A Case for Bayesian Grading

Craig Zilles
University of Illinois
Computer Science
Urbana, IL, United States
zilles@illinois.edu

Chenyan Zhao
University of Illinois
Computer Science
Urbana, IL, United States
chenyan4@illinois.edu

Yuxuan Chen
University of Illinois
Computer Science
Urbana, IL, United States
yuxuan19@illinois.edu

Evan Michael Matthews
University of Illinois
Computer Science
Urbana, IL, United States
evanmm3@illinois.edu

Matthew West
University of Illinois
Mechanical Science & Engineering
Urbana, IL, United States
mwest@illinois.edu

Abstract

Academic integrity continues to be an issue in education. Students' grades are often computed using a collection of evidence that varies in its trustworthiness (e.g., a proctored exam can be trusted more than an out-of-class programming project) due to practical constraints. When a student cheats, their trusted and less trustworthy scores are inconsistent, which presents instructors a choice between rewarding the cheating behavior and the burden of investigating / making cheating allegations.

In this position paper, we propose that Bayesian inference might be a useful tool in assigning grades derived from trusted and less trusted evidence. Rather than compute grades by performing arithmetic on both trusted and untrusted assessments, we instead try to infer a latent variable, the student's mastery of the course material, from these observed performances and their potential for cheating. Key to this approach is that grades can be assigned that discount suspicious work without needing to explicitly make a cheating allegation. A logical conclusion of this approach is that the needed amount of trusted assessments for a given student depends on how inconsistent are their trusted and untrusted assessments.

CCS Concepts

• **Social and professional topics → Student assessment**.

Keywords

grading, Bayesian inference, cheating, trust

ACM Reference Format:
Craig Zilles, Chenyan Zhao, Yuxuan Chen, Evan Michael Matthews, and Matthew West. 2024. A Case for Bayesian Grading. In *Proceedings of the 2024 ACM Virtual Global Computing Education Conference V. 1 (SIGCSE Virtual 2024), December 5–8, 2024, Virtual Event, NC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3649165.3703624

SIGCSE Virtual 2024, December 5–8, 2024, Virtual Event, NC, USA

ACM ISBN 979-8-4007-0598-4/24/12
https://doi.org/10.1145/3649165.3703624

1 Introduction

Few if any instructors relish grading. Most would rather help students learn. Nevertheless, grades serve both to communicate to outside audiences the student's mastery of the material [8] and can serve as a motivation for students to attain that mastery.

It is likely that, as long as grades have existed, so has cheating. Student cheating has been long documented [4, 12, 15, 19] and the capabilities of large language models [5, 9] facilitate cheating by making it quick, cheap, and something that can be done without interacting with another human.

Because cheating isn't new, instructors have developed means of assessing student ability in a trusted manner. This trust is achieved by putting the student in a situation where the student's product is likely generated by the student themselves without communication with others or access to disallowed materials. Two common mechanisms include proctored written exams and oral exams. Historically, proctored written exams have been conducted on paper, but in computer science have also included "lab exams" [1, 2] and computer-based testing facilities [6, 20, 21] where students take digital exams on institutional computers. Oral exams are really a special case of proctored exams where the exam is adaptive, presented and answered verbally, and graded as it is taken.

Trusted assessments, however, can be expensive for both course and student and have many practical constraints. Running a proctored exam typically requires significantly more effort than a unproctored one. Oral exams typically magnify this burden because they are done with one or a few students at a time. For the student, taking a trusted written exam presents a burden due having to be physically present.[1] These costs mean that typically only a portion of the course's material is assessed in a trustable fashion.

Furthermore, proctoring presents constraints on the kind of work that can be assessed in a trusted manner. Larger works, like programming projects, design documents, and research papers, require different skills than their smaller-scale counterparts, but we can't sequester students away for days while they complete them. We assign these less trustable assessments because we believe in their value in our students' development. For these two reasons (cost and larger works), many courses include a collection of less trustworthy assessments in their grade computations.

