive in terms of providing individual support through feedback and other communication (Hill et al., 2004; 2008; Jacob et al., 2017). In approaching assessment, teachers may consider a number of contextual factors when evaluating student work. In light of this, several questions emerge in terms of how assessment should be conducted to ensure fairness among students, particularly as researchers are moving to develop automated methods that attempt to mimic teacher grading practices.

Our goal in this work is to explore teacher assessment pertaining to open-ended questions. Using data collected from students working in a learning system in pre-COVID-19 classroom settings, we report on a pilot study to examine and explore teachers' approach to assessing open-ended work and how student identity may be considered in the grading process. We build on the study by conducting an exploratory analysis examining whether student-level attributes are predictive of teacher-provided grades when controlling for answer-level descriptors. Finally, we explore whether teacher grading fatigue poses risks to the fairness of student assessment.

In consideration of exploring factors that may affect fair student assessment, this paper addresses the following research questions:

1. Do teachers grade students differently when the students are anonymized?
2. When controlling for answer-level features, are factors of prior student performance and effort a reliable predictor of teacher-provided assessment scores?
3. Does the order in which teachers assess students appear to impact their grading?

To address these questions, we conduct 3 studies exploring various factors that are likely to affect or potentially bias teacher assessment. In the first study, we examine whether the anonymization of student identifiers affects how teachers score their own students. In the second study, we conduct a regression analysis to examine how measures of knowledge and effort correlate with teacher-provided assessment scores for student open-ended work. Finally, we explore the potential effects of grading fatigue on teacher-provided scores.

## 2. Background

Growth and innovation in Educational Technology (Ed-Tech) have broadly influenced the adaptation and regular usage of CBLPs in classrooms. Researchers and developers approached the design of learning platforms to consider students' various learning needs (Graham et al., 2013; Koedinger et al., 1997; Rafferty et al., 2016) to developing generic platforms (Arroyo et al., 2014; Burstein et al., 2013; Calvo et al., 2010; Heffernan et al., 2014). Systems developed to improve writing (Allen et al., 2016; Burstein et al., 2013; Calvo et al., 2010; Roscoe et al., 2019), mathematics (Arroyo et al., 2014; Heffernan et al., 2014), and programming (Price et al., 2017; Wiggins et al., 2015) are among the many examples of learning systems. Within these systems, a variety of features and supports are developed to support different aspects of learning, including the crowdsourcing of problems and solutions (Bhatnagar et al., 2016; Denny et al., 2008), and the availability of on-demand hints (Cambre et al., 2018; Williams et al., 2016).

Perhaps the most prominent feature of CBLPs is the ability to offer immediate feedback to students. Traditionally assessments are administered in objective and subjective forms. In many domains such as mathematics, students typically work through close-ended problems that can be assessed by simply comparing student answers with a finite set of acceptable correct responses (often with a simple "exact-match" approach), but across domains, the use of open-ended questions, allowing students to utilize language to explain their reasoning, have

made it more difficult for CBLPs to immediately score. While examples of automated scoring tools exist (e.g. Baral et al., 2021; Erickson et al., 2020, Allen et al., 2016; Burstein et al., 2013), many CBLPs still rely on teachers/graders for manual assessment. A primary challenge with subjective assessment associated with this manual grading process is its susceptibility to bias.

As introduced in the previous section, the Halo Effect has been the focus of prior research within the context of subjective assessment of student work (Nisbett et al., 1977; Nieva et al., 1980). Prior research has found stereotyped biases interacting with gender (Spear, 1984; Wen, 1979 Martin, 1972; Roen, 1992), ethnicity (Piche et al., 1977, Wen, 1979; Fajardo, 1985), "likeability" and attractiveness (Landy et al., 1974, Cardy et al., 1986), student names (Lebuda et al., 2013), and perceived ability (Babad et al., 1975; Babad, 1980). Other studies exploring the Halo Effect found effects persisting across multiple assignments from the same student (Dennis, 2007), and that this effect was specifically identified in cases where teachers were assessing student writing samples (Forgas, 2011). One possible solution proposed to mitigate teacher bias is the anonymization of student identity during subjective grading (Malouff et al., 2013; 2014), motivating the current work.

