Debiasing Evaluations That are Biased by Evaluations (Extended Abstract)

Jingyan Wang, Ivan Stelmakh, Yuting Wei, Nihar B. Shah
Carnegie Mellon University
{jingyanw, stiv}@cs.cmu.edu, ytwei@stat.cmu.edu, nihars@cs.cmu.edu

It is common to aggregate information and evaluate items by collecting ratings of these items from people. For example, many universities use student ratings as one of the mechanisms for teaching evaluation. However, such ratings can inherit biases from people that are irrelevant to the metric of interest. Going back to the example of teaching evaluation, numerous studies have shown that student ratings can be affected by the grading policy of the instructor (e.g., [6, 7, 3]). In this work, we study the problem of correcting such biases in ratings. In particular, we focus on the bias introduced by people's observable outcome or experience from the item under evaluation, and we call it the "outcome-induced bias". Before describing the exact problem formulation, we first describe in more detail two applications that are prevalent in an academic context – teaching evaluation and peer review – and the corresponding outcome-induced bias involved.

It is well known that student ratings are highly unreliable in reflecting the teaching quality truthfully. Indeed, in some circumstances the association between student ratings and teaching effectiveness is negative [3], and student ratings also serve as a poor predictor on the follow-on course achievement of the students [5, 4]:

> *"...teachers who are associated with better subsequent performance receive worst evaluations from their students."* [4]

The outcome we consider in teaching evaluation is the grades that the students receive in the course under evaluation[1]. The student ratings, as one may expect, are known to be substantially influenced by the students' received grades or perceived grade expectations (e.g., [7, Chapter 3] and references therein):

> *"...the effects of grades on teacher–course evaluations are both substantively and statistically important, and suggests that instructors can often double their odds of receiving high evaluations from students simply by awarding A's rather than B's or C's."* [7]

An analogous issue arises in conference peer review, where conference organizers conduct surveys of authors to rate their received reviews in order to understand the quality of the review process. The outcome we consider in peer review is the reviewers' scores or the final paper decisions [9, 1]. Again, as one may expect, authors are more likely to give higher ratings to a positive review than a negative one [10, 9, 1]:

> *"Satisfaction had a strong, positive association with acceptance of the manuscript for publication... Quality of the review of the manuscript was not associated with author satisfaction."* [10]

Notably, an author feedback experiment was conducted at the PAM 2007 conference. and it was observed that

> *"some of the TPC members from academia paralleled the collected feedback to faculty evaluations within universities... while author feedback may be useful in pinpointing extreme cases, such as exceptional or problematic reviewers, it is not quite clear how such feedback could become an integral part of the process behind the organization of a conference."* [9]

[1] We use the term "grades" broadly to include both letter grades and numerical scores. We do not distinguish the difference between evaluation of a course and evaluation of the instructor teaching the course, and use them interchangeably.

Although the existence of such bias is widely acknowledged, student and author ratings are still widely used [2], and such usage can be problematic. By taking the outcome-induced bias as part of the estimated quality, it can be unfair to reviewers or instructors who are rigorous in reviewing and grading. This unfairness is exacerbated by the fact that author ratings can be a factor for selecting reviewer awards [1], and student ratings can be a heavily-weighted component for salary or promotion and tenure decision of the faculty members [2, 5, 3]. Therefore, one can imagine that naively using these ratings or simply taking their mean or median is not sufficient; interpreting and correcting these ratings properly is an important practical problem.

Moreover, the outcome-induced bias also introduces incentives for the instructors to improve the student evaluation by manipulating the outcome, raising concerns such as inflating grades, reducing content [5], and "teach to the test" [4], so that students spend less effort but get higher grades, mistakenly equating that to better teaching quality. Despite the higher ratings, these incentives are actually counter-productive for improving teaching quality; correcting the outcome-induced bias will in turn reduce such undesirable incentives.

The goal of this work is to correct such outcome-induced bias in ratings (to some extent). Incidentally, in teaching evaluation and peer review, the "outcome" that people (students or authors) encounter in the process is the evaluation they receive (grades from instructors or reviews from reviewers), and hence we call it "evaluations that are biased by evaluations". However, the general problem we consider here is applicable to other settings with outcomes that are not necessarily evaluations. For example, in evaluating whether a two-player card game is fair or not, the outcome can be whether the player wins or loses the game [8].

