TheWebConf 2020 Tutorial on Fairness and Bias in Peer Review and other Sociotechnical Intelligent Systems (Part II on Peer Review)

Nihar B. Shah and Zachary Lipton
Carnegie Mellon University
nihars@cs.cmu.edu zlipton@cmu.edu

Abstract

Peer review is the backbone of scholarly research, but it faces a number of challenges pertaining to bias and unfairness. There is an urgent need to improve peer review. This TheWebConf tutorial (part 2) discusses several problems, empirical studies, proposed solutions, and open problems in this domain. This document serves to provide a summary and references for the tutorial.

1 Introduction

Peer review is a cornerstone of academic practice today and also for years to come (Price and Flach, 2017). The peer review process is highly regarded by the vast majority of researchers and considered by most to be essential to the communication of scholarly research (Mulligan et al., 2013; Nicholas et al., 2015; Ware, 2008). However, there is also an overwhelming desire for improvement (Smith, 2006; Ware, 2008; Mulligan et al., 2013).

The following quote from Rennie (2016), in a Nature commentary titled "Lets make peer review scientific" provides an apt summary of the state of peer review today:

> *"Peer review is touted as a demonstration of the self-critical nature of science. But it is a human system. Everybody involved brings prejudices, misunderstandings and gaps in knowledge, so no one should be surprised that peer review is often biased and inefficient. It is occasionally corrupt, sometimes a charade, an open temptation to plagiarists. Even with the best of intentions, how and whether peer review identifies high-quality science is unknown. It is, in short, unscientific."*

The need to improve peer review is particularly urgent due to the explosion in the number of submitted papers in various fields. Conferences in machine learning and artificial intelligence are experiencing a near-exponential growth in the number of submissions. The increase in number of submissions is also large in many other fields beyond computer science: according to McCook (2006) *"Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint"*.

Peer review is particularly known to hinder novel and interdisciplinary research. Quoting Travis and Collins (1991): *"interdisciplinary research, frontier science, areas of controversy, and risky new departures are all more likely to suffer from cognitive cronyism than is mainstream research."* See also Church (2005); Porter and Rossini (1985); Lamont (2009). Naughton (2010) has makes a noteworthy point: *"Today reviewing is like grading: When grading exams, zero credit goes for thinking of the question. When grading exams, zero credit goes for a novel approach to solution. (Good) reviewing: acknowledges that the question can be the major contribution. (Good) reviewing: acknowledges that a novel approach can be more important than the existence of the solution."*

Problems in peer review have consequences much beyond the outcome for a specific paper, particularly due to the widespread prevalence of the Matthew effect ("rich get richer") in academia (Merton, 1968). As noted by Triggle and Triggle (2007) *"an incompetent review may lead to the rejection of the submitted paper, or of the grant application, and the ultimate failure of the career of the author."* (See also Thorngate and Chowdhury, 2014; Squazzoni and Gandelli, 2012.)

Lee (2015) thus asks: *"In public, scientists and scientific institutions celebrate truth and innovation. In private, they perpetuate peer review biases that thwart these goals... what can be done about it?"*

The importance of peer review and the urgent need for improvements, behooves research on principled approaches towards addressing problems in peer review, particularly at scale. In this tutorial, we outline a few directions of research, and emphasize that this is just the tip of the iceberg.

For concreteness we restrict attention to (conference) peer review of scholarly research, but emphasize that research on this topic has implications for a wide variety of applications such as crowdsourcing, A/B testing, peer grading, recommender systems, hiring, college admissions, and many others. The common thread among these applications and peer review is that they involve distributed human evaluations—a set of people need to evaluate a set of items, but every item is evaluated by a small subset of people and every person evaluates only a small subset of items.

In the following sections, we discuss the following issues related to unfairness in peer review: biases; noise; dishonest behavior; miscalibration; subjectivity; and norms and policies. We draw conclusions in the final section of this document.

2 Biases

We begin with a discussion on issues related to biases with respect to certain groups of people. There is a lot of debate on whether peer review should be single blind (i.e., reviewers know authors' identities) or double blind (i.e, reviewers do not know authors' identities), and different communities follow different approaches. A primary argument against single blind is that it may cause the review to be biased with respect to the gender/race/fame or other attributes of the authors. For example, a paper submitted by two women authors to PLOS ONE received a review: *"It would probably be beneficial to find one or two male researchers to work with (or at least obtain internal peer review from, but better yet as active co-authors)"* (Bernstein, 2015). This debate can be made more informative via experiments and data collection about this topic, which in turn requires the design of appropriate tools and techniques to do so.

