# Editorial: Improving Publication Quality by Reducing Bias with Double-Blind Reviewing and Author Response

Kathryn S. McKinley
The University of Texas at Austin
mckinley@cs.utexas.edu

## 1. Introduction

Because scientific progress depends on peer-reviewing, it behooves researchers to ensure that evaluations are as error free as possible and of the highest possible quality. However, scientists are human and humans have emotions and biases. The reviewing process acknowledges our humanity. In particular, *single-blind* reviewing never reveals the reviewers' identity to the authors, in order to protect reviewers from author retribution. All the conference, journal, and grant processes of which I am aware use at least single-blind reviewing. Some also use *double-blind* reviewing. In addition to not revealing reviewer identities, the authors' identities are not known to the reviewers, for most of the double-blind reviewing process. The purpose of double-blind reviewing is to focus the evaluation process on the quality of the submission by reducing human biases with respect to the authors' reputation, gender, and institution, by not revealing those details.

Compared to single-blind reviewing, every study so far shows double-blind reviewing improves the outcome of the process. Many ACM conferences sponsored by SIGPLAN, SIGARCH, SIGMETRICS, SIGMICRO, and SIGMOD [1, 6], some computer science journals, such as TODS [4], and many journals in other disciplines, such as *The Journal of Finance* [2], successfully use double-blind reviewing.

The SIGPLAN PLDI community significantly prefers double-blind reviewing. In the 2007 PLDI attendee survey, we had 148 responses from 334 attendees (a 44% response rate). Respondents indicated double-blind reviewing was: very useful (40), useful (50), neutral (30), not useful (14), or harmful (4). Only 19% were opposed to it and 60% support it. I recommend that SIGPLAN require all its conferences and journals to use double-blind reviewing for evaluating research submissions, and furthermore that SIGPLAN advocates for an ACM wide policy that requires double-blind reviewing.

In the remainder of this editorial, I point to some of the literature and scientific studies on reviewing, discuss the types of biases that double-blind reviewing helps minimize, and suggest implementation strategies. These strategies include (1) author response, (2) an external review committee (instead of ad hoc external reviews and in addition to the program committee), and (3) to minimize errors, revealing authors before making final decisions, but of course after review submissions and scoring. In *author response*, also called *rebuttal*, reviewers enter their reviews, authors read the reviews, enter a response, and then reviewers make their final decisions. The purpose is to provide authors a forum to correct and directly address issues raised in the reviews. All the above processes seek to improve reviewing quality, minimize some of the objections to double-blind reviewing, reduce errors, and ultimately improve the outcomes of the process.


*SIGNOTICES* August 1, 2008.


## 2. Reducing Bias Improves Quality

A number of scientific studies have examined nepotism and gender bias in the scientific evaluation processes. I summarize three studies here [2, 8, 7] and recommend Snodgrass'06 for a more extensive literature analysis [3].

As a community of scientists, we all benefit if our conferences and journals publish the "best" submitted work, evaluated using established community standards on originality, quality, and methodology. However, if our evaluations are influenced by nepotism, gender, researcher reputation, or institution reputation, the quality of our science is degraded.

Unfortunately, men and women still express systematic bias against women. Consider the 2005 European Young Investigator Awards (EURYI) [7] and Wenneras and Wold's analysis of the 1995 Swedish Medical Research Postdoctoral Fellowship competition for biomedical research [8]. In the first EURYI competition, the European Science Foundation (ESF) awarded 3 of 25 (12%) fellowships to women, although 25% of applicants were women [7]. ESF has not provided further data for analysis, so it is possible the men were better. However, the data for the 1997 Swedish Medical Research (SMR) postdoctoral fellowship is available. Wenneras and Wold forced SMR to provide the data by appealing to the Administrative Court of Appeal in Sweden, which ruled that the information fell under the Freedom of the Press Act [8]. In 1997, there were 114 applicants (62 men and 52 women) for 20 fellowships, which were awarded to 16 men and 4 women. Wenneras and Wold analyze applicant success rates based on gender, publication record, position in author list, quality of publication venue, quality of PhD granting institution, research area, and affiliations with selection committee members.

