Accounting for Peer Reviewer Bias with Bayesian Models

Ilya M. Goldin

Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, USA
goldin@cmu.edu

Abstract. Instructors and researchers of peer review would benefit from a consistent way of characterizing peer review among students. One factor that can affect peer review is reviewer bias. For example, students may give biased assessments if some reviewers are lenient and others stringent. Accordingly, statistical models of peer review should account for reviewer bias. We present work in progress comparing alternative models of bias, including models that postulate that bias may be a multidimensional construct and models that consider whether bias is a useful predictor of instructor assessment. We find that a model that represents bias as multidimensional and as predicting instructor assessment is the best-fitting model on one of two real-world datasets.

Keywords: peer review in education, reviewer bias, Bayesian models

1 Introduction

Among the many ways to configure a collaborative learning exercise [1], one in particular that deserves the attention of instructors and researchers is computer-supported peer review. Peer review is a family of collaborative learning exercises that all involve the same minimal process: students create a first draft of some assignment, exchange drafts, and review other students' drafts. (This process may be extended, e.g., by asking students to create a second draft after they receive peer feedback.)

Instructors and researchers of peer review would benefit from a consistent way of characterizing peer review among students. For instance, an instructor may wish to know to what extent the students' works show mastery of domain content.

Complexity of peer review makes this difficult. One issue is that reviewers vary in terms of a general tendency to lenience or strictness. For instance, a student whose work is assessed by reviewers who happen to be lenient may be misinformed about his own abilities. Further, differences among reviewers may lead an author to doubt the validity of peer review. [2] We call a reviewer "biased" if he assesses peer works systematically higher or lower than other reviewers. Related psychometric research explores constructs such as "response set" (e.g., [3]).

The minimal peer review process may be summarized using statistical models. Related work in statistics has explored sophisticated models of survey respondent bias (e.g., [4]), but has considered neither connections to rubric dimensions in peer review in education, nor connections to instructor assessment. Prior peer review research has developed methods to evaluate reviewers. For instance, reviewers may be evaluated in

terms of bias ("systematic difference"), consistency and spread, and reviewer evaluations may act as differential weights on the scores given by the reviewers to produce a summative peer assessment of an author's work. [5] Reviewer evaluations and the quality of the peer author works can be estimated in a joint estimation process. [6, 7]

Our new family of models examines alternative representations of reviewer bias, including models that postulate that bias may be a multidimensional construct and models that consider whether bias is a useful predictor of instructor assessment. We report how the new models perform on two real-world peer review datasets.

2 Methods

We define the Minimal Peer Review Process as follows:

1. For the purpose of some instructor-assigned exercise, some students are designated as Peer Authors and some as Peer Reviewers.
2. Each of the Peer Authors produces a first draft, and circulates the draft to Peer Reviewers for assessment.
3. Each of the Peer Reviewers assesses the drafts according to a well-defined Assessment Rubric.

Some ways to extend the Minimal Process are to make the sets of Peer Authors and Peer Reviewers the same; to automatically circulate drafts for review with computer support [5]; to increase the number of reviewers for reliability of assessment [2]; to train reviewers in applying the rubric [8]; to select reviewers randomly or based on substantive criteria [9, 10]; and to de-identify the drafts [11] so that student relationships do not influence assessment.

Independently of these extensions, the result of MPRP is that each Peer Author's first draft receives assessments from one or more Peer Reviewers, along the one or more dimensions embedded in the Assessment Rubric.

One way to summarize a peer review exercise is to say that the true quality of each draft is equivalent to an average of all the assessments from all of the reviewers. However, this is an oversimplification. First, averaging is inappropriate for rubric dimensions that are semantically distinct, as they should be. [12] For example, a submission to an academic conference that is evaluated as low in rigor and but high in novelty is not of the same quality as a submission that is mediocre in both. Second, depending on the author, some assessment dimensions may be more challenging than others, e.g., rigor may be harder to achieve than novelty for some authors, and vice versa for others. Each dimension may be said to have a baseline level of difficulty, and the quality of the drafts should be considered with respect to this baseline. Third, there may be different individual biases among reviewers. Reviewer bias is a systematic difference whereby some reviewers tend to give higher ratings than other reviewers. Fourth, the summary could be made more useful by incorporating external information beyond the peer assessments themselves, e.g., the instructor's assessments of the same works. Thus, a more nuanced summary of a peer review exercise would distinguish assessments along different dimensions, account for dimension difficulty and reviewer bias, and it would incorporate external information.

