Better Peer Grading through Bayesian Inference

Hedayat Zarkoob[1], Greg d'Eon[1], Lena Podina[1,2], Kevin Leyton-Brown[1]

[1] Department of Computer Science, University of British Columbia
[2] Cheriton School of Computer Science, University of Waterloo
{hzarkoob, gregdeon, kevinlb}@cs.ubc.ca, lpodina@uwaterloo.ca

Abstract

Peer grading systems aggregate noisy reports from multiple students to approximate a "true" grade as closely as possible. Most current systems either take the mean or median of reported grades; others aim to estimate students' grading accuracy under a probabilistic model. This paper extends the state of the art in the latter approach in three key ways: (1) recognizing that students can behave strategically (e.g., reporting grades close to the class average without doing the work); (2) appropriately handling censored data that arises from discrete-valued grading rubrics; and (3) using mixed integer programming to improve the interpretability of the grades assigned to students. We show how to make Bayesian inference practical in this model and evaluate our approach on both synthetic and real-world data obtained by using our implemented system in four large classes. These extensive experiments show that grade aggregation using our model accurately estimates true grades, students' likelihood of submitting uninformative grades, and the variation in their inherent grading error; we also characterize our models' robustness.

1 Introduction

Peer grading is a powerful pedagogical tool. It benefits students by helping them to internalize evaluation criteria by applying them critically to peer work (Lu and Law 2012); offering them feedback from equal-status learners (Topping 2009); and giving them exposure to others' perspectives. Just as importantly, it gives instructors a way to make classes more scalable by shifting (some) grading workload away from course staff; this again benefits students by giving them more opportunities for their work to be evaluated within a course's staffing constraints.

In order for peer grading systems to be both useful to instructors and acceptable to students, they must produce grades that are sufficiently similar to those that an instructor would have given. This is a challenging task because individual peer graders will be biased (consistently give generous or harsh grades); noisy (the same grader could grade an assignment differently on different days); and potentially strategic (some students will enter insincere peer grades unrelated to a submission's quality if they can get away with it). Addressing these interrelated challenges has been a topic of academic study in Computer Science for at least the last two decades.

The first methods for aggregating peer grades—and many others introduced more recently—produce *point estimates* of each assignment's grade and each grader's quality (Walsh 2014; Chakraborty, Jindal, and Nath 2018; Prajapati et al. 2020; de Alfaro and Shavlovsky 2014; Hamer, Ma, and Kwong 2005). At their best, methods that produce point estimates maximize the likelihood of the data given a model, e.g., by assigning each grader a "reliability" parameter and iteratively updating these parameters to best describe the reported grades. (At worst, they do not even maximize likelihood. In this case they can produce grades and reliabilities that are inconsistent with each other, such as giving high weights to graders who are judged unreliable.) Even when they do maximize likelihood, such point estimates can be overly confident. This can matter for model accuracy: e.g., the data might show that only one of several students is reliable without offering evidence about which is which, making it likely that the model will commit to the wrong explanation. It can also limit the way such models are used in real classes: e.g., an instructor may not want to trust student grades until the system is *confident* that a peer grader is reliable.

These problems can be addressed by inferring distributions over grade and reliability estimates rather than point estimates. A seminal paper due to Piech et al. (2013) introduced the first such system, using graphical models to simultaneously determine distributions over both grades for student submission and accuracy assessments for each grader. Their core "PG1" model assumes that each assignment has a latent true grade and each peer grader has a latent bias and reliability; these parameters can be estimated from grading data through Bayesian inference, producing posterior distributions (and, hence, confidence intervals) on each assignment's grade and each student's grading abilities. Piech et al. also introduced PG2 and PG3 models that respectively permit graders' biases to change over time and students' grading reliability to be correlated with their own assignments' grades. Follow-up work by others further extended these models to allow for more complex reliability-grade correlations (the PG4 and PG5 models of Mi and Yeung 2015) and to more explicitly account for differences between a single grader's reported grades (the PG6 and PG7 models of Wang et al. 2019).

A key issue in peer grading is that while students are asked to expend effort in grading each other's work, it is difficult to assess whether they did expend this effort. For example, students can subvert systems that assess grading quality by comparing individual grades to each other if they coordinate on all reporting the same grade. In response to this issue, an extensive line of work in the mechanism design literature focuses on incentivizing high quality reporting in peer grading and other crowdsourcing environments (Prelec 2004; Jurca and Faltings 2009, 2005; Faltings, Li, and Jurca 2012; Witkowski and Parkes 2012; Witkowski et al. 2013; Radanovic and Faltings 2013, 2014; Riley 2014; Kamble et al. 2015; Kong, Ligett, and Schoenebeck 2016; Shnayder et al. 2016; Liu and Chen 2018; Goel and Faltings 2019; Gao, Wright, and Leyton-Brown 2019; Zarkoob, Fu, and Leyton-Brown 2020). Work in this area is mostly centered around the idea of *peer prediction*, developing mechanisms that incentivize graders to grade carefully by rewarding them based on comparing their peer grades to others'. While our model of low-effort grading is inspired by work in this area, these mechanisms typically rely on restrictive modeling assumptions, making them inapplicable to most practical peer grading systems. Further, Burrell and Schoenebeck (2021) found that rewards from out-of-the-box peer prediction mechanisms do not accurately reflect grader effort levels on realistic simulated data.