They found nepotism and gender bias were significant factors in the evaluation process. To be judged as good as their male counter parts, female applicants had to be 2.5 times more productive. For example, if you were a woman, you needed 3 more *Nature* or *Science* articles or 20 more articles in specialized, prestigious journals to be judged equal to a man. Although the SMR prohibited reviewers from evaluating applicants with which they had a conflict, e.g., their own PhD students or students from their institutions, that was insufficient to protect against nepotism. The other committee members systematically scored these applicants higher. For example, if

you had a conflict of interest with a committee member, you accrued an advantage of the equivalent of 3 *Nature* or *Science* articles compared to your peers. If the SMR committee had awarded fellowships without these biases, the quality of fellowship recipients would have improved dramatically.

A widely used metric for research quality and influence is citation count, i.e., the number of other publications that refer to a given paper. Relative to other papers within a discipline, the more highly cited papers are more influential. Laband and Piette use this quality metric to study single and double-blind reviewing outcomes [2]. They examine 28 economic journals which used both single (non-blinded) and double-blind (blinded) reviewing and find:

> *Articles published in journals using blinded peer review were cited significantly more than articles published in journals using non-blinded peer review, controlling for a variety of author, article, and journal attributes.*

They conclude that reviewers are better at applying objective criteria on submission quality with double-blind because the articles accepted under this system have a higher citation rate than articles accepted using single-blind reviewing.

There is every reason to believe these and other results on how human nature affects the outcomes of scientific evaluation hold regardless of the scientific discipline and venue.

## 3. Advice for Authors on Blinding Submissions

Common sense and careful writing can easily preserve anonymity without detracting from the submission. To make your submission double-blind, do not reveal the identity of any author in the text. For example, do not include author names, funding sources, or personal acknowledgments. Do not put your name in the submission document name; do not submit a file called McKinley.pdf whether or not your name is McKinley.

Do not eliminate essential self-references or other references. However, limit self-references only to papers that are relevant for reviewing the submitted paper. Always use the third person when referring to your prior work. For example, if you are Smith write: "We build on the prior work by Jones and Smith [JS 2003]." Do not reference technical reports (or URLs for downloading versions) of your submission, software, or publications. If you must provide supplementary materials, email it to the program chair.

If you have a concurrent related submission, reference it as follows: "Closely related, concurrently submitted work shows how to use this pointer analysis for testing [Anonymous 2007]." with the corresponding citation: "[Anonymous 2007] Under submission. Details omitted for double-blind reviewing." You must send this related submission and submission venue to the program chair. Even following these guidelines, closely building on your own prior work may indirectly reveal your identity. Double blind is not perfect, just better.

## 4. Double-Blind Implementation Issues

In my role as program chair for ASPLOS 2004, PACT 2005, and PLDI 2007, I implemented a double-blind reviewing process. Double-blind reviewing requires more work on the part of the program committee chair and program committee.

***Software support and conflicts.*** The submission software should automate tracking and enforcing conflicts. Authors, committee members, and the program chair must enter conflicts. The software must ensure conflicted committee members never see reviews, rankings, or the reviewers of their conflict submissions. The program chair must also ensure committee members leave the room during discussions of their conflict submissions. I recommend that the submission software requires authors to select all the committee members with whom they have a conflict or no conflict, and to enter in a separate list other institutional and personal conflicts. For example, my conflicts would include: All UT, Steve Blackburn (ANU), Emery Berger (UMass), etc. The committee members should also provide such a list. I do not recommend asking the committee members to select from a list of submitting authors since this list reveals information, especially if the submission pool is small.

At the beginning of the reviewing process, the program chair should remind the reviewers that knowing the authors names and institutions before reading a submission can introduce positive and negative bias. The reviewers current opinions and experiences (or lack thereof) with work from any individual should not influence the evaluation of the current submission. Reviewers should not endeavor to discover the authors, but should endeavor to read any related work needed to determine the novelty of the submission.

***Staged author unblinding.*** I suggest *staged author unblinding* to minimize the impact of human mistakes by authors, reviewers, and the program chair. For example, the corresponding author may forget to enter conflicts for their co-authors. If committee members ask colleagues for additional reviews, they may make mistakes as well. Some mistakes will reveal themselves immediately and occasionally compromise double-blind reviewing. However, other mistakes cannot be revealed without revealing the authors to the reviewers.

