Board 56: Using anonymous grading for high-stakes assessments to reduce performance discrepancies across student demographics

Dr. Neha B. Raikar, University of Maryland Baltimore County

Dr. Raikar is a Lecturer at the University of Maryland, Baltimore County in the Chemical, Biochemical, and Environmental Engineering department. She has taught both undergraduate and graduate-level courses. Dr. Raikar also has 3 years of industry experience from working at Unilever Research in the Netherlands.

Dr. Nilanjan Banerjee

Nilanjan Banerjee is an Associate Professor at University of Maryland, Baltimore County. He is an expert in mobile and sensor systems with focus on designing end-to-end cyber-physical systems with applications to physical rehabilitation, physiological mon

Work-In-Progress: Feasibility of anonymous grading for reducing performance discrepancies across student demographics

Neha Raikar[1] and Nilanjan Banerjee[2]
[1]Department of Chemical, Biochemical, and Environmental Engineering
[2]Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County

Introduction/Motivation

Exams and quizzes are critical tools for evaluating the classroom performance of students. As a result, grading methods for these exams and quizzes are central to determining student ranks, final letter grades, and cumulative grade point averages (GPA). The student GPA and letter grades are important metrics used to gauge student success. Unfortunately, a human (either a teaching assistant or an instructor) performs grading in most classes. Humans suffer from implicit bias, and grading is no exception. Conscious and unconscious bias in grading is a common problem in academia. Unconscious bias can take various forms, such as gender, ethnicity or performance bias. Due to this bias, differences in performance have been reported in underrepresented minorities [1]. For example, researchers have noticed the *Halo effect,* where positive perceptions about a student from past work result in a higher grade [2]. These include but are not limited to politeness, good performance in previous courses, asking questions, or hard work. In contrast, a Horn effect exists where negative student traits like absence from classes, disruptive behavior, etc., lead to poor grades [3]. Additionally, preconceived stereotypical notions about specific ethnic groups can manifest as bias in grading. We hypothesize that anonymous grading can lead to a reduction, if not elimination, of implicit bias during grading. Moreover, anonymous grading improves the *fairness perception* among students, especially underrepresented minority students.

Literature on anonymous grading is focused on peer assessments [4],[5] and student papers. While there is some literature on the efficacy of anonymous grading in different fields, such as medicine [6], there is a lack of studies that focus on understanding the usefulness of anonymous grading for ***in-person exams or quizzes*** in engineering. While learning management systems such as Blackboard offer anonymous grading, the tool only applies to online or electronic submission exams. Tools like the Akindi bubble sheet only offer anonymous grading for multiple-choice questions [7]. Auto graders are available for programming-based assignments and are anonymous. But they need the submission to follow stringent requirements. An e-testing platform has been employed for essay-based grading that offers anonymity [8]. However, there are no tools, to the best of our knowledge, that can allow anonymous grading for in-class paper exams and quizzes which form a majority of exams on campus.

Approach and Plan of Work

Our proposed work has three distinct components listed below.

- Development of a mobile system that helps instructors perform anonymous grading for paper exams

- Data collection in courses and statistical analysis to understand grade differences using anonymous and non-anonymous grading.
- Self-reporting data collection to understand the student and faculty perspective on anonymous grading.

Mobile system that helps instructors perform anonymous grading

The proposed workflow for the mobile application of anonymous grading is shown in figure 1. A web application will input the class roster and generate a mapping between the student's name and an alphanumeric code stored in a backend database. The web app will also allow the printing of a barcode and student name stickers for affixing on the student's exams. The barcodes and student names will be affixed on separate pages, and the name page will be collected after handing out the exam to the corresponding student. We have so far implemented this part of the system.