The issue of such biases in peer review is investigated in many prior works (Reinhart, 2009; Budden et al., 2008; Webb et al., 2008; Okike et al., 2016; Bernard, 2018; Bennett et al., 2018; Seeber and Bacchelli, 2017; Snodgrass, 2006; Madden and DeWitt, 2006; Tung, 2006; Swim et al., 1989; Blank, 1991; Lee et al., 2013), primarily in journals and in non-computer-science fields.

In computer science, and particularly in the conference-review setting, a remarkable experiment was conducted at the WSDM 2017 conference by its program chairs (Tomkins et al., 2017). The reviewers were split uniformly at random into two groups – a single blind group and a double blind group – and each paper was assigned two reviewers each from both groups. This allowed for a direct comparison of single blind and double blind reviews for each paper while requiring a number of reviews only as much as what would occur in a non-experimental setting. In a nutshell, their results found a significant bias towards famous authors, top universities, and top companies. They also found a high effect size but not statistically significant bias against papers with at least one woman author (a meta-analysis in combination with other studies was statistically significant). The experiment did not find evidence of bias with respect to papers from the United States, when reviewers were from the same country as the authors, and for/against academic (versus industrial) institutions. The WSDM conference moved to double blind from the following year.

A subsequent work (Stelmakh et al., 2019) offers a note of caution that the peer review process has a number of peculiar characteristics due to which any experimental setup or statistical test requires a careful design. It offers a number of possible scenarios which can break the tests used in the WSDM experiment and designs a new experimental setup and statistical tests with rigorous guarantees.

Open problems: The tests of Stelmakh et al. (2019) have only asymptotic guarantees on its power, and finite sample guarantees on power for this problem remain open. Moreover, this test requires a semi-randomized controlled trial; the design of tests (and quantification of needed assumptions) to test for biases from observational data incorporating the idiosyncracies of peer review remains an important open problem. Finally, there is need for many more such experiments that can help inform the discourse on peer review and make it more "scientific".

3 Noise

By noise, here we mean poor reviews due to inappropriate choice of reviewers. Data from people is often noisy due to lack of expertise. In peer review, the assignment of the reviewers to papers determines the expertise of the reviewer who will review any paper. Indeed, the importance of the reviewer-assignment stage of the peer-review process cannot be overstated: quoting Rodriguez et al. (2007), *"one of the first and potentially most important stage is the one that attempts to distribute submitted manuscripts to competent referees."* A survey of researchers McCullough (1989) indicated that the top reason for author dissatisfaction was that *"Reviewers or panelists not expert in the field, poorly chosen, or poorly qualified"*.

The assignment of reviewers to papers in most large conferences (such as ICML, NeurIPS, AAAI and others) is performed in an automated fashion. There are two stages in the assignment procedure. The first stage involves computing a "similarity score" between every reviewer-paper pair (Mimno and McCallum, 2007; Liu et al., 2014; Rodriguez and Bollen, 2008; Tran et al., 2017; Charlin and Zemel, 2013). A higher similarity scores means a better envisaged quality of review. The second stage then uses these similarity scores to assign reviewers to papers in a manner that maximizes some function of the similarities of the assigned reviewer-paper pairs.

The most popular assignment method is to maximize the total sum of the similarities of all assigned reviewer-paper pairs (Goldsmith and Sloan, 2007; Tang et al., 2010; Charlin et al., 2012; Long et al., 2013). This method is followed in the Toronto Paper Matching System (Charlin and Zemel, 2013) which is widely used in many conferences and is also followed in conference management systems such as EasyChair (`https://easychair.org`) and HotCRP (`https://hotcrp.com/`).

The aforementioned approach of maximizing total sum of similarities, however, can result in unfairness to certain papers (see Stelmakh et al., 2018 for an example). An alternative approach is to optimize for the paper with the minimum sum similarity, and subject to that, optimize for the paper with the next smallest sum similarity and so on (Stelmakh et al., 2018; see also Garg et al., 2010; Benferhat and Lang, 2001; Hartvigsen et al., 1999). Empirical evaluations for such an approach in three major conferences are available in Kobren et al. (2019).

