DOI:10.1145/3208157

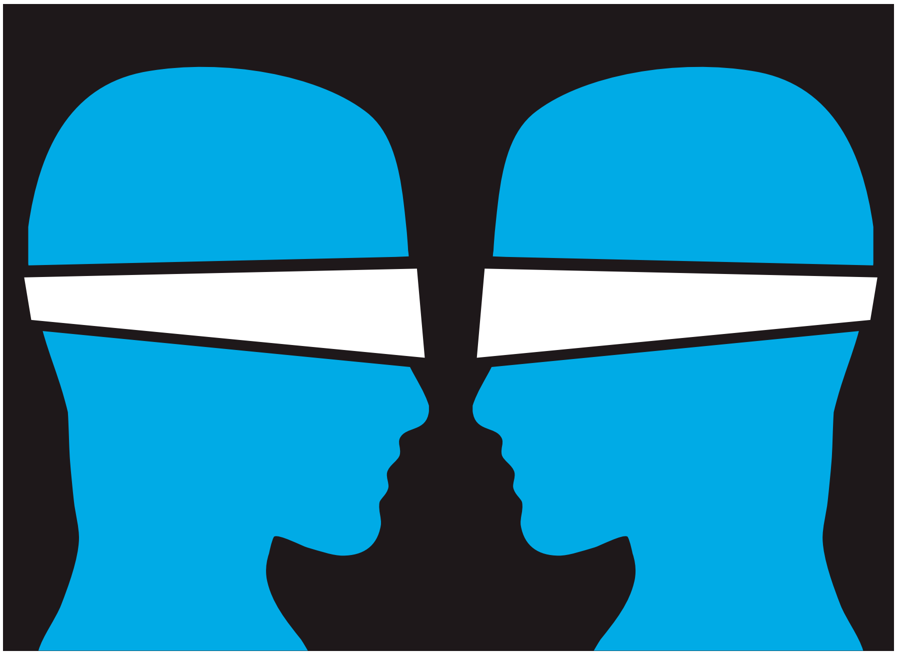
C. Le Goues, Y. Brun, S. Apel, E. Berger, S. Khurshid, and Y. Smaragdakis

Viewpoint
Effectiveness of Anonymization in Double-Blind Review

Assessing the effectiveness of anonymization in the review process.

Peer review is a cornerstone of the academic publication process but can be subject to the flaws of the humans who perform it. Evidence suggests subconscious biases influence one's ability to objectively evaluate work: In a controlled experiment with two disjoint program committees, the ACM International Conference on Web Search and Data Mining (WSDM'17) found that reviewers with author information were 1.76x more likely to recommend acceptance of papers from famous authors, and 1.67x more likely to recommend acceptance of papers from top institutions.[6] A study of three years of the Evolution of Languages conference (2012, 2014, and 2016) found that, when reviewers knew author identities, review scores for papers with male-first authors were 19% higher, and for papers with female-first authors 4% lower.[4] In a medical discipline, U.S. reviewers were more likely to recommend acceptance of papers from U.S.-based institutions.[2]

These biases can affect anyone, regardless of the evaluator's race and gender.[3] Luckily, double-blind review can mitigate these effects[1,2,6] and reduce the perception of bias,[5] making it a constructive step toward a review system that objectively evaluates papers based strictly on the quality of the work.



Three conferences in software engineering and programming languages held in 2016—the IEEE/ACM International Conference on Automated Software Engineering (ASE), the ACM International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA), and the ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)—collected data on anonymization effectiveness, which we[a] use to assess the degree to which reviewers were able to successfully deanonymize the papers' authors. We find that anonymization is imperfect but fairly effective: 70%–86% of the reviews were submitted with no author guesses, and 74%–90% of reviews were submitted with no correct guesses. Reviewers who believe themselves to be experts on a paper's topic were more likely to attempt to guess author identities but no more likely to guess correctly. Overall, we strongly support the continued use of double-blind review, finding the extra administrative effort minimal and well worth the benefits.

[a] Sven Apel and Sarfraz Khurshid were the ASE'16 PC chairs, Claire Le Goues and Yuriy Brun were the ASE'16 review process chairs, Yannis Smaragdakis was the OOPSLA'16 PC chair, and Emery Berger was the PLDI'16 PC chair.