[1] We don't consider proctored online written exams to be trusted assessments due to their significant potential for cheating [11, 16–18] We suspect even online oral exams will be effectively spoofed by AI technology within a few years.

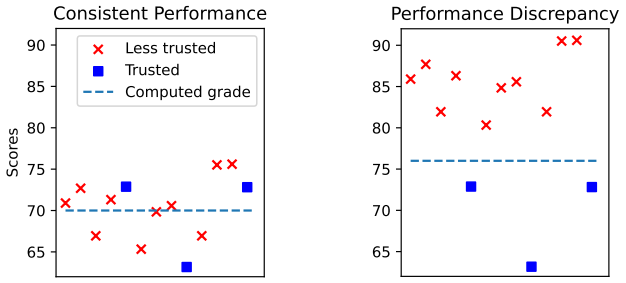

Figure 1: Arithmetic grade computations when scores are coherent (a) vs. incoherent (b). In the latter case, the student's higher performance on less trustworthy assessments buoy their grade by 6 points above their actual ability (70).

In this position paper, we demonstrate that traditional arithmetic grading schemes are susceptible to students cheating on these less trustworthy assessments and, then, show that using Bayesian inference-based methods have the potential to be less susceptible. This case is made through a series of illustrative examples using simulated data.

2 Illustration of the Problem

Grade computations are typically done through weighted averages. This solution works when the trusted and less trustworthy grades are in alignment (e.g., Figure 1a). While there is some variation in the assessment scores, the weighted average implicitly estimates that the student's overall mastery is in the middle of their scores.

The problem with such grade computations is that they are susceptible to cheating. Figure 1b shows a scenario where the student's less trusted assessment scores are inconsistent with their trusted ones. In this scenario, the less trusted scores end up buoying their grade significantly above their trusted assessment scores. If an instructor expects scores from these two classes of assessments to be aligned, then this is problematic, leading instructors to concede these higher grades or requiring them to conduct a cheating investigation.

Method: In this paper, all of the data is synthetic, generated using simulation, so that we can compare a known latent variable (student ability) to that computed by grading techniques. The models are motivated by our previous work that has compared trusted and untrusted assessments [3, 13].

For this short paper, we use a simple model with the following assumptions. First, we assume that students have a constant latent ability with the material and their (non-cheating) assessment scores are drawn from a normal distribution with mean equal to their ability and a standard deviation of 3.5 percentage points [14]. Second, when a student cheats, their score is shifted up 15 points, i.e., mean is 15 points higher, but same standard deviation. Third, students can't cheat on trusted assessments and either cheat on all less trusted assessments or none. We assume a traditional grade computation that gives 60% weight to trusted assessments and 40% to untrusted, with assessments within each class being given equal weight. We consider a hypothetical student with an ability of 70.

Figure 2 shows the grade distributions from such a student that chooses to cheat and not to cheat, resulting from 2,000 simulations of each. Assuming this equal likelihood cheating and not cheating,

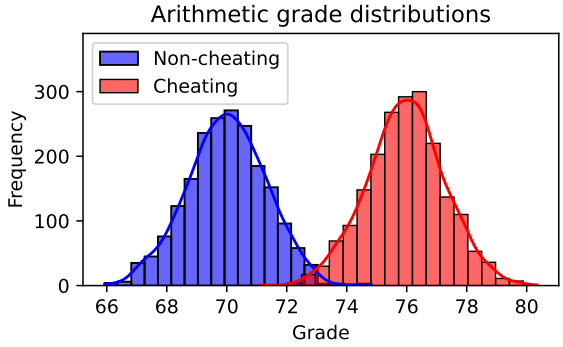

Figure 2: Bimodal distribution of arithmetic grade computations from 4,000 samples of a simulation of a student with ability $\mu = 70$, with half benefiting from a $\chi = 15$ point bonus on less trusted assessments from cheating.

this arithmetic grading has a mean absolute error (MAE) of 3.5 and a root mean square error (RMSE) of 4.41, relative to the latent ability of 70. Without the cheaters, these values are 1.04 and 1.31, respectively.