Open problems: Among the assignment algorithms in the literature, there is a tradeoff between the fairness guarantees and the computational complexity of the assignment algorithm (Stelmakh et al., 2019; Kobren et al., 2019), and designing assignment algorithms that are computationally faster and have strong fairness guarantees is an important open problem. The second direction pertains to a better computation of the similarity scores, taking into account the various aspects of peer review, or furthermore jointly compute the similarity and assignment (Mimno and McCallum, 2007; Rodriguez and Bollen, 2008; Charlin and Zemel, 2013; Liu et al., 2014; Tran et al., 2017). Third, many conferences adopt a "bidding" procedure before the assignment stage, in which reviewers can bid for the papers they wish or don't wish to review. The bidding procedure is one of the most under-studied phases of the review process, and there is much to be done to make it more fair and efficient (Fiez et al., 2019).

4 Dishonest behavior

Peer-review is susceptible to strategic manipulations. A reviewer may be able to increase the chances of acceptance of their own submissions by manipulating the reviews (e.g., providing lower scores) for other papers. A recent empirical study Balietti et al. (2016) examined the strategic behavior of people in competitive peer review, and concluded that *"...competition incentivizes reviewers to behave strategically, which reduces the fairness of evaluations and the consensus among referees."* See Akst (2010); Anderson et al. (2007); Langford (2008) for more anecdotes. As Thurner and Hanel (2011) posit, even a small number of selfish, strategic reviewers can drastically reduce the quality of scientific standard.

It is thus highly important to protect peer review from any possible strategic manipulations. We define strategyproofness in terms of a "conflict graph", which is a fixed graph given to us. A conflict graph is a bipartite graph with all reviewers and papers as its vertices, and has an edge between a reviewer vertex and a paper vertex if the reviewer has a conflict with the paper. Examples of conflicts include authorship conflicts (e.g., the reviewer is an author of that paper), institutional conflicts, etc. Then strategyproofness means that no reviewer must be able to influence the final ranking of her/his conflicted papers by manipulating the reviews that she/he provides.

A number of past works (Alon et al., 2011; Holzman and Moulin, 2013; Bousquet et al., 2014; Fischer and Klimm, 2015; Kurokawa et al., 2015; Kahng et al., 2017) consider designing strategyproof procedures of "peer grading" in MOOCs and classrooms. There are two key differences between these peer-grading settings and the peer-review setting. First, the peer grading setting involves conflict graphs of degree at most 1, that is, every reviewer conflicts with at most one paper and every paper has at most one author. On the other hand, even if one considers only authorship conflicts in conference peer review, every author may submit multiple papers and any paper may have multiple authors, thus requiring strategyproofness with respect to more general graphs. Second, these prior works do not account for "heterogeneity" in the papers and reviewers with the motivation that all students in peer grading take the same course. On the other hand, conference papers and reviewers are more diverse in terms of their expertise and subject matter. Hence any peer-review framework must have significant flexibility to accommodate the various intricacies. These differences make the peer-review setting strictly more general and significantly more challenging.

The partitioning-based method is used for the peer review setting by Xu et al. (2019). In addition to theoretical guarantees, Xu et al. (2019) also perform an empirical analysis on data from ICLR 2017 and 2018.

Open problems: Is strategyproofing possible when conflict graph cannot be partitioned (Xu et al., 2019; Aziz et al., 2019)? What is the maximum efficiency under strategyproofness, where efficiency may be defined as the quality of the reviewer-paper assignment (Xu et al., 2019)? Finally, how can one detect and/or prevent other forms of dishonest behavior (Ferguson et al., 2014; Gao and Zhou, 2017; Langford, 2012a)?

5 Miscalibration

There are many applications which ask people to provide ratings. However, it is well known (Mitliagkas et al., 2011; Ammar and Shah, 2012; Griffin and Brenner, 2008; Freund et al., 2003; Harzing et al., 2009) that the same rating score may have different meanings for different individuals. For instance, if reviewers are asked to provide scores in the interval $[0, 1]$, some reviewers may be lenient and always provide scores greater than 0.5 whereas some others may be strict and rarely give scores above 0.5. Or some reviewers are more moderate whereas others provide scores at the extremes of the allowed interval. Such mismatches cause additional difficulty in the final acceptance decisions as well as lead to unfairness, as noted by Siegelman (1991): *"the existence of disparate categories of reviewers creates the potential for unfair treatment of authors. Those whose papers are sent by chance to assassins/demoters are at an unfair disadvantage, while zealots/pushovers give authors an unfair advantage."* Miscalibration may also be due to mismatched expectations of the "bar" for acceptance. In NeurIPS 2016, there was a significant difference between the expected scores and the scores given by reviewers (Shah et al., 2018).