### *2.1 Open-Ended Problems in ASSISTments*

In this paper, we analyze the teachers grading open-ended mathematics problems in ASSISTments. ASSISTments is a CBLP that allows teachers to assign content (primarily in middle-school mathematics) and monitor student progress. The system provides students with immediate correctness feedback on close-ended problems and offers computer-provided help in the form of on-demand help and scaffolding. Open-ended problems in ASSISTments are available in two formats: (a) sub-problem of a multipart problem or (b) primary problem. Figure 1(a) shows an open-ended problem presented as a subpart of a multipart problem, and figure 1(b) shows an open-ended problem presented as the main problem. A multipart problem can have more than one open-ended sub-problem as well.

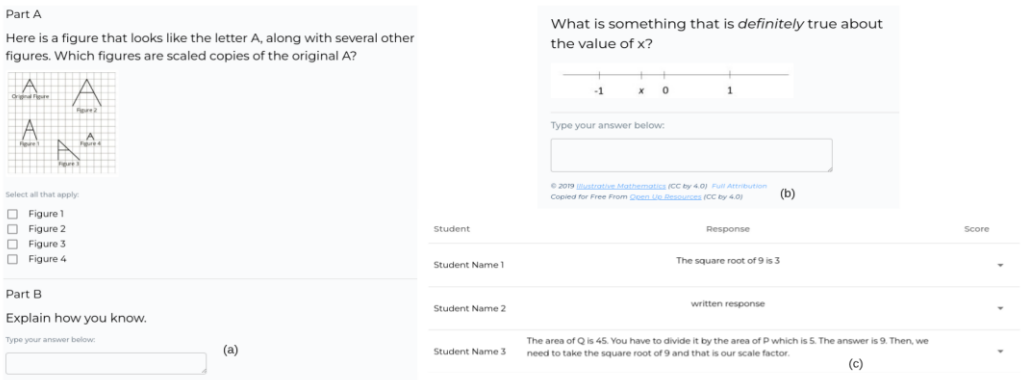


*Figure 1.* The different types of open-ended problems in ASSISTments.

Once students have completed their assignments, teachers can grade student responses. Figure 1 (c) shows the interface teachers can use to grade their students' responses. Teachers have the option to anonymize their students during the grading process where their identity is hidden, and the responses are shuffled, but the default behavior is to show identifying information of the students.

## 3. Study 1: Examining Grading Differences When the Student is Anonymized

Our first analysis explores whether teachers assess their students differently when they know the student's identity compared to when students are anonymous. In a purely unbiased, impartial scenario, teacher grading behavior would be solely dictated by the quality of the response as determined by how well the student was able to articulate their thoughts and demonstrate their knowledge of the concepts; this may include aspects such as grammar, use of mathematical terms (in mathematics contexts), and overall completeness. However, we posit there may be other factors that teachers consider including effort and prior performance. Even from a motivational perspective, a teacher may be inclined to bias

their grade in a positive direction for students who normally under-perform but applied notable effort as a way to encourage similar behavior in the future (i.e. an arguably positive example of bias). The danger, however, is in whether these perceptions, implicit or explicitly applied during the grading process, impact the types of opportunities that may be available to a student.

We conducted a study involving 9 teachers who commonly assigned and graded open response problems in ASSISTments (14 teachers were part of a larger study, but only 9 participated in this portion). We selected 3 problems containing an open-response sub-part that was assigned within the month prior to the study. Each teacher had already assigned and graded their student work for at least one of the three problems using the default scoring paradigm (i.e. teachers scored while knowing the identity of each student). In the month after this scoring was completed, we collected all student responses from across all teachers, anonymized them, and then randomly selected approximately 25 student responses to present to each teacher, ensuring that at least 10 of the responses were from each respective teacher's own students (if fewer than 10 were randomly selected, the difference was sampled from that teacher's student responses and added to the 25 given to the teacher); duplicate responses (e.g. empty responses or answers of "I don't know") were removed so that each teacher had a set of unique answers (resulting in some teachers having fewer than 10 responses from their own students if one was a duplicate). Each teacher was then asked to grade the given set of student answers, such that, for their own students, they would be anonymously grading the same set of answers as they had non-anonymously in the past (the additional responses and amount of time between grading reduced the likelihood that teachers would recognize their own students' responses). This presented the opportunity to measure each teacher's intra-rater reliability (i.e., how well they agreed with their past selves) and whether their grading was biased in a particular direction when they knew the student compared to not.