Despite the considerable intellectual progress just described, there remain obstacles to deploying AI-based peer grading systems in practice. Statistically rich methods based on graphical models and economically informed mechanism design approaches have been developed independently; we are not aware of any system that unifies the two by providing both Bayesian parameter estimates and meaningful incentives for students to invest effort in peer grading. Furthermore, the statistical literature allows both students and the peer grading system itself to assign real-valued grades, whereas real instructors tend to use coarse-valued rubrics, particularly when eliciting grades from students. This can harm inference and also requires the instructor to find a way of mapping real-valued grades back onto their course's rubric. Even if this mapping produces accurate grades, students must be able to understand it in order to trust the system (Kizilcec 2016).

This paper addresses all of these problems, introducing extensions to probabilistic peer grading systems that can detect (and hence enable the disincentivization of) low-effort, strategic behavior by students; improve inference quality in the presence of discrete grading rubrics; and output interpretable, discrete final grades that closely approximate maximum a posteriori estimates. In what follows we begin by introducing notation and formally defining the baseline PG1 model (Section 2). We then introduce our novel methods for modeling grader effort, modeling discrete grade reports as censored observations, and outputting interpretable discrete grades via mixed-integer programming (Section 3). We evaluate our contributions in two ways. First, using real data from four offerings of a large class, we show that each of our effort and censoring model extensions improved likelihood on held-out data and that our method for generating interpretable grades closely tracked MAP estimates (Section 4). Second, using simulated data generated using the hyperparameters fit in the previous section, we assessed our model's ability to recover grades and grader reliabilities as a function of dataset size and in the presence of hyperparameter misspecification (Section 5). We conclude by discussing ways in which our methods can be leveraged in the classroom (Section 6).[1]

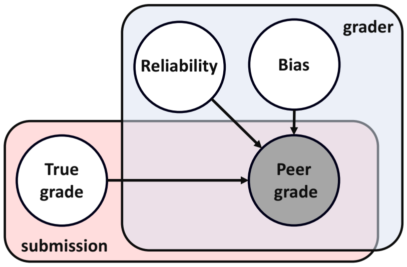

Figure 1: The PG1 graphical model.

2 Technical Setup

We will use the following notation throughout the paper. Let $\mathcal{U}$ be the set of submissions and $\mathcal{V}$ be the set of graders. It is common for instructors to provide graders with a *rubric*: a decomposition of the overall grade into a set of separate components. To capture this, we assume that each submission is graded on C components $\mathcal{C} = \{1, 2, \dots, C\}$, with possible grades of $\{0, 1, \dots, M\}$ for each. We write $\mathcal{N}(\mu, \sigma^2)$ to denote the normal distribution with mean μ and variance σ^2.

We now define PG1 (Piech et al. 2013), a key graphical model from the literature, which models students as having reliabilities and biases. The PG1 model supposes three sets of latent variables. Each submission $u \in \mathcal{U}$ has a *true grade* $s_{u,c} \in \mathbb{R}$ for each rubric component $c \in \mathcal{C}$. Each grader $v \in \mathcal{V}$ is described by a *reliability* $\tau^v \in \mathbb{R}^+$, which captures the consistency of their grading, and a *bias* $b^v \in \mathbb{R}$, which describes their tendency to give generous or harsh grades. An ideal grader would have a high reliability and 0 bias. Then, when a grader v grades a submission u, they give a peer grade $g^v_{u,c} \in \mathbb{R}$ for each component c. Concretely, the PG1 data generating process, depicted in Figure 1, is:

$$\begin{aligned}
&\text{(True grades)} && s_{u,c} \sim \mathcal{N}(\mu_s, 1/\tau_s);\\
&\text{(Reliabilities)} && \tau^v \ \sim \mathrm{Gamma}(\alpha_\tau, \beta_\tau);\\
&\text{(Biases)} && b^v \ \sim \mathcal{N}(0, 1/\tau_b);\\
&\text{(Peer grades)} && g^v_{u,c} \sim \mathcal{N}(s_{u,c} + b^v, 1/\tau^v).
\end{aligned}$$

This model has five hyperparameters: μ_s and τ_s fix the prior distribution of true grades; α_τ and β_τ fix the prior over reliabilities (gamma-distributed to ensure that reliabilities are positive); and τ_b is the precision of the bias distribution.

Armed with a dataset of peer grades, the goal of the model is to infer true grades for each submission and reliabilities and biases for each grader. Such a complex model does not give rise to a closed-form expression for the posterior distribution over these parameters. Instead, the posterior must be estimated numerically. A good option is *Gibbs sampling* (Geman and Geman 1984): initializing each variable randomly,

[1] Open-source implementations of our models are available at https://github.com/hezar1000/mta-inference-public.

repeatedly sampling new beliefs about a single variable in the model conditional on beliefs about all other variables, and reporting the long-run distributions of these samples. This approach is particularly attractive for PG1, as the true grade, bias, and reliability priors are conjugate priors for the normally-distributed peer grade likelihoods, giving each of the Gibbs updates a simple closed form. We present these update equations in Appendix A.[2]

3 Methods

We now present our main conceptual contributions. First, we show how low-effort grading behavior can be disincentivized by augmenting the probabilistic model to include latent variables describing graders' *effort*. Second, we show how to better handle the common case where graders select discrete grades from a coarse rubric by modeling these reports as *censored* observations of an underlying real value. Third, we introduce a mixed-integer programming method for identifying *interpretable* weighted averages of the peer grades that are faithful to the model's posterior beliefs. We present and evaluate these features as extensions to PG1, which we found most applicable to our own class, but they could be applied to any of the PG* models in the literature. In Appendix B, we show how these extensions could be added to the PG5 model of Mi and Yeung (2015), additionally modeling correlations between students' submission grades and reliabilities.