3 Bayesian Grading

The key idea of this paper is that we should explore using Bayesian inference as part of grade computation. Bayesian inference is a form of statistical inference that uses Bayes' theorem to update a probability distribution as additional evidence is collected. Starting from a prior distribution, Bayesian inference computes a posterior distribution based on observed data and a statistical model for the observed data.

In this paper, we use the same statistical model that was used to generate the data[2]: a normal distribution with a known standard deviation ($\sigma = 3.5$) and known benefit from cheating ($\chi = 15$) but with an unknown mean (the student's ability, μ) and unknown cheating activity ($c \in \{0, 1\}$ where $(c = 0), (c = 1)$ indicate not cheating and cheating, respectively). For trusted assessments, we set χ to zero to model an inability to cheat on those assessments.

With this model, we can compute the probability of observing a given series of assessment scores a_i for a given μ, c pair, using the following equation (where C is a normalization constant):

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

Using Bayes' theorem, we can use the probability of the observations given the parameters to compute the probability of the parameters given the observations:

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

We operationalize Bayes' theorem for this paper by computing probabilities for integer student abilities from 0 to 100 and both cheating activities. Assuming a uniform distribution of student performances as a prior, we transform Equation (2) into

$$P(\mu, c \mid \text{obs}) = \frac{P(\text{obs} \mid \mu, c)}{\sum_{\mu', c'} P(\text{obs} \mid \mu', c')} \tag{3}$$

[2]This is obviously a best-case scenario and is done for simplicity of exposition.

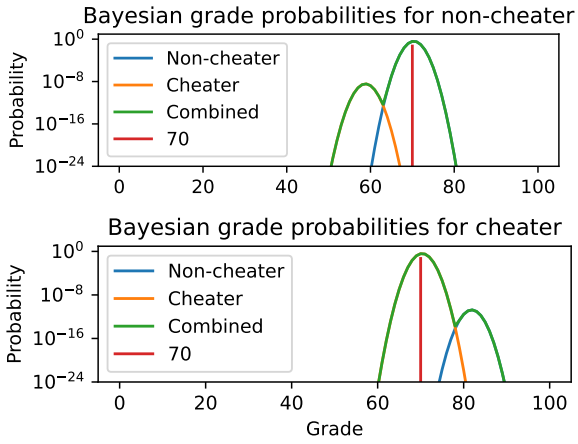

Figure 3: Appropriate grade probability distributions based on latent mastery of material inferred for student exam scores in Figures 1a and b, respectively. Whether or not the student cheats on the less trusted assessments, the method correctly infers the student's latent ability to have the highest probability.

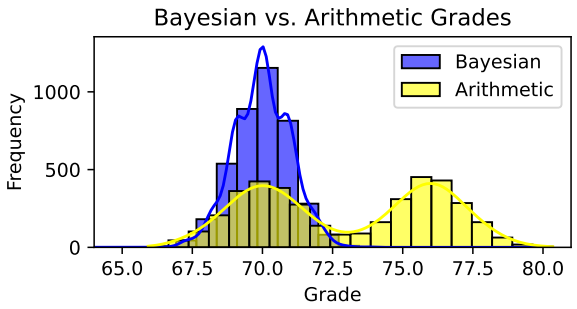

Figure 4: Unimodal Bayesian grading distribution for the same students in Figure 2.

where our probabilities are normalized by the sum of previously computed probabilities $P(\text{obs} \mid \mu, c)$.

Figure 3 shows the inferred probability distributions of how much course material was mastered by the two students in Figure 1. In these log-linear plots, the probability distribution is dominated by the appropriate [non-cheater, cheater] distribution in each case. By taking the argmax of the distribution, we can assign a score. Figure 4 shows that this approach results in a uni-modal distribution around the students latent ability, independent of whether or not they cheated. This results in an MAE of 0.8 and an RMSE of 1.04.