In the literature, there are two popular approaches towards this problem miscalibration. The first approach (Paul, 1981; Flach et al., 2010; Roos et al., 2011; Ge et al., 2013; Baba and Kashima, 2013; MacKay et al., 2017) is to make simplifying assumptions on the nature of the miscalibration, for instance, assuming that these miscalibration is linear or affine. The research following this approach designs algorithms to learn "parameters" of the miscalibration.

The simplistic assumptions described above are known to be frequently violated (see Brenner et al., 2005; Griffin and Brenner, 2008 and references therein). These algorithms based on these assumptions can then be "significantly harmful" in practice (Langford, 2012b). With this motivation, a second approach (Rokeach, 1968; Freund et al., 2003; Harzing et al., 2009; Mitliagkas et al., 2011; Ammar and Shah, 2012; Negahban et al., 2012) towards handling miscalibrations is to either directly elicit rankings from reviewers or convert the scores into rankings. This approach is often believed to be the only resort when the underlying miscalibration may be arbitrary. However, it is shown in Wang and Shah (2019b) that in contrast to this folklore belief, even if the miscalibration is arbitrary or adversarially chosen, ratings can yield better results than rankings. The estimators proposed in this work, however, are randomized and tailored for the worst case.

Open problems: An important open problem is to design practically useful calibration algorithms that accommodate non-parametric, non-linear models (i.e., weaker than parametric assumptions of some past literature) but not as weak as the adversarial assumptions of Wang and Shah (2019b), e.g., using permutation-based models which have several benefits as compared to traditional models in various applications (Shah et al., 2017, 2019b, 2016, 2019a; Shah and Wainwright, 2018; Heckel et al., 2016). Moreover, we need the designed algorithms to be amenable to the small sample sizes that are typical of peer review, perhaps achieved via different means of data elicitation or a more relaxed space for outcomes (e.g., not necessarily outputting a total ranking or parameter values).

6 Subjectivity

It is known that different reviewers have different, subjective opinions about the relative importance of various criteria in judging papers (Church, 2005; Lamont, 2009; Bakanic et al., 1987; Hojat et al., 2003; Mahoney, 1977). On the other hand, in order to ensure fairness, every paper should ideally be judged by the same yardstick. For instance, suppose three reviewers consider "improvement of at least 10%" as most important, whereas most members of the community have a high emphasis on "novelty". Then a highly novel paper that yields a 5% improvement over the state of the art may be rejected if reviewed by these three reviewers but would have been accepted by any other set of reviewers. Indeed, as revealed in the survey by Kerr et al. (1977), more than 50% of reviewers say that even if the community thinks a certain characteristic of a manuscript is good, if the reviewer's own opinion is negative about that characteristic, it will count against the paper; about 18% say this can also lead them to reject the paper. Lee (2015) calls this issue "commensuration bias."

Noothigattu et al. (2018) propose an approach to alleviate this problem. They model the problem as that of "learning" a mapping from individual criteria to a final score, that is common to the set of all reviewers. Marrying machine learning with social choice theory, they take an axiomatic approach towards designing the learning algorithm in a principled manner. They also present an analysis on peer-review data from IJCAI 2017.

Open problems: What are the statistical properties of the above problem and the algorithm of Noothigattu et al. (2018)? How can one evaluate the performance of any peer review systems or algorithms, particularly since there is no ground truth in terms of which papers are actually of higher quality than others? How can the the various aforementioned issues — biases, noise, dishonest behavior, miscalibration, and subjectivity — which may not be separable in the data be handled together?

7 Norms and policies

Issues of biases and unfairness also arise due to the norms and policies followed by certain communities or conferences.

(a) Biases due to alphabetical ordering: Einav and Yariv (2006) study biases due to alphabetical ordering in the field of Economics, where they find a significant bias towards researchers with last names earlier in the alphabet. Economics follows the norm of listing authors in papers in alphabetical order of their last names. In contrast, they find no such bias in the related field of Psychology where the ordering is typically done in terms of the authors' contributions. (See also Hilmer and Hilmer, 2005; Van Praag and Van Praag, 2008.)

