Peering Inside Peer Review with Bayesian Models

Ilya M. Goldin and Kevin D. Ashley

Intelligent Systems Program and Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA, USA 15260
{goldin,ashley}@pitt.edu

Abstract. Instructors and students would benefit more from computer-supported peer review, if instructors received information on how well students have understood the conceptual issues underlying the writing assignment. Our aim is to provide instructors with an evaluation of both the students and the criteria that students used to assess each other. Here we develop and evaluate several hierarchical Bayesian models relating instructor scores of student essays to peer scores based on two peer assessment rubrics. We examine model fit and show how pooling across students and different representations of rating criteria affect model fit and how they reveal information about student writing and assessment criteria. Finally, we suggest how our Bayesian models may be used by an instructor or an ITS.

Keywords: computer-supported peer review, evaluation of assessment criteria, Bayesian models.

1 Introduction

Increasingly, instructors are turning to peer review as a teaching aid. [1, 2] Peer review has important benefits beyond shifting some of the burden of assessment from the instructor to students. By giving and receiving feedback from peers, students may improve their own work, and practice a useful professional skill. If instructors state their criteria explicitly, this helps make assessment rigorous and objective. Students can focus on these criteria as they write their essays and evaluate peer work. By spending less time on assessment, instructors can help struggling students. It has been shown that combined evaluations from multiple reviewers estimate essay quality reliably, and that students may respond better to feedback from peers rather than the instructor. [3, 4] Perhaps, most importantly, peer review enables instructors to assign writing exercises they might not otherwise assign for lack of time to prepare in-depth critiques, especially in very large classes.

Instructors and students would benefit even more, if peer reviewers used assessment rubrics that were conducive to all of the potential benefits. Rubrics are the heart of assessment. If reviewers assess aspects of peer work that are wrong for the exercise, then feedback will not be beneficial to authors, and the reviewers will have wasted their time providing it. Generic review criteria such as “flow”, “logic”, and “insight” [6] may be appropriate for some writing exercises, but not all. Peer feedback may be solicited in a structured way on the issues raised in an assignment (i.e., with

G. Biswas et al. (Eds.): AIED 2011, LNAI 6738, pp. 90–97, 2011.

problem-specific support to reviewers) and on more general but still domain-relevant aspects of the writing (domain-relevant support). [8]

A peer review system that prompts reviewers to assess authors' understanding of specific conceptual issues may provide aggregate estimates of students' grasp of these issues in the peer-review exercise thanks to a statistical model. The estimates are based on the feedback that students give to each other when in peer review. Gathering independent perspectives of multiple peer reviewers increases reliability. Modeling students' exchange of feedback may provide an instructor with a more informed view concerning how well students have grasped conceptual issues in a writing exercise. The modeling may also evaluate the criteria that students used to assess each other. Problems with peer assessment criteria may be indicative of curricular issues (e.g., if material is not covered in an optimal sequence), or of student comprehension of the criteria. Ultimately, the modeling aims to make peer-review exercise transparent to the instructor, who can use the information to guide and modify his or her teaching and make appropriate midcourse adjustments to the curriculum.

We compare our models on two types of peer assessment criteria. The only inputs to the model are the instructor's and the peers' scores of student work. We begin by describing computer-supported peer review and some artifacts of the peer review process, and explain how they may be related to assessment. We then develop several hierarchical Bayesian models relating these artifacts, and evaluate the models. Finally, we discuss the lessons learned from this modeling and explain how a Bayesian model may be used by an instructor or an Intelligent Tutoring System. We leave the actual generation and evaluation of reports to instructors for future work.