The key insight we use in this work is that the outcome (e.g., grades and paper decisions) is observed by the people who conduct the evaluation (e.g., universities and conference organizers). These observed outcomes provide directional information about the manner that people are likely to be biased. Intuitively, for example, students receiving higher grades are likely to give higher ratings to the course instructor than students receiving lower grades (if all other things being equal). This directional information is key to our model and proposed estimators.

The contributions of this work are as follows.

- **Formulation** We assume the ratings by the students (in the running example of teaching evaluation) is the summation of the true quality of the course, a bias term dependent on the outcome (i.e., grades), and an i.i.d. noise term. We model the directional information from the bias as a partial ordering constraint on the biases across students and courses, where students receiving higher grades have bias towards giving higher ratings. Hence, in this model the inter-difference (i.e., difference in means) across the items is due to the true value of the items (e.g., course quality), and the intra-difference within each item is due to a combined effect of bias and noise. This formulation is general, without making parametric assumptions on the relation between the bias and the outcome that can be application-specific. The goal is to estimate the true quality of all the courses.

- **Proposed approach** Our approach consists of a class of estimators and a cross-validation procedure. We propose a class of estimators in the form of solving an optimization formulation. This class of estimators jointly minimizes a loss function over the true quality of the courses, and biases subject to the partial ordering constraint. The loss function consists of a "fitting error" – the squared ℓ_2 difference between the observed ratings and the sum of the estimated qualities and biases – and a standard ℓ_2-regularization. The class of estimators is parameterized by a hyper-parameter λ controlling the extent of the ℓ_2-regularization. To determine an appropriate value of this hyperparameter, we further propose a cross-validation approach that chooses this hyperparameter by minimizing a carefully-defined fitting error on a validation set. The estimator and the cross-validation method do not operate under specific shape assumptions on the bias or the noise, and hence are completely data-dependent.

- **Theoretical guarantees** We show that the estimator is provably minimax-optimal at $\lambda = 0$ and $\lambda = \infty$ respectively for two extremal cases (when there is only bias, and when there is only noise). We then show that for a certain class of partial orderings, the cross-validation method correctly converges these optimal solutions in probability in these two extremal cases.

- **Simulation** We supplement our theoretical results by simulation in more general settings – in cases where both the bias and the noise are present, and in cases for general partial orderings outside the class analyzed in our theoretical results.These simulation results demonstrate the effectiveness of our proposed estimator and cross-validation procedure.

References

[1] Serge Belongie Aditya Khosla, Derek Hoiem. Analysis of reviews for CVPR 2012. 2013.

[2] William E. Becker and Michael Watts. How departments of economics evaluate teaching. *The American Economic Review*, 89(2):344–349, 1999.

[3] Anne Boring, Kellie Ottoboni, and Philip B. Stark. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 0(0):1–11, 2016.

[4] Michela Braga, Marco Paccagnella, and Michele Pellizzari. Evaluating students' evaluations of professors. *Economics of Education Review*, 41:71 – 88, 2014.

[5] Scott E Carrell and James E West. Does professor quality matter? evidence from random assignment of students to professors. Working Paper 14081, National Bureau of Economic Research, June 2008.

[6] AG Greenwald and GM Gillmore. Grading leniency is a removable contaminant of student ratings. *The American psychologist*, 52(11):1209–1217, November 1997.

[7] Valen E. Johnson. *Grade Inflation: A Crisis in College Education*. 2003.

[8] Mario D. Molina, Mauricio Bucca, and Michael W. Macy. It's not just how the game is played, it's whether you win or lose. *Science Advances*, 5(7), 2019.

[9] Konstantina Papagiannaki. Author feedback experiment at pam 2007. *SIGCOMM Comput. Commun. Rev.*, 37(3):73–78, July 2007.

[10] Ellen J Weber, Patricia P Katz, Joseph F Waeckerle, and Michael L Callaham. Author perception of peer review: impact of review quality and acceptance on satisfaction. *JAMA*, 287(21):2790–2793, 2002.