Bayesian Grading: Analysis of Scores in the Context of Cheating

Victor Zhao
chenyan4@illinois.edu
University of Illinois
Urbana-Champaign

Yuxuan Chen
yuxuan19@illinois.edu
University of Illinois
Urbana-Champaign

Evan Matthews
evanmm3@illinois.edu
University of Illinois
Urbana-Champaign

ABSTRACT

TODO: Abstract

ACM Reference Format:
Victor Zhao, Yuxuan Chen, and Evan Matthews. 2024. Bayesian Grading: Analysis of Scores in the Context of Cheating. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

TODO: Introduction

Something about automated assessments and testing, resource limitations, unproctoring [**?**], untrustable grades, and our problem.

2 BACKGROUND

2.1 Computer-Based Assessments

In recent classroom settings, computer-based assessments has been observed to have many advantages over the traditional paper-based assessments [**?**]. One property of computer-based assessments is the ability to autograde students' work and provide instant feedback. A large body of research has shown that providing feedback in time can be beneficial to student learning [**?**]. In formative contexts, the ability to autograde problems and provide instant feedback also enables a mastery approach in courses [**?**]. With computer-based assessments, students also have the power to work on problems and submit their work asynchronously, enabling more flexibility in classrooms [**?**]. Automatic Item Generation [**?**] allows computers to generate problems that are similar to each other based on item models, providing a large problem pool students can practice with. The combination of Automatic Item Generation and autograding also allows for more courses to be offered at scale through online open assessment platforms, as the number of course enrollments increase significantly in recent years [**?**]. New technologies in development such as Automatic Short Answer Grading [**?**] are expanding the use for computer-based assessments to more types of problems. These features can enable computer-based assessments in almost any discipline.

2.2 Computer-Based Testing

Computer-based assessments can be used in summative testing contexts as well. Dishonesty is usually a large factor of the effectiveness of summative assessments [**?**]. Compared to the traditional pen and paper exams, students have more control over the environment in which they are taking the exam, but plagiarism and cheating prevention can also be more challenging because students can potentially have more resources (legal or illegal) available to them [**?**]. Some of the initial proctoring methods is to Proctoring the exams at a large scale poses a lot of logistical problems such as the trouble students need to go through to setup a testing environment using lockdown browsers or other similar technologies [**?**], and the difficulty in conducting frequent testing for classes and dealing with accomodations or conflict exams [**?**]. Training the course staff to be responsible proctors could also be time-consuming.

One solution proposed by researchers is to have a dedicated computer laboratory specifically for computer-based testing. This method was adopted and used to build the Computer-Based Testing Facility (CBTF) at the University of Illinois Urbana-Champaign, which has been utilized for an extended period of time [**?**]. The CBTF was built in a laboratory using computers configured for testing environments and hiring highly-trained proctors. This method has shown to be reasonable in terms of cost [**?**], and also effective in preventing cheating on a large scale [**?**].

However, with more classes incorporating computer-based assessments and computer-based testing, there might not be enough resources to have all exams proctored through the CBTF [**?**].

2.3 Bayesian?

3 EXPERIMENT

3.1 Context

The data for this study was drawn from a lower-division undergraduate computer science (CS) course over two semesters: Fall 2021 and Fall 2023. Students were expected to have completed an introductory algebra course as a prerequisite prior to taking this course. The course included 4 unproctored quizzes, 3 proctored exams, and 1 comprehensive proctored final exam. Each quiz and exam had a duration of 50 minutes, except for the longer, cumulative final exam. The quizzes were given one week before the exams and had overlapping

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

content, allowing exams to serve as a rough mirrored comparison to the quiz just before the exam. Students took the proctored exams within the CBTF. We gathered exam records for the 695 students enrolled in Fall 2021 and the 423 students enrolled in Fall 2023. After excluding students with test accommodations or those who didn't complete all the exams or quizzes, the final dataset included 622 student records from Fall 2021 and 405 student records from Fall 2023.

3.2 Method

Our proposed method analyzes student scores through policy simulation of computed Bayesian probabilities. First, probabilities of student scores are computed with respect to their observed ability $\mu \in \{0, 1, ..., 100\}$ and cheating boolean $c \in \{0, 1\}$ where $(c = 0), (c = 1)$ indicate not cheating and cheating, respectively. These parameters are used to compute

$$P(\text{obs} \mid \mu, c) = \frac{1}{N} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{a_i - \mu - \chi c}{\sigma} \right)^2 \right) \quad (1)$$

where a_i is the ith assessment of a student with parameters (μ, c), σ is the standard deviation of a student's assessments, and χ is a nonnegative "penalty" applied when a student has cheated. In order to distribute our data across student ability and/or cheating booleans, we use Bayes' Theorem to invert the relationship of our computed probabilities:

$$P(\mu, c \mid \text{obs}) = \frac{P(\text{obs} \mid (\mu, c)) \cdot P(\mu, c)}{\sum_{\mu', c'} P(\text{obs} \mid \mu', c')} \quad (2)$$

where $P(\mu, c)$ is the probability of a student's performance amongst their class [1], and our probability is normalized by the sum of previously computed probabilities $P(\text{obs} \mid \mu, c)$. The follow computations returns a probability array of shape (mu, c) from which we can visualize the distributions of cheating and non-cheating students.

finally, we define a policy $\pi_\theta(p)$ for simulating random variable $S = (\mu, c)$ and testing whether the correct outcome was returned:

$$\pi_\theta(p) = \begin{cases} 0 & p \leq \theta \\ 1 & p > \theta \end{cases} \quad (3)$$

The result of performing this simulation m times is a 2×2 truth table indicating percentages of true and false outcomes of cheating and non-cheating for the data.

3.3 Analysis

4 RESULTS

5 DISCUSSION

5.1 Limitations

While existing datasets of student assessments confirm the viability of Bayesian analysis, such information greatly reduces the extent to which the statistics can be considered trustworthy. In particular, our aforementioned analysis relies on the data being summative and free of statistically significant cheating attempts. Bayesian analysis on formative assessments is entirely possible, but the computation time drastically increases with respect to the context for which the analysis is performed. A weekly topic-by-topic basis is likely to provide more results with finer details given enough time, but two to three summative assessments can return similar results by individual problems while significantly decreasing computation time. Additional cheating attempts can also reduce the initial trustworthiness of scores and diminish Bayesian analysis's effectiveness. As such, our results appear highly volatile or inconclusive without a subset of reasonable scores for analysis.

6 CONCLUSIONS

[1] A uniform distribution of student performances can also be used here.