2 Study, Methods, and Data Sets

We report a new analysis of the datasets described in [8]. All 58 participants were second or third year law students in a course on Intellectual Property law. Students were required to take an open-book, take-home midterm exam and to participate in the subsequent peer-review exercise. The exam comprised one essay-type question, which students had 3 days to answer. Answers were limited to no more than four 1.5-spaced typed pages. The question asked students "to provide advice concerning [a particular party's] rights and liabilities" given presented a fairly complex factual scenario. The instructor designed the facts of the problem to raise issues involving many of the legal claims and concepts (e.g., trade secret law, shop rights to inventions, right of publicity, passing off) that were discussed in the first part of the course. Each claim involves different legal interests and requirements and presents a different framework for viewing the problem. Students were expected to analyze the facts, identify the claims and issues raised, make arguments pro and con resolution of the issue in terms of the concepts, rules, and cases discussed in class, and make recommendations accordingly. Since the instructor was careful to include factual weaknesses as well as strengths for each claim, the problem was ill-defined; strong arguments could be made for and against each party's claims.

With peer review systems such as CPR [5] and SWoRD [6], (1) students in a class write essays on a topic assigned by the instructor, and (2) the system distributes the essays to a group of *N* student peers for review. (3) Using review criteria and forms prepared by the instructor, the peer reviewers assess the student authors' papers along the criteria and submit their feedback via the system. It is important that reviewers provide written justifications of their numeric ratings. [7] The authors (4) may indicate whether or not the feedback was helpful, and (5) revise their drafts. In our study, the participants completed steps (1) through (4) of this peer review process. Students were randomly assigned to one of the two conditions in a manner balanced with respect to their LSAT scores. Each student gave feedback to and received feedback from four others, who had to be in the same condition. We collected ratings according to Likert scales (7 points, grounded at 1,3,5,7). The conditions differed in the rating prompts used by the reviewers, either domain-relevant or problem-specific. The former dealt with legal writing skills (i.e., issue identification, argument development, justifying an overall conclusion, and writing quality). The latter addressed criteria concerning five legal claims or issues raised by the problem's facts. This yielded two datasets: domain-relevant and problem-specific.

3 Overview of Bayesian Data Analysis

Our several different statistical models representing the domain of peer review use an expert's scores of the students' essays as the response variable; the models differ in the explanatory variables they use and in the hierarchical structure. We ask if it is possible to approximate the instructor scores by using the artifacts of peer review. We consider whether the additional complexity required for sophisticated modeling is a worthwhile trade-off for the inferences supported by the models.

We use a statistical modeling technique called Bayesian data analysis. [9] Bayesian models can incorporate prior beliefs about the parameters; for example, aggregate peer ratings may be said to be normally distributed. By combining prior beliefs with data and with formulations of likelihood, a Bayesian model yields posterior estimates for the parameters of interest and describes each estimate in terms of a probability distribution rather than just a point value.

While Bayesian modeling has long been applied in educational research, as far as known, our use of it is a novel contribution to the study of peer assessment in education. From the perspective of statistical analysis, peer review is fairly complex. It involves repeated measures (multiple reviews of every paper), sparse data (any student reviews only a few papers), and hierarchy (authors may receive feedback according to multiple reviewing criteria). By using Bayesian data analysis, we can enter these relationships among the data into our model in a straightforward way, and we can compare different models based on our intuitions about model structure. Furthermore, a single Bayesian computation estimates all the quantities of interest at once, bringing to bear all the available data. This means that the different parameters help estimate each other according to the expression of likelihood we enter.

Given two models that fit the data equally well, one may prefer the simpler one (e.g., complex models can be prone to overfitting) or the more complex one (e.g., it may embody knowledge about domain structure). We compare models in each condition in terms of Deviance Information Criterion (DIC), a metric that rewards well-fitting models, and penalizes models for complexity. Model fit is defined as deviance, similar to generalized linear models. Model complexity involves the effective number of parameters in the model. This is computed at model “run time” as a function of how information is pooled across groups in a multilevel model, rather than at “compile time” from the mathematical model structure. Lower DIC is better. DIC values may be compared on one dataset but not across datasets.

4 Hierarchical Bayesian Models of Peer Review

In each model, we regress the instructor score on peer ratings. Our baseline model 5.1a uses the simplest representation that maps from peer ratings to an instructor's score; it averages all ratings an author receives. It ignores the distinct rating dimensions reviewers used, and treats students as independent. In model 5.1b, we do not treat students as independent, pooling model parameters so that what we know about students as a group helps us understand individual students, and vice versa. In model 5.2b, we represent the ratings dimensions separately rather than together. We also developed models (not described here) that include inbound back-review ratings as a predictor and seek out trustworthy reviewers by comparing reviewer opinions.