4 Varying the number of trusted assessments

Using Bayesian inference in grading naturally provides guidance on how much trusted assessment is necessary to confidently assign grades. This enables reducing the student burden of trusted assessment when it isn't necessary. Our thinking here is motivated by two scenarios. First, students taking online classes may be required to take in-person proctored exams or online oral exams as part of generating trustworthy grades [7]. Second, future "generative AI"-based tutors may continuously assess student mastery through process as well as product [10], but even this wealth of student data

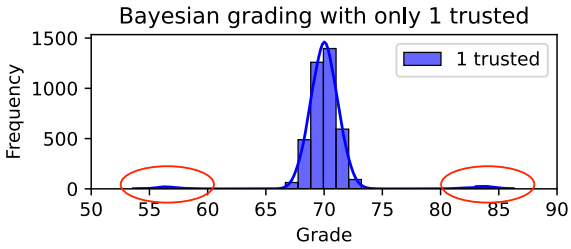

Figure 5: Bayesian grading distribution with only 1 trusted assessment. The bulk of simulated students are still categorized around with a similar distribution around the latent ability ($\mu = 70$), but two smaller modes appear at $\mu = 57$ and $\mu = 83$ as highlighted.

would need to be validated to ensure that it was conducted by the student in question and without other genAI tools.

The key idea is that only enough trusted information needs to be collected to validate that the untrusted information can be correctly interpreted. Figure 5 shows Bayesian grading for the same simulations as above but with only one trusted assessment instead of three, resulting in an MAE of 1.18 and an RMSE of 2.43. While Bayesian grading with one trusted assessment is still better than the arithmetic grading with three, it is quite a bit worse than than Bayesian with three. This is due to mischaracterized students highlighted with ovals in the figure: non-cheating students whose trusted assessment is much lower (by chance) than their less trusted assessments (making them appear to be cheating with a much lower ability) and cheating students who luckily got a high score on their trusted assessment (making them appear to be not cheating with a much higher ability).

When these outliers occur, they predominantly involve a misalignment between the trusted and untrusted scores, which shows up via a non-trivial amount of probability mass (e.g., $> 10\%$) in the cheating ($c = 1$) part of the distribution. If we force students to take additional trusted assessments until their $P(c = 1 \mid \text{obs}) < 10\%$ or until they've taken three trusted assessments, then we can improve the accuracy ($MAE = 0.86$, $RMSE = 1.29$) with few additional trusted assessments. Less than 7% of non-cheaters need to take more than one trusted assessment, and only 0.35% of non-cheaters need to take all three, representing a large reduction of trusted assessment.

5 Conclusions

Our intention in this paper is to highlight the potential of using Bayesian inference in grading and motivate future research. Because the results shown in this position paper use synthetic data, they should be considered only as illustrative examples. We believe that modeling real cheating behavior will require somewhat more sophisticated models than are included in this paper. Future research identifying these models, estimating ability and cheating predilection priors, making this grading transparent to students, and developing models resilient to adversarial attack is encouraged.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2121424.

References

[1] Mary Elaine Califf and Mary Goodwin. 2002. Testing Skills and Knowledge: Introducing a Laboratory Exam in CS1. In *Proceedings of the 33rd SIGCSE Technical Symposium on Computer Science Education* (Cincinnati, Kentucky) *(SIGCSE '02)*. ACM, New York, NY, USA, 217–221. https://doi.org/10.1145/563340.563425

[2] Jacabo Carrasquel, Dennis R. Goldenson, and Philip L. Miller. 1985. Competency testing in introductory computer science: the mastery examination at Carnegie-Mellon University. In *SIGCSE '85*.

[3] Binglin Chen, Sushmita Azad, Max Fowler, Matthew West, and Craig Zilles. 2020. Learning to Cheat: Quantifying Changes in Score Advantage of Unproctored Assessments Over Time. In *Proceedings of the Seventh ACM Conference on Learning @ Scale* (Virtual Event, USA) *(L@S '20)*. Association for Computing Machinery, New York, NY, USA, 197–206. https://doi.org/10.1145/3386527.3405925

[4] Binglin Chen, Colleen M Lewis, Matthew West, and Craig Zilles. 2024. Plagiarism in the Age of Generative AI: Cheating Method Change and Learning Loss in an Intro to CS Course. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 75–85.