3.1 Modeling Grader Effort

In order for a statistical model to be able to accurately recover true grades from peer reports, students must invest the effort required to grade as well as they can. Typically, students are incentivized to do so in part by receiving explicit grades for their peer grading prowess. However, it is not easy to determine whether a student has done a good job of peer grading when there are no instructor or TA grades to which their evaluation can be compared. The main alternative method of providing an incentive—called peer prediction—is based on comparing students to each other. When all other students grade as accurately as they can, it is often possible to design reward systems that incentivize a given student to do the same (i.e., making effortful reporting an equilibrium). Unfortunately, however, other equilibria also exist in which students coordinate on the same grade without reading the assignment (Jurca and Faltings 2009; Waggoner and Chen 2014; Gao et al. 2014).

We can reduce students' incentives for such low-effort behavior by explicitly modeling it, helping us to avoid assigning high reliabilities to low-effort students. Each time a grader v grades a submission u, we assume they make a binary decision about whether to make an *effort* z_u^v on the submission. If they choose to make an effort, they produce a noisy grade as usual; otherwise, they choose a random grade from a fixed "low-effort" distribution D_ℓ. For simplicity, we model these effort decisions as being independent of the content of the submission. Then, each student has an effort probability e^v, describing the fraction of submissions where they exert high effort. Formally, adding this feature to PG1 produces the model:

$$\begin{aligned}
&\text{(True grades)} && s_{u,c} \sim \mathcal{N}(\mu_s, 1/\tau_s);\\
&\text{(Reliabilities)} && \tau^v \quad \sim \mathrm{Gamma}(\alpha_\tau, \beta_\tau);\\
&\text{(Biases)} && b^v \quad \sim \mathcal{N}(0, 1/\tau_b);\\
&\text{(Effort prob.)} && e^v \quad \sim \mathrm{Beta}(\alpha_e, \beta_e);\\
&\text{(Efforts)} && z_u^v \quad \sim \mathrm{Ber}(e^v);\\
&\text{(Peer grades)} && g_{u,c}^v \sim \begin{cases} \mathcal{N}(s_{u,c} + b^v, 1/\tau^v), & z_u^v = 1;\\ D_\ell, & z_u^v = 0.\end{cases}
\end{aligned}$$

Regardless of our choice of D_ℓ, Gibbs sampling remains straightforward: both efforts and effort probabilities yield closed-form updates, and all other parameter updates simply exclude grades that a given sample calls low effort.

So, which low-effort distribution D_ℓ should we choose? Hartline et al. (2020) showed that the most robust "low effort" strategy is to report the class average, because it minimizes the expected distance to an effortful report. A point-mass low-effort model would be extremely brittle, so a natural D_ℓ is thus a normal distribution centered on the class average. Some low-effort students may adopt other, idiosyncratic strategies such as assigning everything a low or high grade. We would prefer to model such outliers as low-effort behavior rather than having them drive our reliability estimates, so we chose our final D_ℓ to be a mixture between this normal distribution and a uniform distribution:

$$D_\ell = \begin{cases} \mathcal{N}(\mu_s, 1/\tau_\ell), & \text{with probability } 1-\epsilon;\\ \mathrm{Uniform}(0, M), & \text{with probability } \epsilon.\end{cases}$$

Adding effort to a model introduces four new, tunable hyperparameters: α_e and β_e parameterize a prior over grades' effort probabilities, which is beta-distributed to ensure that these probabilities are between 0 and 1; τ_ℓ describes the amount of noise in graders' reports when they put in low effort; and ϵ specifies the probability with which a low-effort grader reports a grade uniformly at random.

3.2 Discrete Rubrics as Censored Observations

In practice, submissions are usually graded on coarse, discrete rubrics such as five-point scales (and virtually no class allows graders arbitrary decimal precision). PG1 and all of its successors nevertheless assume that both true and reported grades are real valued. They do this for two good reasons. First, continuous distributions like Gaussians are realistic models of true grade distributions, and arguably reported grade distributions are just discretizations of the same continuous distributions. Second, such discretizations typically produce non-conjugate priors, and without such special structure, Gibbs updates can be computationally intractable.

However, a model's posterior beliefs can be skewed by failing to model the fact that only integer values can be reported by graders. Treating discrete grades as real-valued observations can add statistical bias to graders' reliability estimates, both by overestimating (e.g., graders appear to be in perfect agreement when their rounded grades are interpreted as draws from a continuous distribution) or underestimating

[2] Our appendix is available at https://arxiv.org/abs/2209.01242.