We centered the peer ratings about the mean (and centered separately within each rating dimension for model 5.2b). We ran each model separately for the students in the two datasets, because it would not be sensible to compute the contribution of problem-specific information for students in the domain-relevant condition and vice versa. We fit each model 3 times and examined whether the chains converged in their posterior estimates. Each fit was allowed 6000 iterations, with 1000 initial iterations discarded to avoid bias due to randomly determined starting values.

4.1 Model 5.1a: Contribution of Inbound Peer Ratings

Model 5.1a is a regression of the midterm scores as a function of the pupils' inbound peer ratings only. We treat students as randomly drawn from a single population, and we do not distinguish between rating dimensions.

The multiple ratings that each student receives are exchangeable with each other (i.e., not tied to particular reviewers), and constitute repeated measures of each student. We treat midterm and ratings as normally distributed.

The inbound peer ratings are taken as normally distributed and sufficiently described by each author’s ratings mean and variance. In such a model, the means and variances are hyperparameters and estimated simultaneously with other parameters during MCMC sampling. This yields both point estimates and posterior distributions with credible intervals indicating the model's certainty in the parameter estimate.

Formally, the model is as follows. The per-pupil instructor score Y_p is distributed normally, with a mean that is the per-pupil knowledge estimate μ_p and overall variance estimate σ^2.

$$Y_p \sim N(\mu_p, \sigma^2)$$

We fit a per-pupil intercept α_p. We also compute $\mu_p^{[IPR]}$, the mean of inbound peer ratings for pupil p ignoring criteria distinctions, and we give this a weight β.

$$\mu_p = \alpha_p + \beta * \mu_p^{[IPR]}$$

Finally, we say that a pupil's inbound peer ratings are distributed normally according to the pupil's individual mean $\mu_p^{[IPR]}$ and individual ratings variance $\sigma^2_{p[IPR]}$.

$$IPR_p \sim N(\mu_p^{[IPR]}, \sigma^2_{p[IPR]})$$

The prior distribution for $\mu_p^{[IPR]}$ was said to be uninformative, normally distributed with a mean of 0 and a variance of 1000.

This "no pooling" regression model does not share information across pupils. [9] Each pupil is described via individual intercept α_p, between-pupils variance σ^2, individual mean peer rating $\mu_p^{[IPR]}$ and individual ratings variance $\sigma^2_{p[IPR]}$. In other models, below, we consider that information could be pooled across students, and what we learn about one student could help describe a different student.

Model 5.1a is a plausible first attempt to establish if the ratings that peers give each other approximate instructor assessment. It asks if the cumulative opinion of the reviewers (i.e., the mean inbound peer rating) corresponds to an instructor's grade, and if the peer reviewers tend to agree (i.e., measuring the ratings' variance). Additionally, it incorporates normal prior distributions for the response and the ratings. Whether or not this baseline differs from the alternative models, its evaluation should still be helpful in understanding peer review.

4.2 Model 5.1b: Contribution of Information Pooling

In model 5.1b, we use information learned about one student to inform our understanding of other students. This is accomplished in two ways. First, we stipulate that all individual intercepts α_p are not independent, but drawn from a common distribution. Each student's information is then used to estimate this distribution's hyperparameters μ_α and σ^2_α, and the distribution in turn constrains the estimation of the individual students' intercepts.

$$\alpha_p \sim N(\mu_\alpha, \sigma^2_\alpha)$$

Second, in a similar fashion, we constrain the estimation of individual students' inbound peer rating means $\mu_p^{[IPR]}$ via hyperparameters $\mu_{[IPR]}$ and $\sigma^2_{[IPR]}$.

$$\mu_p^{[IPR]} \sim N(\mu_{[IPR]}, \sigma^2_{[IPR]})$$

All hyperparameters were given uninformative prior distributions.

4.3 Model 5.2b: Contribution of Rating Dimensions

Models 5.1a and 5.1b treat all inbound peer ratings as though they correspond to one rating dimension, no matter that they were elicited via different prompting questions. Model 5.2b represents the distinct dimensions of the peer ratings.