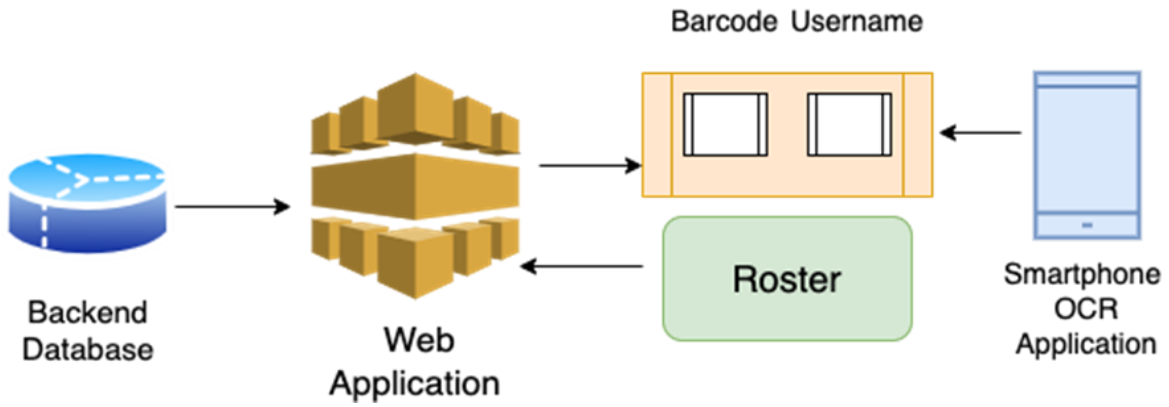

Figure 1: The figure illustrates our proposed tool for anonymous grading.

The next step will happen after exams have been administered and graded. The graded exams need to be decoded. An OCR (Optical Character Recognition) based app will be used to read the barcode and the grade and create a grade sheet after reverse mapping the barcode with the student's name. This part is still under development. For our current data analysis, we performed manual decoding. The manual reverse mapping is time-consuming and not scalable for large classes. Hence, we will be focusing on the development of the reverse mapping application next.

Perform data collection and statistical analysis

Figure 2 highlights the average course GPA for a Spring 2022 course in Chemical Biochemical and Environmental Engineering course that one of the authors taught. The figure shows that there is a difference in the mean GPA across various ethnicities. The GPA difference can be a result of various factors such as prior preparation, semester course load, jobs, etc., in addition to bias in grading. The goal of this project is to reduce the performance discrepancies by reducing the grading bias.

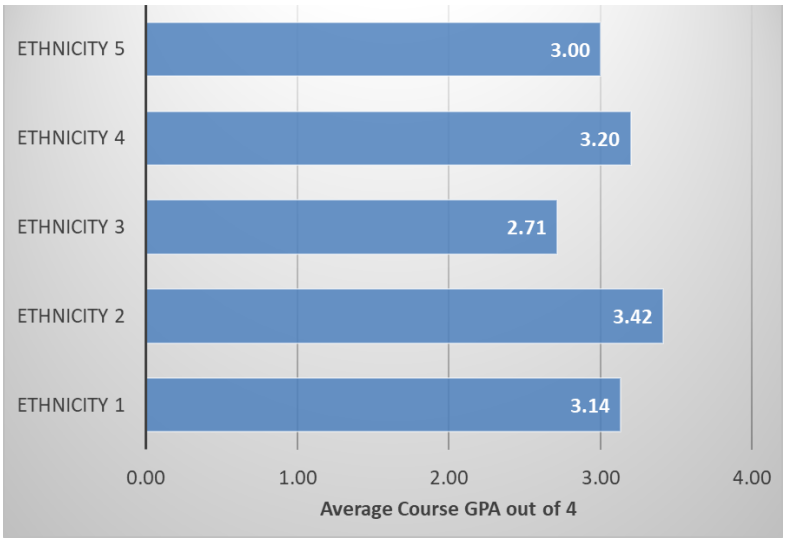

Figure 2: Average course GPA across different ethnicities for a Spring 2022 course.

During the Fall 2022 semester, we started data collection with anonymized barcodes. Since the class was small, anonymous exams were administered to all students. Figure 3 shows the demographic distribution for one of the classes in Chemical, Biochemical, and Environmental Engineering (ENCH 620) for which anonymous grading was administered. This is a graduate-level engineering course. The exams for the course use free response and calculation questions and is administered in person. The class size we tested was small (14 students), but it was a good size to test out the initial implementation of the system. The system was also implemented in two other classes, CMSC 621 (graduate) and ENCH 225L (undergraduate), but the analysis shown is only for ENCH 620.