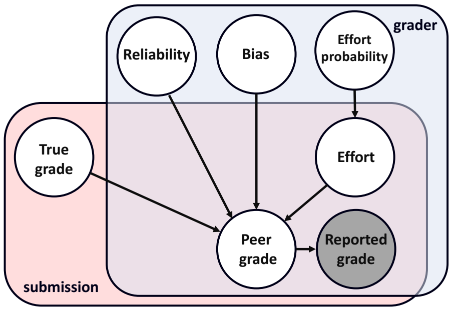

Figure 2: Our complete graphical model.

(e.g., two graders who correctly assess that a true grade is close to the midpoint between two integers can appear to disagree substantially after rounding). Of course, degraded reliability estimation implies degraded true grade estimation.

We propose an approach for extending PG-style models to realistic discrete grade distributions that maintains the tractability of Gibbs updates. Let $G \subset \mathbb{Z}$ be a set of legal discrete grades (e.g., integers between 0 and 5), and let $n_G : \mathbb{R} \to G$ be a function mapping a grade to its nearest value in G, rounding up. We continue to model a grader v grading a submission u as sampling a real-valued peer grade $g^v_{u,c}$, but we now treat these real-valued peer grades as latent variables, with the student reporting the discrete grade $r^v_{u,c} = n_G(g^v_{u,c})$: a *censored observation* of the real-valued peer grade. Figure 2 shows the resulting graphical model. Observe that it correctly leads us to consider grader disagreement to be more likely when an assignment's true grade is 3.51 than when it is 3.0.

Performing naive Gibbs sampling on the latent peer grade variables would lead to extremely sample-inefficient inference: the posterior distribution has multiple modes where the true grade nearly matches one latent peer grade, and Gibbs sampling rarely moves between these modes. To avoid this problem, we instead marginalize over each peer grade, integrating over all of its possible values. While we are no longer able to apply Gibbs updates by evaluating a closed-form expression, we can still straightforwardly compute the likelihood of a reported grade for any setting of the submission's true grade and the grader's reliability and bias, enabling a discrete approximation of the Gibbs updates. To update a true grade variable, we thus consider a uniform grid of possible grades (ranging from 0 to a grade slightly above the maximum grade M), compute an unnormalized posterior probability for each value, renormalize these probabilities to sum to 1, and take a sample from the resulting discrete distribution. The reliability and bias updates are similar, testing uniform grids of plausible reliabilities and biases. We provide full details of these Gibbs updates in Appendix A.

3.3 Explaining Discrete Grades via MIP

A key advantage of a Bayesian approach to reasoning about peer grades is that it yields distributional posterior beliefs about each quantity of interest rather than point estimates. However, students still expect to receive discrete grades rather than probability distributions. Furthermore, in settings where the course staff use the same rubric as the students to grade submissions and where TA grades replace peer grades after an appeal, assigning real-valued final grades to students incentivizes half of them to ask for regrades (e.g., if 3.6 is their true grade, the rounded TA grade would be 4).

How should we turn posterior distributions into discrete grades? One might map the mean of the Gibbs samples to the nearest rubric element, but this can lead to rounding errors. A better option is to choose the rubric element corresponding to the continuous grade interval having the highest mass in the posterior distribution, which is the *maximum a posteriori (MAP)* grade. While this approach is statistically sensible, it leaves students with little insight about how their peer grades influenced the calculation; this can lead to reduced trust in the system (Kizilcec 2016) and more appeals. It can also sometimes produce final grades larger or smaller than any peer grade (e.g., when the model assigns biases having the same sign to all graders). In our experience, students find such grades confusing and unfair; they instead expect to receive final grades that interpolate their received grades, such as averages weighted by each grader's perceived trustworthiness. However, students also tend to fixate on low-quality peer reviews and strongly prefer for such graders to receive zero weights rather than small positive weights.

We propose a novel mixed integer programming (MIP) formulation that maps posterior grade distributions to discrete final grades that can be explained as rounded weighted averages of reported peer grades. Our starting point is to assign a weight for each grader in proportion to their reliability and effort estimates. We then allow the MIP to adjust these weights in two ways to maximize the posterior probability of the resulting rounded weighted average. First, we allow the MIP to deviate from each grader's initial weight by up to a constant S to improve the likelihood of the resulting rounded grade. Second, we prevent the MIP from putting small, nonzero weights on relatively uninformative grades by requiring weights either to be zero or to exceed a minimum threshold T. Notice that the resulting weighted averages can never produce a grade outside of the range of the peer grades. We define our MIP formulation formally in the Appendix C.

4 Validation Experiments on Classroom Data

We now evaluate our contributions on real peer grading data, gathered between September 2018 and December 2021 from four offerings of an undergraduate-level computer science course on ethically evaluating the societal impacts of computing. In each offering of this course, approximately 120 students wrote 11 weekly essays. Each grade consisted of discrete values between 0 and 5 for each of four components (structure; evidence; subject matter; English). Overall, our experiments show that our model extensions (grader effort and censored observations) improved fit and that the explainable grades output by our MIP rarely differed from the (non-explainable) MAP estimates.