To incorporate information on dimensions, we say that each observed inbound peer rating is normally distributed with mean $\mu_{ip}^{[IPR]}$ that is equal to the average of the ratings received by author p for rating dimension i, and with a variance $\sigma^2_{i[IPR]}$ for dimension i that is shared across all pupils. (There are n=4 rating dimensions in the domain-relevant condition, and n=5 in the problem-specific condition.)

$$Y_p \sim N(\mu_p, \sigma^2)$$

$$\mu_p = \alpha + \Sigma_1^n \beta_i * \mu_{ip}^{[IPR]}$$

$$IPR_{ip} \sim N(\mu_{ip}^{[IPR]}, \sigma_{i[IPR]}^2)$$

Within each dimension, we pool individual pupils' means of inbound peer ratings $\mu_{ip}^{[IPR]}$ by stipulating a common distribution across students. These have uninformative prior distributions, normal with a mean of 0 and a variance of 1000.

Model 5.2b estimates a coefficient β_i for each rating dimension i. A partially pooled α_p in this model had problems with convergence, so we substituted a single completely pooled intercept α. Distinguishing the ratings by dimension leads to fewer observed ratings per pupil, per dimension. For example, rather than 20 peer ratings per pupil in the problem-specific condition (5 rating dimensions times 4 reviewers), there are ratings from 4 reviewers per dimension. This precludes estimation of individual per-dimension variances; instead, we estimate per-dimension variance parameters $\sigma_{i[IPR]}^2$ pooled across all students.

5 Results and Discussion

There are two key findings. First, partial pooling (model 5.1b) can improve significantly on the baseline (5.1a) for both domain-relevant and problem-specific datasets. (Table 1) Second, distinguishing the different rating criteria (5.2b) improves the fit for the problem-specific dataset, but actually hurts the fit for domain-relevant.[1]

Table 1. Model fit (DIC) for domain-relevant and problem-specific datasets

Model	domain-relevant DIC	problem-specific DIC
5.1a	1416	2423
5.1b	1305	2237
5.2b	1347	1732

In all three models, the intercepts α_p (5.1a, 5.1b) and α (5.2b) were estimated to be close to the mean of the instructor-assigned midterm scores. With ratings centered, the intercept represents the predicted midterm score for a student whose inbound peer ratings averaged to zero. Pooling made intercept estimates an order of magnitude tighter, e.g., for the problem-specific dataset, to ±0.8 points with 95% confidence on the instructor's scoring scale. Pooling for $\mu_p^{[IPR]}$ also allowed model 5.1b to share information across students, but the effect was less pronounced given that 5.1a already had tight intervals on these parameters. We find that partial pooling can be an effective technique for these models.

[1] We have seen DIC scores vary ±15 points over a dataset given different random starting points. Despite high autocorrelation for some parameters, Rhat values for all parameters in all models were below 1.2, i.e., the chains converged in their estimates. The chains mixed well, suggesting that samplers did not get stuck. Thus, we conclude that the results are stable.

Table 2. β coefficients, domain-relevant (DR) and problem-specific (PS), * marks significance

Model	β	β_1	β_2	β_3	β_4	β_5
5.1a-DR	-0.14					
5.1a-PS	-0.07					
5.1b-DR	0.99*					
5.1b-PS	1.46*					
5.2b-DR		1.48*	-0.50	0.56	-1.01	
5.2b-PS		2.10	0.86*	0.03	0.32*	0.40

The β coefficients could be said to represent the importance of the averaged inbound peer ratings (per-dimension or collapsing dimensions) to estimating the instructor's score. Under 5.1a, the credible intervals for β included zero, implying that peer ratings were not significant predictors of instructor scores for that model. With model 5.1b, estimates of β with 95% confidence did show that average peer ratings predicted instructor scores, emphasizing the value of pooling.