After the review is submitted, or after the rebuttal period, or at the committee meeting, authors (and reviewers) should be revealed. The purpose of this step is to expose any conflicts that may have been missed due to human error. No person with a conflict with the submission should see the reviews, reviewer names, ranking, or stay in the room during the discussion of the paper. This element is key to single-blind as well as double-blind reviewing and ensures the privacy of the reviewer to be frank without fear of reprisals. This process introduces biases later, but very few or hopefully no reviewers will then raise or lower scores based solely on this new knowledge. Since the scores are the primary factor in discussion order and impact decision making, this process protects against errors and bias.

It may be the case that the error rate is so low that reducing bias is more important than uncovering errors. My personal experience is that error rates are low, but that there is always at least one person who should not be in the room for a discussion and who gets revealed by exposing the author list at the committee meeting. If we adapted a software system that automated all conflict tracking across the field, I think we could eliminate staged author unblinding, but in the current systems, I think it resolves the conflict between protecting reviewer identities and reducing bias.

***Review committee.*** In addition to program committee reviews, many SIGPLAN conferences obtain ad hoc outside reviewers on a per-submission bases. The goal of the outside review is to generate a thoroughly expert review. With single-blind reviewing, the process of selecting ad hoc reviewers can be distributed among committee members. With double-blind reviewing, the same process is very error prone. For PLDI 2007, PLDI 2008, and ASPLOS 2006, the program chair took on this task, which consumed an enormous amount of time and email bandwidth.

I recommend instead a formal *review committee* to solve this problem, as pioneered at ISMM 2008 by Steve Blackburn. The program chair selects the review committee to complement and extend the expertise of the program committee. The review committee applies the same reviewing standards, but they review fewer papers and do not attend the program committee meeting. The review committee should be sized about the same or slightly larger than the program committee. Whereas a SIGPLAN program committee typically generates three reviews per submission, the review

committee need generate only one review per submission, and thus will read about one third the number of submissions. Since they wont have to travel, are selected, and acknowledged together with the program committee in the proceedings and on the web page (see `http://www.cs.kent.ac.uk/people/staff/rej/ismm2008/`), they should be willing to serve, as was the case for ISMM 2008.

The program chair should use the conference software to apply the same submission assignment process, including conflict of interest procedures, to the review and program committee. Compared to obtaining ad hoc external reviews for each submission, the review committee reduces the chance for errors and eases the burden on the program chair. Hopefully, it also improves reviewer quality because: (1) The review committee is transparent. (2) They are systematically selected to improve breadth and depth of expertise. (3) Since they review more than one paper, they can make relative judgments. (4) They can be made just as accountable by including them in the author response cycle (see below). On the ISMM 2008 survey, 15 thought the review committee was a good idea, 7 had no opinion, and 0 thought it was a bad idea. 32 of 51 ISMM attendees responded. The survey also asked them to justify their choice. Many thought there were "more chances of getting expert reviews." Other comments included: "Expanding the pool helps - it structures the essential extra reviewing." "Hard to tell since I don't know which reviewer was on the PC and which on RC. However, some of my reviews were notable better than others, so maybe the RC worked." "In comparison to other review processes, the feedback was in depth, leading me to believe reviewers had more time." "I didn't submit, but this seems like a wonderful/innovative idea!"

The review committee could also handle program committee submissions in a separate process, addressing the nepotism problem in conferences that allow program committee submission. As far as I know, no SIGPLAN conference has tried this process.

## 5. Objections to Double-Blind

Snodgrass presents a number of objections and frequently asked questions about double-blind reviewing, which I recommend reading [5]. I discuss three objections. Reviewers complain that it eliminates part of the benefit of program committee membership. In my experience, the most objections come from prolific established researchers, who believe it is ineffective to double-blind submissions and/or that it works against prolific authors.

SIGPLAN program committee work usually requires reading 15 to 30 papers. Part of the benefit of this service work is gaining some global knowledge about the field. By removing authors information, the reviewer no longer learns who is doing the best and worst work, although the very best work is hopefully revealed in the conference proceedings. Revealing the authors after rebuttal or during the committee meeting solves this problem.

Some people object to double-blind reviewing because they believe that as reviewers they can identify authors based on the submission, even if authors follow the above guidelines. Research bears out that authors cite themselves more than other authors, and thus established, prolific researchers can often be identified through their citation list [1]. However, the very act of omitting author details on the paper has two distinct benefits. First, it reminds authors that they should endeavor not to reveal themselves through their citations or otherwise. Second, it reminds reviewers that they should judge the paper on its merits rather than based on whomever they guess the authors might be.