As the grading was done on a 5 points scale of 0-4; we applied Weighted Cohen's Kappa with linear weights to measure the variation in teachers grading behavior per response and found that the agreement coefficient as low as k=0.2293 and as high as k=0.7368, indicating that there was a large degree of disagreement between the two-time points, as shown in Table 1. These resulting scores were notably lower than we had initially hypothesized, suggesting that there were large differences in how teachers approached the grading of these when students were anonymized.

Table 1. *Exploring the grading behavior of teachers when they had access to students' identity vs. when students were anonymized using Linear Weighted Cohen's Kappa.*

| Teacher | N | Intra-rater reliability (Weighted Cohen's Kappa) | Intra-rater reliability (Relaxed Cohen's Kappa) | Avg. grade diff (Initial - anonymized) |
|---|---|---|---|---|
| Teacher1 | 10 | 0.3878 | 0.8077 | -0.2 |
| Teacher2 | 10 | 0.7368 | 1 | 0.2 |
| Teacher3 | 10 | 0.3939 | 0.6269 | -0.2 |
| Teacher4 | 10 | 0.4737 | 0.7368 | 0.3 |
| Teacher5 | 11 | 0.2293 | 0.443 | 0.27 |
| Teacher6 | 19 | 0.3596 | 0.5662 | 0.57 |
| Teacher7 | 9 | 0.3571 | 0.3571 | 0.44 |
| Teacher8 | 10 | 0.4531 | 0.5312 | 0.3 |
| Teacher9 | 9 | 0.6539 | 0.761 | -0.66 |

In addition to the Kappa measure, we also observed a relaxed calculation of Weighted Kappa. Given that grades are given on a 5-point scale and a teacher's assessment may reasonably vary by a small degree, we observe intra-rater agreement with an off-by-one adjustment (e.g. a scoring difference of one in either direction is treated as the same score when calculating Kappa). This adjustment resulted in notably higher Kappas, suggesting that the overall difference of scores was not as large as the first Kappa value suggested; while there was low precise agreement, teachers were relatively consistent within a grade point of themselves when the student was anonymized.

Pairing these Kappa scores with the average grade difference (the right-most column of Table1), we see that there is apparent bias in a particular direction. The positive difference exhibited by the majority of teachers suggests that teachers were more likely to grade anonymous students lower, on

average, than when they knew the students. We conducted a permutation test across teachers to estimate the average difference in teacher grading behavior when students were anonymous versus not. We observed that teachers on average were more likely to give higher scores, 0.163 with 95% CI= [-0.1367, 0.4627] when students were not anonymized; while not statistically significant, likely due to our small sample size, this result is suggestive that there was some bias observed in the study.

We followed this empirical analysis with a set of semi-structured qualitative interviews first with the teachers as a group, and then with 2 teachers individually for an extended session to gain better context as to how they approach grading and why they believed their grades changed during the study. Overall, the teachers unanimously described considering contextual information of the student when approaching the grading process. Several teachers mentioned that they consider motivational aspects (e.g., trying to encourage students to apply effort) when determining grades, as we had initially hypothesized. Several teachers also mentioned attempting to consider student effort when grading but did not use action-level reports in the system to do so (suggesting that it would be too time-consuming). The interviewed teachers similarly acknowledged potential risks to fairness highlighted by the observed differences, with one teacher expressing the intention to always grade anonymously following the study; others disagreed with this course of action citing perceived benefits of understanding the context of student work in order to provide better feedback to students in conjunction with the grade. While arguably anecdotal due to the limited sample size of teachers involved in this study, these conversations highlight the presence of bias in the form of a Halo Effect where teachers have an inherent motivation to give higher grades to perceived lower-performing students (again, not necessarily in a manner that affects fairness, but is still a form of bias).