Ordering authors alphabetically results in biases due to several reasons. First, primacy effects imply that the reader will tend to remember the authors listed earlier in the ordering. Moreover, many communities use the "first author et al." citation format that puts a significantly greater emphasis on the first author. For instance, more than half the papers in STOC, FOCS and EC conferences — which follow the norm of ordering authors alphabetically — used the "first author et al." citation format (Wang and Shah, 2018).

A related application is the lists of people on websites, for instance, lists of students and/or faculty on the websites of universities. These lists are also often ordered alphabetically, resulting in biases due to serial position effects.

A proposed solution to this problem is to randomize the lists of authors on papers (Ray and Robson, 2018) or (dynamically) randomize the ordering of people on websites. Following outreach by Wang and Shah (2018), the Machine Learning Department at Carnegie Mellon University randomizes the lists of people on its website (`https://www.ml.cmu.edu/people/`) since October 24, 2019.

(b) Gender distribution in paper awards and need for transparency: The gender distribution of paper awardees in top computer science conferences is quite skewed (Wang and Shah, 2019a). At the least, this suggests the need for greater transparency in the award processes, for instance, publishing whether the process was double or single blind, or the criteria that was used. This data has stated conversations in a number of research communities (e.g., Erkip, 2019), and the hope is that such data and outreach will stimulate some much-needed changes in the norms and policies adopted by various communities.

8 Conclusions

There are many sources of biases and unfairness in peer review. The need to improve peer review is important and urgent for scholarly research to thrive. There is a lot at stake beyond an individual paper: careers of researchers and the progress of science. The current research on peer review has only scratched the surface of this important and urgent problem domain. There are lots of open problems which are exciting, challenging, impactful, and allow for an entire spectrum of theoretical, applied, and conceptual contributions.

References

Jef Akst. I Hate Your Paper. Many say the peer review system is broken. Here's how some journals are trying to fix it. *The Scientist*, 24(8):36, 2010.

Noga Alon, Felix Fischer, Ariel Procaccia, and Moshe Tennenholtz. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 101–110. ACM, 2011.

Ammar Ammar and Devavrat Shah. Efficient rank aggregation using partial data. In *SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, 2012.

Melissa S Anderson, Emily A Ronning, Raymond De Vries, and Brian C Martinson. The perverse effects of competition on scientists work and relationships. *Science and engineering ethics*, 13 (4):437–461, 2007.

Haris Aziz, Omer Lev, Nicholas Mattei, Jeffrey S Rosenschein, and Toby Walsh. Strategyproof peer selection using randomization, partitioning, and apportionment. *arXiv preprint arXiv:1604.03632*, 2019.

Yukino Baba and Hisashi Kashima. Statistical quality estimation for general crowdsourcing tasks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013.

Von Bakanic, Clark McPhail, and Rita J Simon. The manuscript review and decision-making process. *American Sociological Review*, pages 631–642, 1987.

Stefano Balietti, Robert L Goldstone, and Dirk Helbing. Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 113(30):8414–8419, 2016.

Salem Benferhat and Jerome Lang. Conference paper assignment. *International Journal of Intelligent Systems*, 16(10):1183–1192, 10 2001. ISSN 1098-111X. doi: 10.1002/int.1055.

Katherine Egan Bennett, Reshma Jagsi, and Anthony Zietman. Radiation oncology authors and reviewers prefer double-blind peer review. *Proceedings of the National Academy of Sciences*, 115 (9):E1940–E1940, 2018.

Christophe Bernard. Gender bias in publishing: Double-blind reviewing as a solution? *Eneuro*, 5 (3), 2018.

Rachel Bernstein. PLOS ONE ousts reviewer, editor after sexist peer-review storm. *Science*, 2015.

Rebecca M Blank. The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *The American Economic Review*, pages 1041–1067, 1991.

Nicolas Bousquet, Sergey Norin, and Adrian Vetta. A near-optimal mechanism for impartial selection. In *International Conference on Web and Internet Economics*, pages 133–146. Springer, 2014.

Lyle Brenner, Dale Griffin, and Derek J Koehler. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, 97(1):64–81, 2005.