Our experiments include data from three types of graders. First, most student essays were graded by 4–5 peers, with a handful receiving more or fewer grades, yielding between 6088 and 7068 peer grades per dataset. Second, each course was supported by a team of 3–5 TAs who spot checked between 474 and 644 essays, mostly in response to suspiciously high average grades, high disagreement between peer grades, and graders with poor historical performance. Our TAs were diligent and responsible, so we clamped their effort estimates to 1 (i.e., their grades could never be explained as coming from the low-effort distribution); we fit their bias and reliability parameters from data just as we did for students. Third, our peer grading datasets included between 60 and 84 gold-standard "calibration" submissions that we used to train students; these grades were painstakingly agreed upon by the whole course staff. We model these gold-standard grades as having being given by a special "instructor" grader, for whom we clamped effort to 1 and further clamped the reliability parameter $\tau^{\text{instr}} = 16$ (corresponding roughly to an 80% accuracy of perfectly recovering true grades); we fit this grader's bias parameter from data.

Of course, our real-world datasets give us no way to reason about ground-truth values for any parameters (like true grades, reliabilities, etc), so we cannot evaluate how accurately our models' posteriors recover these parameters' true values. Instead, we evaluated our models based on their ability to predict held out data: specifically, their *held-out likelihood* (Vehtari, Gelman, and Gabry 2016). Ideally, we would have run leave-one-out cross-validation, but this would have required running the model once for each peer grade in the class, which would have been computationally prohibitive given the size of our datasets and the cost of our inference procedure. Instead, we used 10-fold stratified cross-validation. We first split the dataset into 10 groups of $n/10$ peer grades, ensuring that no two peer grades on the same submission were in the same group. Then, for each way of selecting 9 groups from the 10, we ran the model on these selected observations, summing the model's log likelihoods on the remaining group. An exploratory experiment on one of our four datasets confirmed that this approach closely approximated leave-one-out cross-validation. We use paired t-tests to make statistical comparisons between held-out likelihoods of several models on the same dataset.

Each time we fit our model to a dataset, we collected 4 runs of 1,100 Gibbs samples, discarding the first 100 burn-in samples from each run and concatenating the remainder; this took about 8 CPU hours. We found that this number of samples made a good tradeoff between the runtime and sample complexity of our models: e.g., comparing to runs with 10,000 Gibbs samples, our protocol of gathering 4,000 samples caused an average error of 1%, 3%, and 1% on our estimates of true grades, reliabilities, and effort, respectively, and therefore had a small impact on our results.

First, in order to evaluate the effectiveness of our effort and censoring extensions to PG1, we compared models having both, one, or neither of these features. We independently optimized each model's hyperparameters using randomized search, choosing the hyperparameters that maximized the model's held-out likelihood; full details of this hyperparameter search, along with the resulting hyperparameters, are presented in Appendix D. We found that the model using both features had the highest held-out likelihood.

We then ran an ablation experiment, disabling each feature while holding the hyperparameters fixed. We evaluated held-out likelihood of each ablated model along with the average absolute change in each student's grade. Our results, shown in Figure 3, indicate that removing each feature degraded the model's held-out likelihood, with censoring always causing significant drops in performance and efforts having significant effects in one class. We note that in Fall21 and Winter21, the two offerings of the class where modeling effort made the smallest difference in performance, we made greater efforts to detect low-effort behavior; of course, when such behavior is successfully disincentivized, detecting it will have less impact on model performance. All model changes made small but meaningful changes to the final grades, averaging between 0.05 and 0.15 points on our 5-point scale, and very rarely ever exceeding a single point change.

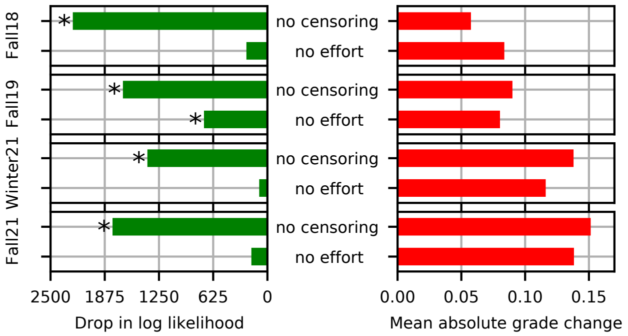

Figure 3: Model ablations. Each bar represents the change relative to the best-performing model, which included both censoring and biases for each dataset. Stars on log-likelihood bars show significance ($p < 0.05$).

Finally, we also investigated variations in model architecture beyond the two extensions introduced in this paper. First, we asked whether we could get good performance without bias terms. We could not; they helped substantially. Second, we asked whether PG5-style correlation between students' grades on their own submissions and their reliabilities was helpful. It was not, regardless of whether we included biases. Details of both experiments are given in Appendix E.

Having settled on our effort + censoring model, we used it to evaluate the extent to which the MIP had to deviate from submissions' MAP grades in order to present them as weighted averages of the graders' reports. We set the MIP constants to the defaults recommended in Appendix C, allowing the graders' weights to change by at most $S = 0.09$, with a minimum non-zero weight of $T = 0.1$. Across the four offerings, replacing the MAP grades with the MIP's output would have caused only 6.5 percent of grades to change.