## 4. Study 2: Exploring Related Factors of Student Assessment

Building on our findings from the pilot study we conduct a quantitative analysis investigating the role of student identity in the grading behavior of teachers. In this section, we attempt to explore the relationship between various answer- and student-level features and teacher-provided grades. Specifically, we seek to address the second and third research questions by exploring 1) whether prior student performance is a strong predictor of grade after accounting for concept-knowledge and other answer-level features (such as, for example, the number of words in the response), and 2) whether a measure of student effort correlates with the grades they ultimately receive while controlling for other measures of knowledge and ability. These analyses are meant to collectively provide insights into what a teacher may consider when assessing student work.

### *4.1 Description of the Dataset*

We collected a dataset of authentic student responses to open-ended problems from the ASSISTments platform. The dataset consists of action logs of students interacting with open-ended problems for the academic years of 2018 through the beginning of 2020 (i.e., up to but excluding the period of remote learning in response to the COVID-19 pandemic). A notable portion of open-ended problems in the data are from open educational resources (OER).

The dataset contains action logs for open-ended problems assigned in the system. Overall, the dataset includes 344,847 action logs from 7,535 students working on 2,268 problems within 2,636 assignments. It is important to highlight that it is additionally the case that problems can contain multiple parts; particularly in the OER content, it is common for open-ended questions to exist as a sub-part of a multi-part problem (e.g., asking students to explain their reasoning after solving a closed-ended question of the same concept). Students worked on 3,404 distinct open-ended problems, reflecting that many problems contained open-response questions for multiple sub-parts. Grades for the student responses observed in this work followed a 5-point integer scale ranging from 0-4.

Following other work that observed response time as a measure of effort (see Gurung et al., 2021), we use student action logs to calculate how much time on task formulating their open response answers as a measure of effort. ASSISTments record three types of actions that are of interest to our analyses: starting a problem, leaving without answering to resume later, and submitting a response. We combine these actions into action pairs to compute the amount of time a student spent formulating their response to the problems, accounting for cases of students leaving and resuming work on the problem. These action pairs can be described using the notation of "(first action, second action)" where they

represent two consecutive actions of a student taken within a session. The time for the action pairs represents the amount of time a student took between the two recorded actions. Our dataset observed two primary action pairs: "Problem Started-Submitted Response" and "Problem Resumed-Submitted Response" distinguished by the action observed prior to students submitting their response.

With this measure of student time-on-task, we apply a log transformation to create a pseudo-normal distribution and remove samples with a z-scores outside the range [-3, 3]; this filtering step attempts to remove very large outliers that may impact or bias our results. We also examined the open response problems and found that teachers graded only 19,446(~20%) of the 97,105 problems. The resulting number of action pairs used in our analyses for graded open responses are in table 2.

Table 2. *Filtered action pairs for students with open responses*

| Action pairs | Graded responses | Ungraded Responses |
|---|---|---|
| (Problem Started, Submitted a response) | 18295 | 73176 |
| (Problem Resumed, Submitted a response) | 1151 | 4483 |

There are several features of student answers that are likely to correlate with teacher-provided grades as identified in Baral et al. (2021). These features include the "Response Category," a categorical variable that indicates that the student response contains only words (positive class) as opposed to a mixture of linguistic and mathematical terms and expressions (negative class), as well as the "Number of Words," a simple count of the number of words (as denoted by spaces) in the student's answer.

In addition to the answer-level features, we calculate several student-level features to describe recent and historic performance measures. These measures include "Prior Percent Correct," the average correctness for all problems attempted by the student prior to beginning the open response problem (representing a long-term measure of general mathematics ability), and "Prior Sub-Part Performance," the average correctness across all prior sub-parts of the problem containing the open response question (representing content-specific knowledge).