IMAGERY BY LUCKY TEAM STUDIO

Methodology

The authors submitting to ASE 2016, OOPSLA 2016, and PLDI 2016 were instructed to omit author information from the author block and obscure, to the best of their ability, identifying information in the paper. PLDI authors were also instructed not to advertise their work. ASE desk-rejected submissions that listed author information on the first page, but not those that inadvertently revealed such information in the text. Authors of OOPSLA submissions who revealed author identities were instructed to remove the identities, which they did, and no paper was desk-rejected for this reason. PLDI desk-rejected submissions that revealed author identities in any way.

The review forms included optional questions about author identities, the answers to which were only accessible to the PC chairs. The questions asked if the reviewer thought he or she knew the identity of at least one author, and if so, to make a guess and to select what informed the guess. The data considered here refers to the first submitted version of each review. For ASE, author identities were revealed to reviewers immediately after submission of an initial review; for OOPSLA, ahead of the PC meeting; for PLDI, only for accepted papers, after all acceptance decisions were made.

Threats to validity. Reviewers were urged to provide a guess if they thought they knew an author. A lack of a guess could signify not following those instructions. However, this risk is small, for example, OOPSLA PC members were allowed to opt out uniformly and yet 83% of the PC members participated. Asking reviewers if they could guess author identities may have affected their behavior: they may not have thought about it had they not been asked. Data about reviewers' confidence in guesses may affect our conclusions. Reviewers could submit multiple guesses per paper and be considered correct if at least one guess matched, so making many uninformed guesses could be considered correct, but we did not observe this phenomenon. In a form of selection bias, all conferences' review processes were chaired by—and this Viewpoint is written by—researchers who support double-blind review.

Figure 1. Papers, reviews, reviewers, and author guesses. Reviewers include those on the program and external committees, but exclude chairs. All papers received at least three reviews; review load was non-uniform.

	ASE	OOPSLA	PLDI
Reviewers	79	37	111
Papers accepted	71	52	48
Papers rejected	263	144	240
Reviews	**1,029**	**636**	**1,154**
Did not contain a correct author guess	90.2%	74.4%	81.0%
Did not contain an author guess	86.4%	70.0%	74.3%
Tried to guess at least one author	14.7%	30.0%	25.7%
Guessed at least one author correctly	9.8%	25.6%	19.1%
All author guesses incorrect	3.8%	4.4%	6.7%
Reviews with a guess	**140**	**191**	**297**
Guess at least one author correctly	72.1%	85.3%	74.1%
Guess all authors incorrectly	27.9%	14.7%	25.9%
Papers reviewed	**334**	**196**	**288**
No one tried guessing authors	66.5%	41.8%	40.6%
Someone guessed an author correctly	24.6%	50.0%	44.1%
All guesses incorrect	9.0%	8.2%	15.3%

Figure 2. Guess rate, and correct guess rate, by self-reported reviewer expertise score (X: expert, Y: knowledgeable, Z: informed outsider).

	ASE		OOPSLA		PLDI	
	Guess	Correct	Guess	Correct	Guess	Correct
X	19.0%	74.7%	33.6%	86.7%	33.7%	74.2%
Y	11.2%	71.2%	29.3%	84.3%	24.6%	69.0%
Z	7.1%	55.6%	21.2%	83.3%	19.7%	48.6%

Anonymization Effectiveness

For the three conferences, 70%–86% of reviews were submitted without guesses, suggesting that reviewers typically did not believe they knew or were not concerned with who wrote most of the papers they reviewed. Figure 1 summarizes the number of reviewers, papers, and reviews processed by each conference, and the distributions of author identity guesses.

When reviewers did guess, they were more likely to be correct (ASE 72% of guesses were correct, OOPSLA 85%, and PLDI 74%). However, 75% of ASE, 50% of OOPSLA, and 44% of PLDI papers had no reviewers correctly guess even one author, and most reviews contained no correct guess (ASE 90%, OOPSLA 74%, PLDI 81%).

Are experts more likely to guess and guess correctly? All reviews included a self-reported assessment of reviewer expertise (X for expert, Y for knowledgeable, and Z for informed outsider). Figure 2 summarizes guess incidence and guess correctness by reviewer expertise. For each conference, X reviewers were statistically significantly more likely to guess than Y and Z reviewers ($p \leq 0.05$). But the differences in guess correctness were not significant, except the Z reviewers for PLDI were statistically significantly correct less often than the X and Y reviewers ($p \leq 0.05$). We conclude that reviewers who considered themselves experts were more likely to guess author identities, but were no more likely to guess correctly.