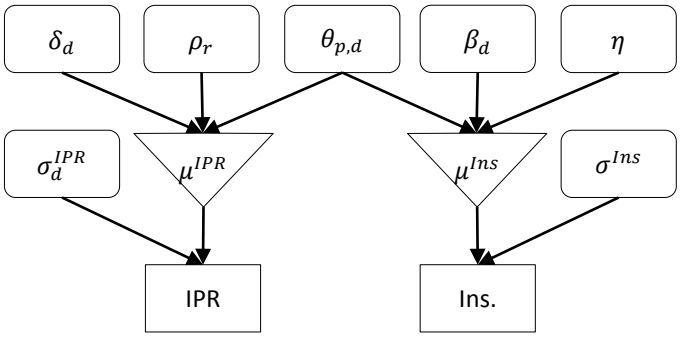

Figure 1: Unidimensional Reviewer Bias as a directed acyclic graph. Latent parameters in rounded-corner boxes, observed data in rectangles; deterministic functions in triangles

2.1 Unidimensional Reviewer Bias

These desiderata for a summary may be incorporated in a statistical model. (Fig. 1) The Unidimensional Reviewer Bias model has two inputs: peer assessments ("IPR", for inbound peer ratings), and external information ("Ins.", for instructor-generated assessments). For each peer rating, we record which author's work is being reviewed, which reviewer gave the rating, and the rubric dimension to which the rating pertains. This observable information allows us to estimate unobservable, latent parameters: the difficulty, δ_d, of the rubric dimension *d*; the reviewer bias, ρ_r, of reviewer *r*; the quality of the work, $\theta_{p,d}$, by author *p* along dimension *d*; and the spread of peer ratings, σ_d^{IPR}, within dimension *d*. Given the instructor assessments of student works, the model also estimates the average and spread of those assessments (σ^{Ins}).

Formally, $IPR \sim N(\delta_d + \rho_r + \theta_{p,d}, \sigma_d^{IPR})$, that is, a peer rating is modeled as drawn from a normal distribution, with the mean determined by a linear combination of the dimension difficulty, reviewer bias and the quality of the peer work along the dimension. The quality of a peer author's work is represented as normal and multivariate in the number of assessment dimensions, $\boldsymbol{\theta_p} \sim N_d(\boldsymbol{\mu^\theta}, \boldsymbol{\Sigma^\theta})$, with partial pooling such that all quality estimates are shrunk towards the mean $\boldsymbol{\mu^\theta}$ of their shared distribution. The dimensions are treated as independent of each other by fixing their covariances at zero in the covariance matrix $\boldsymbol{\Sigma^\theta}$. (The pairwise correlation among the dimensions can still be computed as a separate step after model-fitting.) The reviewer biases are normal, $\rho_r \sim N(\mu^\rho, \sigma^\rho)$, and similarly pooled with a shared mean μ^ρ. The dimension difficulties δ_d are independent of each other, i.e., unpooled. To identify the model, the $\boldsymbol{\theta_p}$ and ρ_r parameters are constrained to sum to zero, while the dimension difficulties are allowed to float. Further, $Ins \sim N(\boldsymbol{\theta_p}\boldsymbol{\beta} + \eta, \sigma^{Ins})$, that is, the instructor assessment of a student's work is modeled as a linear function of an intercept η (the mean of all instructor assessments) and the peer assessments of the quality of the work, differentially weighted for each assessment dimension as per weights vector $\boldsymbol{\beta}$.

This model improves on our previous work [13] in that it differentiates among reviewers by estimating individual reviewer bias. Additionally, it represents the quality of a pupil's work as a multivariate normal parameter rather than separate univariate

parameters. This agrees with the intuition that a work ought to be evaluated in a multidimensional space, and it happens to speed up estimation and reduces memory requirements in the statistical software we use. Finally, this model treats dimension difficulty as its own first-level parameter, as opposed to a somewhat awkward representation where dimension difficulty is a hyperparameter of the quality of the work.