Prolific researchers seem to believe that their submissions will suffer the most from this system. The literature analysis by Snodgrass reveals a conflicted reality [3]. Apparently, some reviewers actually hold prolific authors to higher standards or tire of their work, penalizing them. Whereas other reviewers favor prolific authors.

## 6. Author Response

I also recommend an author response phase to the reviewing process. I believe author response has the following benefits: (1) Because authors have the chance to correct reviews, reviewers tend to be more careful and accurate. (2) Reviewers get their reviews finished well in advance of the program committee meeting, which precludes reviewing in a rush on the airplane traveling to the meeting. (3) The program chair has time to obtain additional reviews if no reviewer is an expert or there seems to be a lot of controversy. The program chair should require reviewers to submit about a week or so before the program committee meeting. The authors should have two or three days in which to compose a response to the reviews, which includes answering reviewer questions, addressing concerns and issues raised in the reviews, and correcting any errors. The reviewers should read the responses and adjust their reviews accordingly before the meeting. At the meeting, the reviewer who leads the discussion should summarize the contents of the response.

## 7. Conclusion

Improving the success rate for authors who clearly present original ideas that move science forward in promising directions, use suitable evaluation methodologies, and make appropriate conclusions benefits researchers (prolific or otherwise), science, and the world. Double-blind reviewing improves the quality of decision making by increasing the focus of the evaluation process on the actual submission, rather than the authors. Is double-blind reviewing perfect? No, but double-blind reviewing improves fairness and quality, and all ACM and SIGPLAN conferences and journals should use it.

## References


[1] S. Hill and F. Provost. The myth of the double-blind review?: Author indentification using only citations. *ACM SIGKDD Explorations Newsletter*, 5(2):179–184, 2003.

[2] D. N. Laband and M. J. Piette. Citation analysis of blinded peer review. *The Journal of the American Medical Association (JAMA): The Second International Congress on Peer Review in Biomedical Publication*, 272(2):147–149, July 1994.

[3] R. T. Snodgrass. Single- versus double-blind reviewing: An analysis of the literature. *ACM SIGMOD Record*, 35(3):8–21, 2006.

[4] R. T. Snodgrass. Editorial: Single- versus double-blind reviewing reviewing. *ACM Transactions on Database Systems (TODS)*, 32(1):1–31, 2007.

[5] R. T. Snodgrass. Frequently asked questions about double-blind reviewing. *ACM SIGMOD Record*, 36(1):60–62, 2007.

[6] A. K. H. Tung. Impact of double-blind reviewing on SIGMOD publication: A more detailed analysis. *ACM SIGMOD Record*, 35(3):6–7, 2006.

[7] D. Watson, A. C. Andersen, and J. Hjorth. Mysterious disappearane of female investigators. *Nature*, 436(7048):174, July 2005.

[8] C. Wenneras and A. Wold. Nepotism and sexism in peer-review. *Nature*, 387(6631):341–343, May 1997.


## Biography

Kathryn S. McKinley is a Professor at The University of Texas at Austin. Her advisor was Ken Kennedy and she received her PhD from Rice University in 1992. Her research interests include compilers, runtime systems, programming languages, debugging, and architecture. She and her collaborators have produced a number of tools that are in wide research and industrial use, e.g., DaCapo Java Benchmarks, the TRIPS Scale compiler, the Hoard memory manager, and the MMTk garbage collector toolkit. Her honors include ACM Distinguished Scientist, IBM Faculty Awards, College of Natural Science's Dean's Fellow, and an NSF CAREER Award. She served as the Treasurer/Secretary for SIGPLAN (1999-2001). She was the program chair for PLDI 2007, ASPLOS 2004, and PACT 2005. She is currently co-Editor-in-Chief of ACM Transactions on Programming Language Systems (TOPLAS). She has mentored in the CRA Distributed Mentor Programs for undergraduate and graduate student woman, and co-led with Daniel Jimenez

the CRAW/CDC Programming Languages Summer School, 2007. She was the Director of the Department of Computer Sciences' First Bytes Program, a one week summer camp to introduce high school girls to computer science. She is currently supervising eight PhD students, and has graduated eight PhDs.