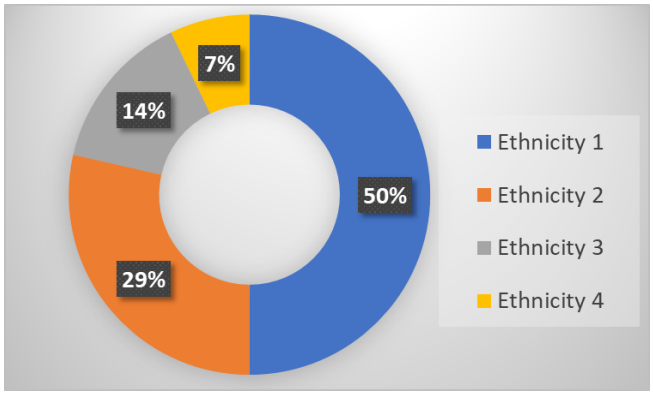

Figure 3: Class demographics for one of the Fall 2022 courses testing the anonymous grading

Preliminary analysis of the average grade on various exams for different ethnicities is compared in figure 4. The figure shows little to no variation in the average grade among the different demographics compared. Exams 1-3 and the final exam were in-person and anonymous for all students. However, due to the format of the coding exam, it was handled anonymously. There is a slight drop in the performance of ethnicity 3 for Exam 2, but that group recovered quickly on subsequent exams.

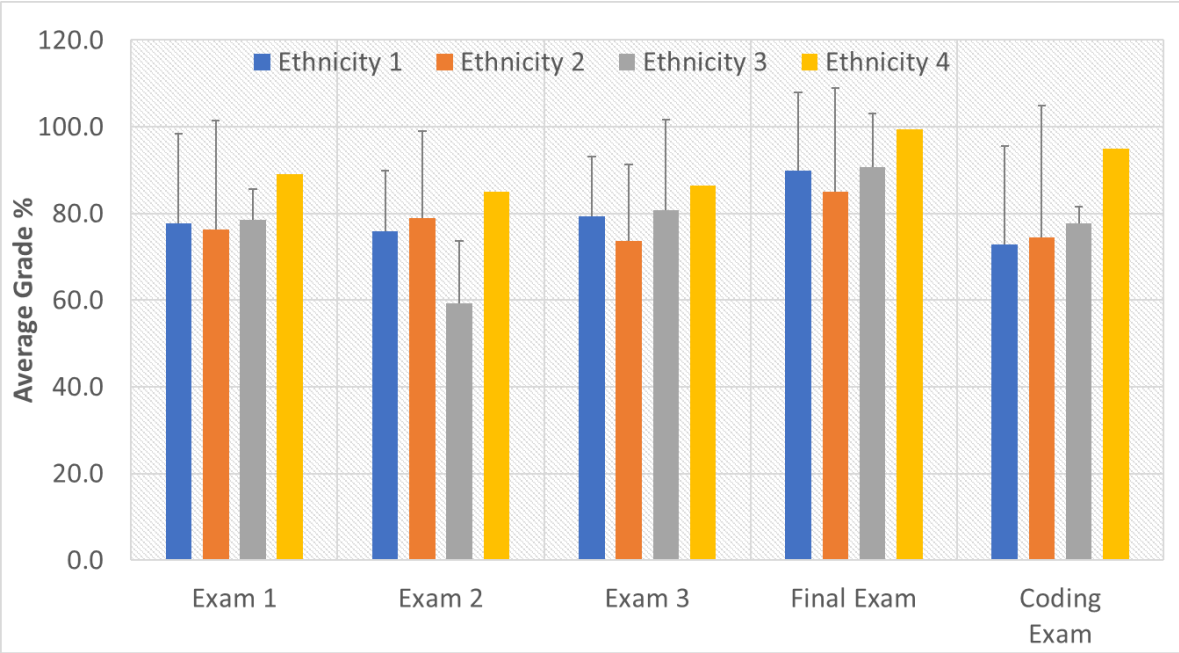

Figure 4: Average grade (%) across various ethnicities compared on different exams. Exams 1 -3 and the final exam were administered anonymously for all students. The coding exam was not anonymous. The error bars exceed 100% due to the extra credit available on the exams. Ethnicity 4 has no standard deviation since there was only one student in that category.

As mentioned earlier, in the Fall semester, we did anonymous testing on all the students due to the smaller class size. However, in the current semester, we want to utilize the A/B testing methodology. We will randomize the students into two groups, one anonymous test-takers, and the other non-anonymous test takers. We will also alternate the two groups across various exams so that all students get to experience anonymous grading for at least two exams. Table 1 shows the proposed layout for this data collection methodology. We will perform statistical analysis on the two groups, namely, a two-sample hypothesis testing to see if there is a significant difference in the performance of the two groups. We will also evaluate the student performance across demographics, ethnicity, gender, etc., similar to the one highlighted in figure 3.