Model 5.2b showed that distinct rating dimensions helped to fit the data using problem-specific criteria but not using domain-relevant criteria. This can be seen from the overall fit (Table 1); additionally the β coefficients show that two of five the problem-specific dimensions helped to estimate the midterm score (a third, β_5, was marginally helpful), versus just one of four domain-relevant ones. Further, the problem-specific β estimates are all positive, suggesting that each dimension adds linearly to the intercept. Some domain-relevant β estimates have negative signs, as if high performance on those dimensions corresponds to a drop in the midterm score, which is counterintuitive. These problems for domain-relevant criteria echo the high pairwise correlation between $\mu_{ip}^{[IPR]}$ for all $\binom{4}{2} = 6$ criteria pairs; problem-specific support had correlation for only 2 of 10 pairs, as reported earlier. [8] High collinearity may cause instability and interactions among β_i coefficients for the domain-relevant rating criteria (without hurting overall model fit). The β_i for the problem-specific ratings may be intuitively interpreted as indicating that criteria differ in their impact on approximating instructor scores.

6 Conclusion

Parameter estimates from these models are likely to provide an ITS or an instructor with actionable information on individual pupils, the whole class, and the assessment rubric itself. Some may even suggest changes in curriculum or assessment. For example, $\mu_{ip}^{[IPR]}$ estimate a student's proficiency with regard to the criteria. Distributions of $\mu_{ip}^{[IPR]}$ can alert an instructor if the criteria differ in difficulty, or if they are poorly anchored. Pairwise correlations among each pupil's $\mu_{ip}^{[IPR]}$ may suggest which criteria are redundant. Inconsistent signs among β_i may hint that the reviewers' rubric differed from the instructor's grading scheme, while consistent β_i show how the criteria differ in their impact on approximating instructor scores. For instance, in the domain-relevant dataset, the $\mu_{ip}^{[IPR]}$ intercorrelation and the signs of β_i

suggest that the instructor should clarify the criteria and concepts to students and revise the criteria for future peer review. All these estimates accommodate missing peer ratings because they "borrow strength" from other students' ratings, and because they are posterior distributions with intervals that speak to the estimates' credibility.

In some cases, peer assessment is an important perspective on a student's work in its own right; in others, its relevance may depend on how well it approximates assessment by an instructor or other expert. Either way, consumers of peer assessment information, whether instructors or tutoring systems, require precise estimates of the key parameters in peer assessment. They also need to know whether or not the estimates are credible. The Bayesian models we have described fill that role.

The old software developers' adage "garbage in, garbage out" applies to peer assessment criteria. Criteria are not all equally useful, clear, or functional. The models we have developed and the results they report are only as good as the criteria. The good news is that peer review provides a built-in facility for evaluating the criteria, which can help instructors to refine them and to communicate them to pupils.

References

1. Strijbos, J., Sluijsmans, D.: Unravelling Peer Assessment. Special Issue of Learning and Instruction 20(4) (2010)
2. Goldin, I.M., Brusilovsky, P., Schunn, C., Ashley, K.D., Hsiao, I. (eds.): Workshop on Computer-Supported Peer Review in Education, 10th International Conference on Intelligent Tutoring Systems, Pittsburgh, PA (2010)
3. Falchikov, N., Goldfinch, J.: Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. Rev. of Ed. Research 70, 287–322 (2000)
4. Cho, K., Chung, T.R., King, W.R., Schunn, C.: Peer-based computer-supported knowledge refinement: an empirical investigation. Commun. ACM 51, 83–88 (2008)
5. Russell, A.: Calibrated Peer Review: A writing and critical thinking instructional tool. Invention and Impact: Building Excellence in Undergraduate Science, Technology, Engineering and Mathematics (STEM) Education. American Association for the Advancement of Science (2004)
6. Cho, K., Schunn, C.D.: Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. Computers and Education 48 (2007)
7. Wooley, R., Was, C.A., Schunn, C.D., Dalton, D.W.: The effects of feedback elaboration on the giver of feedback, pp. 2375–2380. Cognitive Science Society, Washington, DC (2008)
8. Goldin, I.M., Ashley, K.D.: Eliciting informative feedback in peer review: Importance of problem-specific scaffolding. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 95–104. Springer, Heidelberg (2010)
9. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge (2006)