Additional features were calculated including problem difficulty, the number of prior problem sub-parts, the number of prior problems started by the student, and the number of help requests made on earlier sub-parts, but these were found to either be highly correlated with other measures or not correlated with our outcomes of interest and were therefore omitted from our analyses. All pairwise Spearman correlations of features were calculated, omitting features introducing risks of collinearity.

The inclusion of both long- and short-term performance measures helps to distinguish and control for knowledge of the given mathematical concept as compared to general student ability. While initially hypothesized to be highly correlated, these measures ultimately exhibited a low correlation ($r = 0.014$) and could be used in our analyses without introducing risks of collinearity. Presumably, performance on the prior subparts should be a meaningful predictor of student performance on the open response problem given that they both pertain to the same mathematical concept. If while controlling for this it is found that the student's prior percent correct is similarly predictive, while not alone causal, such a result would suggest that a teacher may take prior student ability into account when grading the open response.

### *4.2 Factors Related to Student Grade*

We use regression analysis to investigate the relationship between the described measure of effort (as measured by time-on-task), teacher-provided grades, and answer- and student-level features using a linear regression model (LM). We additionally include an interaction term of prior sub-part performance $\times$ the number of words.

From the regression results reported in Table 3, we found that the model ($R^2 = 0.133$) showed both prior sub-part performance and prior percent correct were reliable and meaningful predictors of student grade. This suggests that both student ability and content knowledge predict student grades. We also found that students with linguistic (word response category) responses correlate with lower scores as compared to responses that contained mathematical terms and expressions. The observed interaction term is found to be statistically reliable, but the low coefficient suggests that the relationship of this term is not very meaningful in comparison to the other more impactful features.

Table 3. *Linear Regression and Mixed-Effect model coefficients observing assessment score.*

| | **LM ($R^2$ = 0.133)** | | **MLM** | |
|---|---|---|---|---|
| | | | Variance | Std.Dev. |
| teacher | | | 0.4523 | 0.6725 |
| | coefficient | std. errs | coefficient | std. errs |
| intercept | 0.2803 | 0.157 | 1.1648*** | 0.2226 |
| response category (Words) | - 0.3927*** | 0.033 | - 0.3924*** | 0.0322 |
| prior sub-part perf. | 1.3684*** | 0.088 | 1.3054*** | 0.0827 |
| words per answer | 0.0332*** | 0.006 | 0.0393*** | 0.0053 |
| prior sub-part perf. × words per answer | - 0.0185* | 0.006 | - 0.0189** | 0.0057 |
| log transformed time on task | 0.0662* | 0.02 | 0.0323 | 0.0202 |
| prior percent correct | 1.275*** | 0.172 | - 0.1497 | 0.2127 |

LM: Linear Model, MLM: Mixed Linear Model, Significance: *** < 0.000, **<0.001, * < 0.05

Additionally, we extended the LM by introducing the teacher as a random effect in a mixed-effect linear model (MLM), also reported in Table 3. The grading process, as reported by the teachers in our pilot study, accounts for students' perceived ability. In this model, the prior percent correct measure was no longer a significant predictor of student grade in the MLM suggesting that longer-term performance is not a prominent factor considered by teachers when assessing students. Prior sub-part performance, as a measure of concept knowledge, however, was still a reliable and meaningful predictor of student score.

## 5. Study 3: Potential Impacts of Fatigue on Grading

To address the final research question, we conduct one last analysis to observe how fatigue may affect teacher grades. In other words, we explore whether the ordering in which teachers grade student work leads to any potential risk of unfairness or bias due to implicit sequence or temporal effects. Particularly when a teacher has a large number of students to grade, it may be difficult to grade consistently for all students even when using a rubric. Particularly with the amount of time and attention needed to assess student open response answers, a teacher may find it difficult to give the same amount of attention to the 50th student as they do to the 1st student on a given assignment. Similarly, teachers may grade more strictly or leniently due to unconscious comparisons with previous students (e.g. a mid-grade student response may look better when assessed after a low-grade student response).