Amber E. Budden, Tom Tregenza, Lonnie W. Aarssen, Julia Koricheva, Roosa Leimu, and Christopher J. Lortie. Double-blind review favours increased representation of female authors. *Trends in Ecology and Evolution*, 23(1):4 – 6, 2008. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2007.07.008. URL `http://www.sciencedirect.com/science/article/pii/S0169534707002704`.

L. Charlin and R. S. Zemel. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

L. Charlin, R. S. Zemel, and C. Boutilier. A framework for optimizing paper matching. *CoRR*, abs/1202.3706, 2012. URL `http://arxiv.org/abs/1202.3706`.

Kenneth Church. Reviewing the reviewers. *Computational Linguistics*, 31(4):575–578, 2005.

Liran Einav and Leeat Yariv. What's in a surname? the effects of surname initials on academic success. *Journal of Economic Perspectives*, 20(1):175–187, 2006.

Elza Erkip. IEEE information theory society diversity and inclusion committee report. *IEEE Board of Governers Meeting*, July 2019. https://www.itsoc.org/people/bog/bog-meeting-isit-2019-paris-france/DICommittee.pdf.

Cat Ferguson, Adam Marcus, and Ivan Oransky. Publishing: The peer-review scam. *Nature News*, 515(7528):480, 2014.

T Fiez, N Shah, and L Ratliff. A SUPER* algorithm to optimize paper bidding in peer review. In *ICML workshop on Real-world Sequential Decision Making: Reinforcement Learning And Beyond*, 2019.

Felix Fischer and Max Klimm. Optimal impartial selection. *SIAM Journal on Computing*, 44(5): 1263–1285, 2015.

Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J. Zaki. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*, 11(2):63–67, May 2010. ISSN 1931-0145. doi: 10.1145/1809400.1809413. URL http://doi.acm.org/10.1145/1809400.1809413.

Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003. URL http://www.jmlr.org/papers/v4/freund03a.html.

Jian Gao and Tao Zhou. Retractions: Stamp out fake peer review. *Nature*, 546(7656):33–33, 2017.

N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. Assigning papers to referees. *Algorithmica*, 58(1):119–136, Sep 2010. ISSN 1432-0541. doi: 10.1007/s00453-009-9386-0. URL https://doi.org/10.1007/s00453-009-9386-0.

Hong Ge, Max Welling, and Zoubin Ghahramani. A Bayesian model for calibrating conference review scores, 2013. URL http://mlg.eng.cam.ac.uk/hong/nipsrevcal.pdf.

Judy Goldsmith and Robert H. Sloan. The AI conference paper assignment problem. WS-07-10: 53–57, 12 2007.

Dale Griffin and Lyle Brenner. *Perspectives on Probability Judgment Calibration*, chapter 9, pages 177–199. Wiley-Blackwell, 2008. ISBN 9780470752937. doi: 10.1002/9780470752937.ch9. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470752937.ch9.

David Hartvigsen, Jerry C. Wei, and Richard Czuchlewski. The conference paper-reviewer assignment problem. *Decision Sciences*, 30(3):865–876, 1999. ISSN 1540-5915. doi: 10.1111/j.1540-5915.1999.tb00910.x. URL http://dx.doi.org/10.1111/j.1540-5915.1999.tb00910.x.

Anne-Wil Harzing, Joyce Baldueza, Wilhelm Barner-Rasmussen, Cordula Barzantny, Anne Canabal, Anabella Davila, Alvaro Espejo, Rita Ferreira, Axele Giroud, Kathrin Koester, Yung-Kuei Liang, Audra Mockaitis, Michael J. Morley, Barbara Myloni, Joseph O.T. Odusanya, Sharon Leiba O'Sullivan, Ananda Kumar Palaniappan, Paulo Prochno, Srabani Roy Choudhury, Ayse Saka-Helmhout, Sununta Siengthai, Linda Viswat, Ayda Uzuncarsili Soydas, and Lena Zander. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*, 18(4):417–432, 2009.

Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions don't help. *preprint arXiv:1606.08842*, 2016.

Christiana E Hilmer and Michael J Hilmer. How do journal quality, co-authorship, and author order affect agricultural economists' salaries? *American Journal of Agricultural Economics*, 87 (2):509–523, 2005.

Mohammadreza Hojat, Joseph S Gonnella, and Addeane S Caelleigh. Impartial judgment by the gatekeepers of science: fallibility and accountability in the peer review process. *Advances in Health Sciences Education*, 8(1):75–96, 2003.