[5] Paul Denny, James Prather, Brett A Becker, James Finnie-Ansley, Arto Hellas, Juho Leinonen, Andrew Luxton-Reilly, Brent N Reeves, Eddie Antonio Santos, and Sami Sarsa. 2024. Computing education in the era of generative AI. *Commun. ACM* 67, 2 (2024), 56–67.

[6] Kelly Downey, Kris Miller, Mariana Silva, and Craig Zilles. 2024. One Solution to Addressing Assessment Logistical Problems: An Experience Setting Up and Operating an In-person Testing Center. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 317–323. https://doi.org/10.1145/3626252.3630902

[7] Bobbie Lynn Eicher. 2016. 3 Ways Online Students Might Take Exams. U.S. News and World Report. https://www.usnews.com/education/online-learning-lessons/articles/2016-05-20/3-ways-online-students-might-take-exams

[8] Joe Feldman. [n.d.]. Grading for Equity. https://gradingforequity.org/

[9] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2024. Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems* 36 (2024).

[10] S. Khan. 2024. *Brave New Words: How AI Will Revolutionize Education (and Why That's a Good Thing)*. Penguin Books Limited. https://books.google.com/books?id=KgPREAAAQBAJ

[11] Mark M Lanier. 2006. Academic integrity and distance learning. *Journal of criminal justice education* 17, 2 (2006), 244–261.

[12] Donald L McCabe, Linda Klebe Treviño, and Kenneth D Butterfield. 2001. Cheating in academic institutions: A decade of research. *Ethics & Behavior* 11, 3 (2001), 219–232.

[13] Jonathan Pierce and Craig Zilles. 2017. Investigating Student Plagiarism Patterns and Correlations to Grades. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (Seattle, Washington, USA) *(SIGCSE '17)*. Association for Computing Machinery, New York, NY, USA, 471–476. https://doi.org/10.1145/3017680.3017797

[14] Michael Scott, Tim Stelzer, and Gary Gladding. 2006. Evaluating multiple-choice exams in large introductory physics courses. *Physical Review Special Topics-Physics Education Research* 2, 2 (2006), 020102.

[15] Judy Sheard, Martin Dick, Selby Markham, Ian Macdonald, and Meaghan Walsh. 2002. Cheating and plagiarism: perceptions and practices of first year IT students. In *Proceedings of the 7th Annual Conference on Innovation and Technology in Computer Science Education* (Aarhus, Denmark) *(ITiCSE '02)*. Association for Computing Machinery, New York, NY, USA, 183–187. https://doi.org/10.1145/544414.544468

[16] Jason M Stephens, Michael F Young, and Thomas Calabrese. 2007. Does moral judgment go offline when students are online? A comparative analysis of undergraduates' beliefs and behaviors related to conventional and digital cheating. *Ethics & Behavior* 17, 3 (2007), 233–254.

[17] Chelley Vician, Debra D Charlesworth, and Paul Charlesworth. 2006. Students' perspectives of the influence of web-enhanced coursework on incidences of cheating. *Journal of Chemical Education* 83, 9 (2006), 1368. https://doi.org/10.1021/ed083p1368

[18] George R Watson and James Sottile. 2010. Cheating in the digital age: Do students cheat more in online courses? *Online Journal of Distance Learning Administration* 13, 1 (2010).

[19] Bernard E Whitley. 1998. Factors associated with cheating among college students: A review. *Research in Higher Education* 39, 3 (1998), 235–274.

[20] Craig Zilles, Matthew West, David Mussulman, and Timothy Bretl. 2018. Making testing less trying: Lessons learned from operating a Computer-Based Testing Facility. In *2018 IEEE Frontiers in Education (FIE) Conference*. San Jose, California.

[21] Craig B Zilles, Matthew West, Geoffrey L Herman, and Timothy Bretl. 2019. Every University Should Have a Computer-Based Testing Facility.. In *CSEDU (1)*. 414–420.