To explore this, we use an expanded dataset collected from 2018 to January 2020 to observe the mean and variance of grades over the course of observed grading sessions. This dataset contains 219,189 graded student open responses across 5562 problems from 3847 assignments.

Understanding that teachers may not grade all students for a given problem or assignment in one sitting, we find the order in which teachers graded student answers for each problem on a given day and plot the distribution of grades over this session ordering. Teachers who graded more than 50 students within a single span of time were omitted due to the sparsity of data. We visualize this trend in Figure 5 to observe whether fatigue appears to exhibit any temporal effects. If fatigue did affect how teachers grade we would expect to see trends that either affect the mean of teacher scores (rising or falling, on average, as teachers grade more students over time) or the standard deviation of scores.

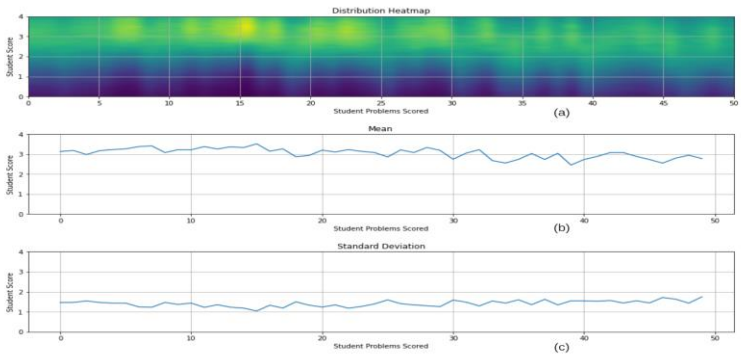


*Figure 5.* Visualizing teacher grading patterns as they grade sequences of student responses

From figure 5, there is little evidence that teachers' grading pattern changes significantly over time both in terms of mean as well as variance (i.e. the width of the distribution does not appear to change significantly, apart from the notable decrease in sample size as the number of problems scored extends above 25-30). While this does not speak to fatigue being an issue for individual teachers or scenarios, it does suggest that fatigue exhibits a low risk in terms of fairness and bias on average.

## 6. Discussion

Across the three studies presented in this work, we have attempted to gain a better understanding of how teachers approach the assessment of student open-ended work and identify factors that may impact given scores. As it pertains to addressing issues of fairness, the collective results of our analyses provide evidence that, somewhat unsurprisingly, teachers do consider contextual information beyond that pertinent to a given answer when assessing student work; this consideration does seem to bias grades in a positive direction, as suggested by the study observing anonymous grading. Qualitative data from our interviews support this, as well as the desire to utilize effort-based measures to even further inform student assessment practices.

From our regression analyses in Study 2, we were able to establish that teachers do not, on average, consider students' general mathematics ability as much as demonstrated knowledge on the given skill (i.e. in that the measure of concept knowledge was still found to be statistically significant after controlling for teacher effects, while the longer-term outcome was not. This is a little surprising particularly because several teachers from our first study described considering historic student performance when approaching scoring, largely from a motivational perspective for students who typically under-perform compared to their peers.

Finally, though we had initially hypothesized that grading fatigue may be a factor that affects how teachers grade, we found little evidence of this. We found that the mean and standard deviations of teacher scores remained relatively consistent while grading multiple students consecutively.

As we do identify some amount of bias in how teachers grade students, it is important to reiterate that this does not necessarily equate to teachers following unfair or inequitable assessment practices. At the current stage of this work, we are ill-equipped to say whether the measured bias is likely to contribute to impacts on the types of opportunities that students may receive; we can say from these studies that teachers, on average within the context observed in this paper, seem to be considerate of a range of factors that describe student performance, and are seemingly not largely impacted by obvious negative factors such as fatigue.

## 7. Limitation & Future work

As we identified certain biases exhibited in teacher assessment, it is important to explore whether these biases are more prevalent among certain student populations than others to identify deeper risks to student fairness. These populations may refer to protected demographic labels such as race, gender, ethnicity, or other geographic descriptors, but also latent groups of students. Similarly, while no meaningful effect of grading fatigue was observed across all teachers, this may still be an issue in more individualized cases.