Ron Holzman and Hervé Moulin. Impartial nominations for a prize. *Econometrica*, 81(1):173–196, 2013.

Anson B Kahng, Yasmine Kotturi, Chinmay Kulkarni, David Kurokawa, and Ariel D. Procaccia. Ranking wily people who rank each other. *Technical Report*, 2017.

Steven Kerr, James Tolliver, and Doretta Petree. Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal*, 20(1): 132–141, 1977.

Ari Kobren, Barna Saha, and Andrew McCallum. Paper matching with local fairness constraints. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

David Kurokawa, Omer Lev, Jamie Morgenstern, and Ariel D Procaccia. Impartial peer review. In *IJCAI*, pages 582–588, 2015.

Michèle Lamont. *How professors think*. Harvard University Press, 2009.

John Langford. Adversarial academia, 2008. URL `http://hunch.net/?p=499`.

John Langford. Bidding problems, 2012a. `https://hunch.net/?p=407` [Online; accessed 6-Jan-2019].

John Langford. ICML acceptance statistics, 2012b. `http://hunch.net/?p=2517` [Online; accessed 14-May-2018].

Carole J Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015.

Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the Association for Information Science and Technology*, 64(1):2–17, 2013. URL `https://pdfs.semanticscholar.org/8731/438cdb50c3242451295ce60acdcdc0812d5c.pdf`.

Xiang Liu, Torsten Suel, and Nasir Memon. A robust model for paper reviewer assignment. In *ACM Conference on Recommender Systems*, 2014.

Cheng Long, Raymond Wong, Yu Peng, and Liangliang Ye. On good and fair paper-reviewer assignment. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 1145–1150, 12 2013. ISBN 978-0-7695-5108-1.

R. S. MacKay, R. Kenna, R. J. Low, and S. Parker. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*, 4(2), 2017. doi: 10.1098/rsos.160760.

Samuel Madden and David DeWitt. Impact of double-blind reviewing on sigmod publication rates. *ACM SIGMOD Record*, 35(2):29–32, 2006.

Michael J Mahoney. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175, 1977.

Alison McCook. Is peer review broken? submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. what's wrong with peer review? *The scientist*, 20(2):26–35, 2006.

Jim McCullough. First comprehensive survey of nsf applicants focuses on their concerns about proposal review. *Science, Technology, & Human Values*, 14(1):78–88, 1989.

Robert K Merton. The Matthew effect in science. *Science*, 159:56–63, 1968.

David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.

Ioannis Mitliagkas, Aditya Gopalan, Constantine Caramanis, and Sriram Vishwanath. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference on Communication, Control, and Computing*, pages 1143–1150, 2011.

Adrian Mulligan, Louise Hall, and Ellen Raphael. Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the Association for Information Science and Technology*, 64(1):132–161, 2013.

Jeffrey Naughton. Dbms research: First 50 years, next 50 years. *Keynote at ICDE*, 2010. `http://pages.cs.wisc.edu/~naughton/naughtonicde.pptx`.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, 2012.

David Nicholas, Anthony Watkinson, Hamid R Jamali, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine. Peer review: still king in the digital age. *Learned Publishing*, 28(1):15–21, 2015.

Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Choosing how to choose papers. *arXiv preprint arxiv:1808.09057*, 2018.

Kanu Okike, Kevin T Hug, Mininder S Kocher, and Seth S Leopold. Single-blind vs double-blind peer review in the setting of author prestige. *Jama*, 316(12):1315–1316, 2016.

S. R. Paul. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34(2):213–223, 1981.

Alan L Porter and Frederick A Rossini. Peer review of interdisciplinary research proposals. *Science, technology, & human values*, 10(3):33–38, 1985.

Simon Price and Peter A Flach. Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3):70–79, 2017.

Debraj Ray and Arthur Robson. Certified random: A new order for coauthorship. *American Economic Review*, 108(2):489–520, 2018.

Martin Reinhart. Peer review of grant applications in biology and medicine. reliability, fairness, and validity. *Scientometrics*, 81(3):789–809, 2009.

Drummond Rennie. Make peer review scientific: thirty years on from the first congress on peer review, drummond rennie reflects on the improvements brought about by research into the process–and calls for more. *Nature*, 535(7610):31–34, 2016.

Marko A. Rodriguez and Johan Bollen. An algorithm to determine peer-reviewers. In *ACM Conference on Information and Knowledge Management*, 2008.