Table 1: A/B Testing methodology to be used for various exams. The group shaded green will receive the test anonymously, and the other group non-anonymously

Exam 1	***Exam 2***	***Exam 3***	***Final Exam***
Group A	Group B	Group A	Group B
Group B	Group A	Group B	Group A

Self-reporting data collection to understand the student and faculty perspective on anonymous grading.

Once we have successfully tested out our platform for anonymous grading, we would like to survey students for their perception of the tool and its efficacy. We believe that anonymous

grading will have a positive reinforcement effect on students as it, by definition, implies that no factors other than the solution of the exam will be used for grading. To test this hypothesis, we will use a questionnaire on student perceptions of anonymous grading and reflections on their performance. Specifically, we will ask the following questions:

1. Do you think or have observed that there is a bias in grading?
2. What did you think of the current implementation?
3. Did you feel your performance changed as a result of the grading anonymity?
4. What were the drawbacks of the implementation, and how can it be improved?

We also want to distribute the tool to other instructors and have them test it out in their respective classes. To get instructor feedback, we will administer similar surveys and focus groups. Programs such as Nvivo or Dedoos will be used to code the free responses. The quantitative questions will be tested for statistical significance as well. These data from students will be analyzed to understand subjectively whether students thought that their performance changed due to grading anonymity and why. We will apply for an IRB approval to conduct the surveys and the focus groups. The IRB will also allow us to perform detailed data analysis.

Our long-term goal is to make the tool available for free to the community and can be used across different disciplines for anonymous grading for in-person exams and quizzes. The data analysis will also be made freely available through the project website. We want to use the data collection to validate our hypothesis. However, even in the event that the data does not show statistically significant differences between anonymous and non-anonymous grading, the data will still be useful to demonstrate that non-anonymous grading is not malignant with implicit bias. The tool will be useful for engineering departments to demonstrate that implicit bias in grading may or may not be an issue.

Summary

The authors would like to highlight that this work focuses on anonymous grading for in-person exams and quizzes. Anonymous grading through the LMS like Blackboard is difficult unless there is a requirement to scan and upload the exams. Using our system, we hope to lower the barrier for instructors to employ anonymous grading. Grading bias, whether conscious or unconscious, can occur across various demographics like ethnicity, gender, sexual orientation, or even based on GPA. We have currently tested anonymous grading on the entire class due to small class sizes. But in the future, we will test our system using an A/B split, where we randomly assign the students with similar demographics to the control and the test group. This will enable us to compare performances on the same exam with or without anonymous grading. The authors are applying for IRB approval for conducting the surveys and focus groups.

Acknowledgments

A part of the work is supported by the Hrabowski Innovation Fund Award, which supports initiatives to enhance teaching and learning at UMBC.

References

1. Addy, Tracie Marcella, et al. *What inclusive instructors do: Principles and practices for excellence in college teaching*. Stylus Publishing, LLC, 2021.
2. Malouff, John M., Ashley J. Emmerton, and Nicola S. Schutte. "The risk of a halo bias as a reason to keep students anonymous during grading." *Teaching of Psychology* 40.3 (2013): 233-237.
3. Satyanarayana, Ashwin, Reneta Lansiquot, and Christine Rosalia. "Using Prescriptive Data Analytics to Reduce Grading Bias and Foster Student Success." *2019 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2019.
4. Kobayashi, Michiko. "Does anonymity matter? Examining quality of online peer assessment and students' attitudes." *Australasian Journal of Educational Technology* 36.1 (2020): 98-110.
5. Panadero, Ernesto, and Maryam Alqassab. "An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading." *Assessment & Evaluation in Higher Education* (2019).
6. Dorsey, J. Kevin, and Jerry A. Colliver. "Effect of anonymous test grading on passing rates as related to gender and race." *Academic medicine* (1995).
7. Akindi website: https://akindi.com/
8. Gusev, Marjan, Magdalena Kostoska, and Sasko Ristov. "A new e-Testing platform with grading strategy on essays." *2017 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2017.