As certain unmeasured recency or other timing effects may impact how teachers score (e.g. as the teachers re-scored student work from a month prior, and had undoubtedly moved on to new content areas), future work could also replicate this study to randomize when anonymization occurs within the study design.

Among the largest limitations of the current work is the correlational nature of the analyses presented. While we have attempted to pair some of our results with insights from experienced teachers, the causal mechanisms impacting student scores, particularly in reference to those factors external to the student's response itself, could be further explored through additional future studies. It is important to understand what is considered when assessing students so that teachers, education researchers, policymakers, and learning system developers can start to address the questions as to what should be considered and what factors may lead to unnecessary risks to fair student assessment.

## 8. Conclusion

While we are able to identify bias and student-level factors that seemingly correlate with student grades, it is uncertain as to whether such bias is truly negative in regard to fair student assessment. While we did not find any obvious risks to fairness within this presented set of analyses, this work represents a step toward identifying and mitigating such risks in educational assessment. This work further attempts to emphasize the distinction between bias and fairness, recognizing that the consideration of contextual information can have many positive benefits in terms of student learning, motivation, and engagement. However, it is equally important to pursue further study of these issues to determine whether such benefits implicitly lead to inequitable opportunities.

## 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

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking. In Intl. Conf. on ITS

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning envs. Review of educational research, 73(3), 277-320.

Ostrow, K. S., & Heffernan, N. T. (2015, June). The role of student choice within adaptive tutoring. In Intl. Conf.on AI in Edu. Springer, 752–755.

Patikorn, T., & Heffernan, N. T. (2020, August). Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In Proceedings of the Seventh ACM conf. on Learning@ Scale.

Biancarosa, G., & Snow, C. E. (2004). Reading next: A vision for action and research in middle and high school literacy: A report from Carnegie Corporation of New York. Alliance for Excellent Edu.

Graham S., MacArthur C. A., & Fitzgerald J. (Eds.). (2013). Best practices in writing instruction. Guilford press.

Douglas N Walton. 1992. Plausible argument in everyday conversation. SUNY Press.

Livne, N. L., Livne, O. E., & Wight, C. A. (2008). Enhanching Mathematical Creativity through Multiple Solution to Open-Ended Problems Online.

Silver, E. A. (1995). The Nature and Use of Open Problems in Mathematics Edu.: Mathematical and Pedagogical Perspectives. Zentralblatt fur Didaktik der Mathematik/Intl. Reviews on Mathematical Edu., 27(2), 67-72.

Cooper, W. H. (1981). Ubiquitous halo. Psychological bulletin, 90(2), 218.

Spear, M. G. (1984). The biasing influence of pupil sex in a science marking exercise. Research in Science & Technological Edu., 2(1), 55-60.

Roen, D. (1992). Gender and teacher response to student writing. In Gender issues in the teaching of English (pp. 126-141). Heinemann.

Wen, S. S. (1979). Racial halo on evaluative rating: General or differential? Contemporary Educational Psychology, 4(1), 15-19.

Fajardo, D. M. (1985). Author race, essay quality, and reverse discrimination. Journal of Applied Social Psychology, 15(3), 255-268.

Lebuda, I., & Karwowski, M. (2013). Tell me your name and I'll tell you how creative your work is: Author's name and gender as factors influencing assessment of products' creativity in four different domains. Creativity Research Journal, 25(1), 137-142.

Babad, E. Y. (1980). Expectancy bias in scoring as a function of ability and ethnic labels. Psychological Reports, 46(2), 625-626.

Babad, E. Y., Mann, M., & Mar-Hayim, M. (1975). Bias in scoring the WISC subtests. Journal of Consulting and Clinical Psychology, 43(2), 268.

Malouff, J. M., Stein, S. J., Bothma, L. N., Coulter, K., & Emmerton, A. J. (2014). Preventing halo bias in grading the work of university students. Cogent Psychology, 1(1), 988937.