5 Robustness Experiments on Synthetic Data

While our experiments on real class data allowed us to test how well our models described real peer-grading behavior, they gave us no way to check the accuracy of the parameter estimates. Of course, giving accurate grades (either for submissions or students' grading abilities) is a primary focus of peer grading systems. We therefore conducted further experiments on *synthetic* data, generating parameters and reported grades according to our best-fitting models from the previous section, and evaluating how well the posteriors recovered the latent parameters' true values. This methodology allowed us to test how our estimates improved as the amount of grading data increased, how sensitive they were to the choice of hyperparameters, and whether explaining grades with our MIP increased their error.

The previous section showed that a model incorporating both efforts and censoring had the best performance on all four datasets, but its optimal hyperparameters varied. We show results here based on a representative dataset, with full results for all four sets of optimal hyperparameters in Appendices F and G. Unless otherwise specified, we simulated courses consisting of 10 weeks, with 120 students each making 1 submission and grading 4 peers' submissions each week. The grading rubric had four components, each of which was given an integer grade between 0 to 5. We also included 3 TAs who grade 25% of submissions; we clamped their effort parameters to 1. We simulated TAs as being more reliable than most students: inspired by our real data, we gave TAs a mean reliability of 2.

We evaluated true grade and bias estimates via *Mean Absolute Error (MAE)*. We also computed accuracy (the fraction of true grade MAP estimates equal to the rounded true grade) and RMSE, finding qualitatively similar conclusions with these measures. We found that some models produced inaccurate reliability and effort estimates, but judged that this was less important because rewarding good grading only requires students to be *ranked* in the correct order. Accordingly, we evaluated our reliability and effort estimates with the Spearman rank-order correlation coefficient, measuring how similarly students were ranked by the estimates and true values. In each case, we report the mean and 95% confidence intervals of each metric across inference runs on 15 simulated datasets. We also compare our true grade MAEs to a hypothetical TA with a reliability of 2 (who achieves a mean absolute error of 0.48), allowing us to ask how much data is required to effectively substitute for a TA.

5.1 Parameter Recovery

We begin by testing how the model's parameter estimates were affected by the amount of grading data available. One obvious way to control the amount of data is to change the number of students in the simulated class. However, this change had surprisingly little impact on the inference problem's difficulty, because as the class size varies, each grader continued to grade a total of 40 submissions, and each submission continued to receive 4 grades. Instead, we control these two dimensions separately, independently varying the number of grades from each student and the number of grades given to each submission.

Varying grades per grader. First, we changed the number of grades given by each student by varying the number of weeks in the class. Here, each assignment always received 4 grades, but the number of grades from each student scaled linearly with the number of weeks. The results (Figure 4a-b) show that increasing the size of the dataset in this way improved grader quality estimates. Reliability estimates had an appreciable correlation of 0.6 after just one week of data, improving substantially to 0.9 after 8 weeks. Effort estimates followed a similar trend, but were much more difficult to estimate: one week of data produced a much poorer correlation of 0.3, with later weeks improving to 0.7. Bias estimates, given in the appendix, also improved with additional data.

Perhaps surprisingly, true grade estimates improved very little as the number of weeks grew. This suggests that, with only 4 grades per submission, most of the inaccuracy in the model's true grade estimates was driven by aggregating a small number of noisy signals, rather than because estimates of graders' reliabilities, biases, and efforts were inaccurate.

Varying grades per submission. Next, we changed the number of graders in each course, holding the number of weekly submissions fixed at 40 and each student's grading workload at 4 grades per week. Adding graders in this way increased the number of grades given to each submission but preserved the amount of data about each grader, isolating the effect of additional information on each assignment. The impact of this change on true grade recovery is shown in Figure 4c. These results indicate that adding additional peer graders on each submission substantially reduced true grade MAE, from 0.52 with a single grader to 0.41 with four—well below the MAE of a TA. Adding more graders decreased true grade MAEs far lower, reaching below 0.2 with 32 graders.

Increasing the amount of data in this way had little effect on the model's ability to recover students' reliabilities, biases, and effort probabilities. This suggests that error in those estimates was driven primarily by noise in the reported grades, not by noise in the underlying true grade estimates.

5.2 Robustness to Incorrect Hyperparameters

While the synthetic experiments we have discussed so far show that we were able to recover the model's parameters with sufficient data, they assumed knowledge of the hyperparameters used to generate this data. We now ask whether it is crucial to set these hyperparameters correctly, or whether the model still robustly recovers parameters of interest when given different hyperparameter settings than those used to generate the data. We tested our models under seven changes to the hyperparameters, varying the true grade mean μ_s; the true grade standard deviation σ_s; the bias standard deviation σ_b; the reliability prior mean α_τ/β_τ; the reliability prior variance α_τ/β_τ^2; the effort probability mean α_e/β_e; and the low-effort precision τ_ℓ.

Overall, we found that many of these changes to the hyperparameters had small and statistically insignificant effects on the inference results; these complete results are shown in Appendix G. Notably, we found that the model's performance was quite robust to changes in the reliability and effort probability priors. We show four exceptions in Figure 4: using an incorrect mean or standard deviation for the true grade prior substantially increased true grade MAE from 0.44 to as high as 0.53; incorrectly specifying the low effort distribution τ_ℓ was very detrimental to the effort probability estimates; and using a bias prior with a standard deviation far below its true value hurt bias estimates.