2.2 Multidimensional Reviewer Bias

The Unidimensional Reviewer Bias model treats reviewer bias as constant across all rubric dimensions. For example, reviewer bias may be a form of response style where the reviewer is generally disposed to be more lenient or more critical. On the contrary, it is possible that a reviewer may assess peer works more strictly in some dimensions than in others.

The Multidimensional Reviewer Bias model relaxes the unidimensionality assumption. Specifically, it represents reviewer bias as $\boldsymbol{\rho_r} \sim N_d(\boldsymbol{\mu^\rho}, \boldsymbol{\Sigma^\rho})$, i.e., normal and multivariate in the number of assessment dimensions, with partial pooling such that all quality estimates are shrunk towards the mean $\boldsymbol{\mu^\rho}$ of their shared distribution. The model treats dimensions as independent of each other by fixing their covariances at zero in the covariance matrix $\boldsymbol{\Sigma^\rho}$.

2.3 Relationship of Reviewer Bias to Instructor Assessment

The Unidimensional and Multidimensional Reviewer Bias models treat a student's bias as unrelated to instructor assessment of the work of that student. An alternative hypothesis is that a reviewer who on average gives lower ratings than other reviewers is more proficient with the subject matter than other reviewers, and the lower ratings reflect a proficient student's ability to notice errors that less proficient students do not. In the case that all students both author and review, as with the datasets analyzed here, we can examine whether a student's bias as a reviewer may also contribute to estimating the instructor assessment of the work of this student. The model testing this explanation requires some assumptions, as follows.

When an instructor creates an assessment rubric to be used for peer review, the rubric is necessarily a subset and a simplification of the instructor's actual assessment criteria. For example, suppose an instructor creates a rubric, sets it aside, and proceeds to assess a student's work. In a set of student papers, the instructor may well encounter something unanticipated, e.g., a new perspective on the problem, or a nuance in the analysis. The instructor will proceed to assess the paper according to substantive criteria that were not reflected in the rubric. Nonetheless, peer reviewers may be constrained to follow the rubric in assessing the same paper. In this way, the instructor's assessment must be a better measure of the quality of the paper than a peer review.

Therefore, if reviewer bias does reflect a student's proficiency, and if peer assessment via a rubric is limited in capturing this proficiency, the student's reviewer bias will be related to the instructor's assessment of the student's work. Specifically, a positive estimate of reviewer bias, i.e., $\rho_r > 0$, signifies a student who is lenient. If

leniency is negatively related to proficiency, then reviewer bias will be negatively related to instructor assessment.

Accordingly, the Unidimensional Bias-Instructor Assessment Relationship model modifies the Unidimensional Reviewer Bias model so that $Ins \sim N(\boldsymbol{\theta_p}\boldsymbol{\beta} + \rho_r\zeta + \eta, \sigma^{Ins})$, i.e., the instructor assessment of a student's work is modeled as a linear function of an intercept η, the peer assessments of the quality of the work, and finally the reviewer bias ρ_r with weight ζ that describes the extent to which the reviewer bias contributes to the instructor estimate. The hypothesized negative relationship between bias and instructor assessment, if true, would be signaled by a negative ζ estimate.

This model can be extended to the case of multidimensional reviewer bias if we hypothesize that a student will be more sensitive to errors in the work under review along a rubric dimension that she has mastered than along a dimension that she has yet to master. The Multidimensional Bias-Instructor Assessment Relationship model modifies the Multidimensional Reviewer Bias model so that $Ins \sim N(\boldsymbol{\theta_p}\boldsymbol{\beta} + \boldsymbol{\rho_r}\boldsymbol{\zeta} + \eta, \sigma^{Ins})$, i.e., the reviewer's bias in each dimension is weighted separately by the vector of coefficients $\boldsymbol{\zeta}$ to estimate its relationship to the instructor assessment.

If the hypothesized explanation of reviewer bias holds, the Bias-Instructor Assessment Relationship models would estimate instructor assessment better than the Reviewer Bias models.