Malouff, J. M., Emmerton, A. J., & Schutte, N. S. (2013). The risk of a halo bias as a reason to keep students anonymous during grading. Teaching of Psychology, 40(3), 233-237.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. Journal for research in mathematics education, 39(4), 372-400.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. The elementary school journal, 105(1), 11-30.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. Intl. Journal of AI in Education, 24(4), 387-426.
Calvo, R. A., O'Rourke, S. T., Jones, J., Yacef, K., & Reimann, P. (2010). Collaborative writing support tools on the cloud. IEEE Transactions on Learning Technologies, 4(1), 88-97.
Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. Intl. Journal of AI in Education, 8(1), 30-43.
Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster teaching via pomdp planning. Cognitive science, 40(6), 1290-1332.
Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. Handbook of automated essay evaluation: Current applications and new directions, 55-67.
Gurung, A., Botelho, A. F., & Heffernan, N. T. (2021, April). Examining Student Effort on Help through Response Time Decomposition. In LAK21: 11th Intl. Learning Analytics and Knowledge Conf. (pp. 292-301).
Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Intl.Journal of AI in Education, 24(4), 470-497.
Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-Based Writing Instruction. Grantee Submission.
Roscoe, R. D., Allen, L. K., & McNamara, D. S. (2019). Contrasting writing practice formats in a writing strategy tutoring system. Journal of Educational Computing Research, 57(3), 723-754.
Price, T., Zhi, R., & Barnes, T. (2017). Evaluation of a Data-Driven Feedback Algorithm for Open-Ended Programming. Intl. Educational Data Mining Society.
Wiggins, J. B., Boyer, K. E., Baikadi, A., Ezen-Can, A., Grafsgaard, J. F., Ha, E. Y., ... & Wiebe, E. N. (2015, February). JavaTutor: an intelligent tutoring system that adapts to cognitive and affective states during
computer programming. In Proceedings of the 46th acm technical symposium on computer science education (pp. 599-599).
Bhatnagar, S., Lasry, N., Desmarais, M., & Charles, E. (2016, September). Dalite: Asynchronous peer instruction for moocs. In European Conf. on Technology Enhanced Learning (pp. 505-508). Springer, Cham.
Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008, September). PeerWise: students sharing their multiple choice questions. In Proceedings of the fourth intl. workshop on computing education research (pp. 51-58).
Cambre, J., Klemmer, S., & Kulkarni, C. (2018, April). Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In Proceedings of the 2018 CHI Conf. on Human Factors in Computing Systems (pp. 1-13).
Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K. Z., Lasecki, W. S., & Heffernan, N. (2016, April).
Axis: Generating explanations at scale with learnersourcing and machine learning. In Proceedings of the Third (2016) ACM Conf. on Learning@ Scale (pp. 379-388).
Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2021). Improving Automated Scoring of Student Open Responses in Mathematics. Intl. EDM Society.
Erickson, J. A., Botelho, A. F., McAteer, S., Varatharaj, A., & Heffernan, N. T. (2020, March). The automated grading of student open responses in mathematics. In Proceedings of the Tenth Intl. Conf. LAK
Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. Journal of personality and social psychology, 35(4), 250.
Nieva, V. F., & Gutek, B. A. (1980). Sex effects on evaluation. Academy of management Review, 5(2), 267-276.
Martin, W. D. (1972). The sex factor in grading composition. Research in the Teaching of English, 6(1), 36-47.
Piché, G. L., Michlin, M., Rubin, D., & Sullivan, A. (1977). Effects of dialect-ethnicity, social class and quality of written compositions on teachers' subjective evaluations of children. Comms. Monographs, 44(1), 60-72.
Dennis, I. (2007). Halo effects in grading student projects. Journal of Applied Psychology, 92(4), 1169.
Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. Journal of Personality and Social Psychology, 29(3), 299.
Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. Journal of applied psychology, 71(4), 672.
Forgas, J. P. (2011). She just doesn't look like a philosopher…? Affective influences on the halo effect in impression formation. European Journal of Social Psychology, 41(7), 812-817