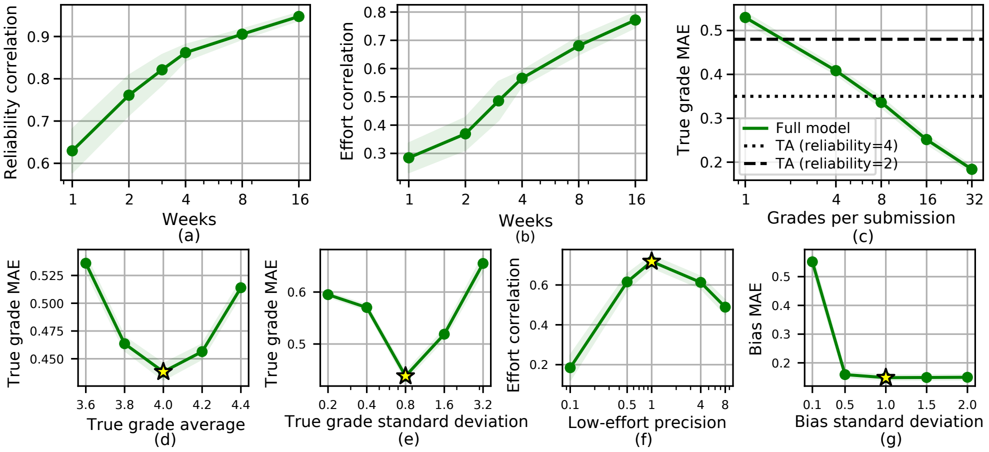

Figure 4: Effects of varying dataset size on (a) reliabilities; (b) effort probabilities; (c) true grades. Robustness of model outputs to misspecified hyperparameters: true grade (d) average and (e) stdev; (f) low-effort precision; (g) bias stdev.

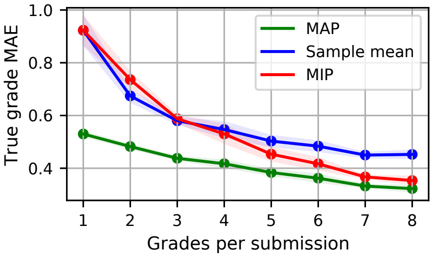

Figure 5: Effect of MIP explanations on true grade MAE.

5.3 MIP Stability

Lastly, we tested the impact of assigning grades based on explanations from our MIP formulation, rather than our model's MAP estimates of true grades, as we varied the number of peer grades given to each submission. The results are shown in Figure 5. With only a single grader, the MIP output was equal to the mean of the Gibbs samples, which is much less accurate than the MAP. However, with additional graders, the MIP gained the flexibility to recover MAP grades, reaching nearly equal MAEs at 7 graders per submission.

6 Conclusions and Practical Considerations

We have shown how probabilistic peer grading systems can be extended to provide incentives for effortful grading; to correctly model discrete peer grades; and to output discrete, interpretable final grades that approximate MAP estimates. We validated our models on four years of real classroom data and investigated both their ability to recover true parameters and their robustness on synthetic data.

Although the peer grading literature has repeatedly shown that Bayesian models can produce accurate grades, tuning them to produce such good performance can be a daunting task for an instructor—our model has 9 hyperparameters! Luckily, our robustness experiments in Section 5.2 showed that the model's posterior beliefs were robust to misspecifying the reliability, bias, and effort priors. Two hyperparameters remain. The first is the true grade distribution, a choice that instructors often make when curving grades. The second is the specification of low-effort behavior, which is important both for boosting model performance and for disincentivizing bad behavior: if the model is good at identifying low effort behavior, students will exhibit this behavior less often. We recommend adapting the specification to capture low-effort behavior observed in spot checks.

Our insistence on providing uncertainty estimates is not just a statistical concern. Our methods work best when they are integrated into the design of the class, giving these uncertainty estimates pedagogical value. For assignment grades, uncertainty estimates can direct TA spot checks towards areas of disagreement. For grader reliability, uncertainty estimates can inform whether students should be trusted to peer grade without TA supervision. They can also help evaluate students' peer grading prowess: in our own class, we initialized the model to be confident that students had poor reliability and required students to do extra grading if the model's pessimistic estimate of their reliability was poor, but scored their peer grading based on the model's *optimistic* reliability estimate. Thus, students got the best grade the model could justify, but students suspected to be weak got additional practice grading, which refined our reliability estimates in turn.

Our parameter recovery experiments in Section 5.1 found that graders' effort probabilities were difficult to estimate: compared to reliabilities, effort probability estimates were much poorer with little data, and converged more slowly as data became available. This is not a surprise: our low-effort graders choose grades that are as difficult as possible to distinguish from effortful graders. The problem is exacerbated by coarse rubrics, which cause many high-effort grades to match the class average exactly. Performance could be improved by tuning the specification of low effort behavior, using an autograding system as another unbiased signal about submissions' grades (Han et al. 2020), or by leveraging other signals of low-effort behavior, such as graders' time spent grading and typing speed (Wang et al. 2019).

Acknowledgments

We thank the reviewers for their helpful comments. This work was funded by an NSERC Discovery Grant, a DND/NSERC Discovery Grant Supplement, a CIFAR Canada AI Research Chair (Alberta Machine Intelligence Institute), a Compute Canada RAC Allocation, awards from Facebook Research and Amazon Research, and DARPA award FA8750-19-2-0222, CFDA #12.910 (Air Force Research Laboratory). Additionally, this work was done in part while two of the authors were visiting the Simons Institute for the Theory of Computing. Research use of our peer grading datasets was authorized by the University of British Columbia's Behavioural Research Ethics Board (BREB #H21-03499).