Marko A Rodriguez, Johan Bollen, and Herbert Van de Sompel. Mapping the bid behavior of conference referees. *Journal of Informetrics*, 1(1):68–82, 2007.

Milton Rokeach. The role of values in public opinion research. *Public Opinion Quarterly*, 32(4): 547–559, 1968. doi: 10.1086/267645.

Magnus Roos, Jörg Rothe, and Björn Scheuermann. How to calibrate the scores of biased reviewers by quadratic programming. In *AAAI Conference on Artificial Intelligence*, 2011.

Marco Seeber and Alberto Bacchelli. Does single blind peer review hinder newcomers? *Scientometrics*, 113(1):567–585, 2017.

Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 2018.

Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632*, 2016.

Nihar B. Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory*, 63(2):934–959, 2017.

Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the NIPS 2016 review process. *The Journal of Machine Learning Research*, 19(1):1913–1946, 2018.

Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. In *IEEE Transactions on Information Theory*, pages 4854–4874. IEEE, 2019a.

Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. Low permutation-rank matrices: Structural properties and noisy completion. In *Journal of Machine Learning Research (to appear)*, 2019b.

Stanley S Siegelman. Assassins and zealots: variations in peer review. special report. *Radiology*, 178(3):637–642, 1991.

Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine*, 99(4):178–182, 2006.

Richard Snodgrass. Single-versus double-blind reviewing: an analysis of the literature. *ACM Sigmod Record*, 35(3):8–21, 2006.

Flaminio Squazzoni and Claudio Gandelli. Saint matthew strikes again: An agent-based model of peer review and the scientific community structure. *Journal of Informetrics*, 6(2):265–275, 2012.

Ivan Stelmakh, Nihar Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *arXiv preprint arxiv:1806.06237*, 2018.

Ivan Stelmakh, Nihar Shah, and Aarti Singh. On testing for biases in peer review. In *NeurIPS*, 2019.

Janet Swim, Eugene Borgida, Geoffrey Maruyama, and David G Myers. Joan mckay versus john mckay: Do gender stereotypes bias evaluations? *Psychological Bulletin*, 105(3):409, 1989.

Wenbin Tang, Jie Tang, and Chenhao Tan. Expertise matching via constraint-based optimization. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010.

Warren Thorngate and Wahida Chowdhury. By the numbers: Track record, flawed reviews, journal space, and the fate of talented authors. In *Advances in Social Simulation*, pages 177–188. Springer, 2014.

Stefan Thurner and Rudolf Hanel. Peer-review in a world with rational scientists: Toward selection of the average. *The European Physical Journal B*, 84(4):707–711, 2011.

Andrew Tomkins, Min Zhang, and William D Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

H. D. Tran, G. Cabanac, and G. Hubert. Expert suggestion for conference program committees. In *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, pages 221–232, May 2017. doi: 10.1109/RCIS.2017.7956540.

G David L Travis and Harry M Collins. New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values*, 16(3):322–341, 1991.

Chris R Triggle and David J Triggle. What is the future of peer review? Why is there fraud in science? Is plagiarism out of control? Why do scientists do bad things? Is it all a case of: "All that is necessary for the triumph of evil is that good men do nothing?". *Vascular health and risk management*, 3(1):39, 2007.

Anthony KH Tung. Impact of double blind reviewing on sigmod publication: a more detail analysis. *ACM SIGMOD Record*, 35(3):6–7, 2006.

C Mirjam Van Praag and Bernard MS Van Praag. The benefits of being economics professor a (rather than z). *Economica*, 75(300):782–796, 2008.

Jingyan Wang and Nihar Shah. Theres lots in a name (whereas there shouldn't be). Research on Research blog. https://researchonresearch.blog/2018/11/28/theres-lots-in-a-name/, 2018.

Jingyan Wang and Nihar Shah. Gender distributions of paper awards. Research on Research blog. https://researchonresearch.blog/2019/06/18/gender-distributions-of-paper-awards/, 2019a.

Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *AAMAS*, 2019b.

Mark Ware. Peer review: benefits, perceptions and alternatives. *Publishing Research Consortium*, 4:1–20, 2008.

Thomas J Webb, Bob OHara, and Robert P Freckleton. Does double-blind review benefit female authors? *Trends in Ecology & Evolution*, 23(7):351–353, 2008.

Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. On strategyproof conference review. In *IJCAI*, 2019.