2.4 Value of Accounting for Reviewer Bias

Even if accounting for reviewer bias as a source of variance should improve model fit, the effort to do so may not be worthwhile. Such accounting requires the estimation of additional parameters (one per reviewer in the unidimensional bias representation, or one per reviewer per dimension in the multidimensional representation), and these extra parameters may lead the model to overfit the data. For comparison, we fit the Ignoring Reviewer Bias model that omits the reviewer bias term altogether.

2.5 Datasets

The models were fit to peer and instructor assessments of a written exam on Intellectual Property law. [13, 14] The exam comprised one essay-type question that asked students "to provide advice concerning [a particular party's] rights and liabilities" given 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 involved different legal interests and requirements and presented 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.

To review each other's exam answers, students used Comrade, a web-based peer review application. Students were assigned to one of two experimental conditions randomly and balanced with respect to LSAT score.[1] The difference between conditions was the assessment rubric that was embedded in the Comrade user interface. The DG rubric (n=29) was designed to be generally applicable to assessment of legal writing; its dimensions pertained to four legal writing skills. The PS rubric (n=28) was tailored to the exam question itself; its dimensions pertained to five legal concepts that the instructor considered to be relevant to the question. Each student was assigned to review the (de-identified) exam answers of four peers. For each exam answer, for each rubric dimension, the reviewer gave a rating on a 7-point scale, and gave a written explanation of the rating. Scale points 1, 3, 5 and 7 were anchored to definitions that clarified the meaning of the scale.

2.6 Model Fitting

Models were compared using Deviance Information Criterion (DIC), and calculating prediction error. The DIC, a Bayesian-modeling analogue of the better-known AIC, rewards models that fit data with low deviance, i.e., that minimize residual error. Additionally, because residual error decreases when parameters are added to a model, DIC also penalizes models that have too many parameters. A model with low DIC is preferred to one with high DIC. DIC can be interpreted as a prediction about which model will perform best on held-out data. DIC uses all available data for a more powerful model comparison than validation on held-out data, but it can overfit the data.

To measure what residual error the models incur on held-out data, we also performed student-stratified cross-validation. Omitting the instructor assessment (but not the peer assessment) for one student's work, we fit the model to the remaining data, and used the model to predict the omitted assessment. This procedure was repeated to predict the instructor assessment for the work of each student. Prediction quality was measured using Root Mean Squared Error (RMSE).

The Bayesian models were fit under OpenBUGS. [15] For DIC, each model was fit with 3 chains of Markov Chain Monte Carlo (MCMC), discarding 1000 initial samples (burn-in), and taking every 25^{th} sample (thinning) to reduce autocorrelation of the subsequent 5000 iterations. Convergence of chains was checked using the Gelman-Rubin *Rhat* statistic. Having checked the models, we fit slightly shorter chains for cross-validation (500 burn-in, 3000 iterations, and thinning of 15).

3 Results and Discussion

Each condition of the peer review exercise produced a dataset. The PS dataset consisted of 527 ratings of the exams of 28 students by 4 peer reviewers each along 5 rubric dimensions. The DG dataset consisted of 453 ratings of the exams of 29 students by 4 peer reviewers each along 4 rubric dimensions.

[1] The LSAT is a standardized test used for admission to American law schools.

Table 1: Model performance; best scores in bold

Model	DIC-PS	RMSE-PS	DIC-DG	RMSE-DG
Ignoring Reviewer Bias	1760	0.684	1384	1.066
Unidimensional Reviewer Bias	**1721**	0.715	**1267**	**1.011**
Multidimensional Reviewer Bias	1739	0.690	1296	1.048
Unidimensional Bias-Instructor Assessment Relationship	1739	0.717	1276	1.092
Multidimensional Bias-Instructor Assessment Relationship	**1723**	**0.677**	**1267**	1.104
Mean of Instructor Assessments		0.964		1.252

According to both DIC and RMSE, the best-performing model on the PS dataset was Multidimensional Bias-Instructor Assessment Relationship, and on the DG dataset, the best performing model was Unidimensional Reviewer Bias. (Tab. 1)

According to both DIC and RMSE, ignoring reviewer bias altogether is inappropriate. Including multidimensional reviewer bias in estimating instructor assessment improved the estimate of instructor assessment on PS over all other models (DIC and RMSE). DIC and RMSE diverge in some cases, but unidimensional bias, with no contribution to instructor assessment, is the preferred model on DG.