References

Burrell, N.; and Schoenebeck, G. 2021. Measurement integrity in peer prediction: a peer assessment case study. *arXiv preprint arXiv:2108.05521*.

Chakraborty, A.; Jindal, J.; and Nath, S. 2018. Incentivizing effort and precision in peer grading. *arXiv preprint arXiv:1807.11657*.

de Alfaro, L.; and Shavlovsky, M. 2014. CrowdGrader: a tool for crowdsourcing the evaluation of homework assignments. In *SIGCSE'14*, 415–420.

Faltings, B.; Li, J. J.; and Jurca, R. 2012. Eliciting truthful measurements from a community of sensors. In *IoT '12*, 47–54.

Gao, A.; Mao, A.; Chen, Y.; and Adams, R. P. 2014. Trick or treat: putting peer prediction to the test. In *EC'14*, 507–524.

Gao, A.; Wright, J. R.; and Leyton-Brown, K. 2019. Incentivizing evaluation with peer prediction and limited access to ground truth. *Artificial Intelligence*, 275: 618–638.

Geman, S.; and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741.

Goel, N.; and Faltings, B. 2019. Deep bayesian trust: A dominant and fair incentive mechanism for crowd. In *AAAI'19*, 1996–2003.

Hamer, J.; Ma, K. T. K.; and Kwong, H. H. F. 2005. A method of automatic grade calibration in peer assessment. In *ACE'05*, 67–72.

Han, Y.; Wu, W.; Yan, Y.; and Zhang, L. 2020. Human-machine hybrid peer grading in SPOCs. *IEEE Access*, 8: 220922–220934.

Hartline, J. D.; Li, Y.; Shan, L.; and Wu, Y. 2020. Optimization of scoring rules. *CoRR*, abs/2007.02905.

Jurca, R.; and Faltings, B. 2005. Enforcing truthful strategies in incentive compatible reputation mechanisms. In *WINE'05*, 268–277.

Jurca, R.; and Faltings, B. 2009. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34: 209–253.

Kamble, V.; Marn, D.; Shah, N.; Parekh, A.; and Ramachandran, K. 2015. Truth serums for massively crowdsourced evaluation tasks. *arXiv preprint arXiv:1507.07045*.

Kizilcec, R. F. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *CHI'16*, 2390–2395.

Kong, Y.; Ligett, K.; and Schoenebeck, G. 2016. Putting peer prediction under the micro(economic) scope and making truth-telling focal. In *WINE'16*, 251–264.

Liu, Y.; and Chen, Y. 2018. Surrogate scoring rules and a dominant truth serum. *arXiv preprint arXiv:1802.09158*.

Lu, J.; and Law, N. 2012. Online peer assessment: Effects of cognitive and affective feedback. *Instructional Science*, 40(2): 257–275.

Mi, F.; and Yeung, D.-Y. 2015. Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. In *AAAI'15*, 454–460.

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*.

Prajapati, S.; Gupta, A.; Nigam, S. K.; and Nath, S. 2020. SwaGrader: An honest effort extracting, modular peer-grading tool. In *COMAD'20*, 312–316.

Prelec, D. 2004. A Bayesian truth serum for subjective data. *Science*, 306(5695): 462–466.

Radanovic, G.; and Faltings, B. 2013. A robust Bayesian truth serum for non-binary signals. In *AAAI'13*, 833–839.

Radanovic, G.; and Faltings, B. 2014. Incentives for truthful information elicitation of continuous signals. In *AAAI'14*, 770–776.

Riley, B. 2014. Minimum truth serums with optional predictions. In *ACM'14 Workshop on Social Computing and User Generated Content*.

Shnayder, V.; Agarwal, A.; Frongillo, R.; and Parkes, D. C. 2016. Informed truthfulness in multi-task peer prediction. In *EC'16*, 179–196.

Topping, K. J. 2009. Peer assessment. *Theory into practice*, 48(1): 20–27.

Vehtari, A.; Gelman, A.; and Gabry, J. 2016. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5): 1413–1432.

Waggoner, B.; and Chen, Y. 2014. Output agreement mechanisms and common knowledge. In *HCOMP'14*.

Walsh, T. 2014. The PeerRank method for peer assessment. In *ECAI'14*, 909–914.

Wang, T.; Jing, X.; Li, Q.; Gao, J.; and Tang, J. 2019. Improving peer assessment accuracy by incorporating relative peer grades. *International Educational Data Mining Society*.

Witkowski, J.; Bachrach, Y.; Key, P.; and Parkes, D. C. 2013. Dwelling on the negative: Incentivizing effort in peer prediction. In *HCOMP'13*.

Witkowski, J.; and Parkes, D. C. 2012. A robust Bayesian truth serum for small populations. In *AAAI'12*, 1492–1498.

Zarkoob, H.; Fu, H.; and Leyton-Brown, K. 2020. Report-sensitive spot-checking in peer-grading systems. In *AAMAS'20*, 1593–1601.