Our prior work showed that peer ratings elicited via the PS rubric were uncorrelated across rubric dimensions, whereas ratings elicited via the DG rubric were highly correlated across dimensions. This new analysis showed that distinctions among PS dimensions lead both to uncorrelated, multidimensional assessments of student works, and to distinctions among the biases of the reviewers using the rubric.

To evaluate the models' predictions in an absolute sense, we also measured RMSE of a baseline that computes the mean of the instructor assessments on each dataset and treats this as the "predicted" instructor assessment for every student. (Tab. 1, last row) We find that all models improve prediction against this baseline on both datasets, with the best model reducing error by 30% on PS and 19% on DG.

4 Conclusions

The findings imply that reviewer bias is an issue for both rubrics, and that statistical methods may be used to account for bias. Most intriguingly to us is that bias is not a mere tendency to assess peer works consistently higher (or lower) than other reviewers. First, bias can function as a multidimensional construct when peer assessments are elicited using an appropriate rubric. Second, because a multidimensional representation of bias improves prediction of instructor assessment, it seems likely that bias is an expression of some cognitive trait. For instance, bias may reflect student proficiency, as hypothesized above. Bias clearly plays a role in peer review, and it may plausibly affect other instructional activities as well. In particular, collaborative learning activities involving peer assessment, whether staged for instruction or research, may need to consider the impact of reviewer bias.

5 References

1. Dillenbourg, P.: Over-scripting CSCL: The risks of blending collaborative learning with instructional design. Three worlds of CSCL. Can we support CSCL. pp. 61–91. Open Universiteit Nederland, Heerlen (2002).
2. Cho, K., Schunn, C.D., Wilson, R.W.: Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. Journal of Educational Psychology. 98, 891–901 (2006).
3. Rorer, L.G.: The great response-style myth. Psychological Bulletin. 63, 129–156 (1965).
4. Muthukumarana, S.: Bayesian Methods and Applications Using WinBUGS, (2010).
5. Cho, K., Schunn, C.D.: Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. Computers and Education. 48, (2007).
6. Hamer, J., Ma, K.T.K., Kwong, H.H.F.: A method of automatic grade calibration in peer assessment. Proceedings of the 7th Australasian conference on Computing education - Volume 42. pp. 67–72. Australian Computer Society, Inc., Newcastle, New South Wales, Australia (2005).
7. Lauw, H.W., Lim, E., Wang, K.: Summarizing review scores of "unequal" reviewers. In Proceedings of the 7th SIAM International Conference on Data Mining (2007).
8. Russell, A.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).
9. Masters, J., Madhyastha, T., Shakouri, A.: ExplaNet: A collaborative learning tool and hybrid recommender system for student-authored explanations. Journal of Interactive Learning Research. 19, 51–74 (2008).
10. Crespo García, R.M., Pardo, A., Delgado Kloos, C.: Adaptive Peer Review Based on Student Profiles. In: Ikeda, M., Ashley, K., and Chan, T.-W. (eds.) Intelligent Tutoring Systems. pp. 781–783. Springer Berlin / Heidelberg (2006).
11. Lu, R., Bol, L.: A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. Journal of Interactive Online Learning. 6, 100–115 (2007).
12. Goldin, I.M., Ashley, K.D.: Eliciting Formative Assessment in Peer Review. Journal of Writing Research. Special Issue: Redesigning Peer Review Interactions Using Computer Tools, (under review).
13. Goldin, I.M., Ashley, K.D.: Peering Inside Peer Review with Bayesian Models. In: Biswas, G., Bull, S., Kay, J., and Mitrović, A. (eds.) Artificial Intelligence in Education. pp. 90–97. Springer Berlin Heidelberg, Berlin, Heidelberg (2011).
14. Goldin, I.M.: A Focus on Content: The Use of Rubrics in Peer Review to Guide Students and Instructors, http://d-scholarship.pitt.edu/8375/, (2011).
15. Lunn, D., Spiegelhalter, D., Thomas, A., Best, N.: The BUGS project: Evolution, critique and future directions. Statistics in medicine. 28, 3049–3067 (2009).