Are papers frequently poorly anonymized? One possible reason for deanonymization is poor anonymization. Poorly anonymized papers may have more reviewers guess, and also a higher correct guess rate. Figure 3 shows the distribution of papers by the number of reviewers who attempted to guess the authors. The largest proportion of papers (26%–30%) had only a single reviewer attempt to guess. Fewer papers had more guesses. The bar shading indicates the fractions of the author identity guesses that are correct; papers with more guesses have lower rates of incorrect guesses. Combining the three conferences' data, the χ^2 statistic indicates that the rates of correct guessing for papers with one, two, and three or more guesses are statistically significantly different ($p \leq 0.05$). This comparison is also statistically significant for OOPSLA alone, but not for ASE and PLDI. Comparing guess rates (we use one-tailed z tests for all population proportion comparisons) between paper groups directly: For OOPSLA, the rate of correct guessing is statistically significantly different between one-guess papers and each of the other two paper groups. For PLDI, the same is true between one-guess and three-plus-guess paper groups. This evidence suggests a minority of papers may be easy to unblind. For ASE, only 1.5% of the papers had three or more guesses, while for PLDI, 13% did. However, for PLDI, 40% of all the guesses corresponded to those 13% of the papers, so improving the anonymization of a relatively small number of papers would potentially significantly reduce the number of guesses. Since the three conferences only began using the double-blind review process recently, the occurrences of insufficient anonymization are likely to decrease as authors gain more experience with anonymizing submissions, further increasing double-blind effectiveness.

Are papers with guessed authors more likely to be accepted? We investigated if paper acceptance correlated with either the reviewers' guesses or with correct guesses. Figure 4 shows the acceptance rate for each conference for papers without guesses, with at least one correct guess, and with all incorrect guesses. We observed different behavior at the three conferences: ASE submissions were accepted at statistically the same rate regardless of reviewer guessing behavior. Additional data available for ASE shows that for each review's paper rating (strong accept, weak accept, weak reject, strong reject), there were no statistically significant differences in acceptance rates for submissions with different guessing behavior. OOPSLA and PLDI submissions with no guesses were less likely to be accepted ($p \leq 0.05$) than those with at least one correct guess. PLDI submissions with no guesses were also less likely to be accepted ($p \leq 0.05$) than submissions with all incorrect guesses (for OOPSLA, for the same test, $p = 0.57$). One possible explanation is that OOPSLA and PLDI reviewers were more likely to affiliate work they perceived as of higher quality with known researchers, and thus more willing to guess the authors of submissions they wanted to accept.

Figure 3. Distributions of papers by number of guesses. The bar shading indicates the fraction of the guesses that are correct.

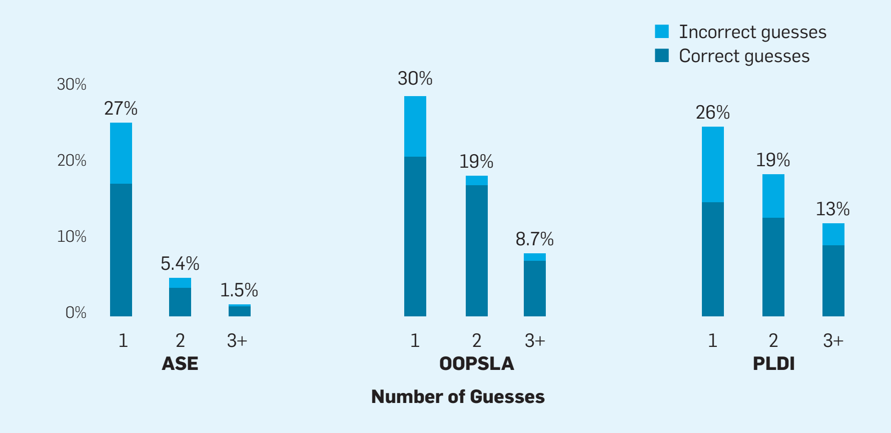


Figure 4. Acceptance rate of papers by reviewer guessing behavior.

Papers with	ASE	OOPSLA	PLDI
No guesses	21.2%	20.7%	6.8%
At least one correct guess	22.0%	31.6%	22.3%
All guesses incorrect	23.0%	25.0%	25.0%
All papers	21.3%	26.5%	16.7%

How do reviewers deanonymize? OOPSLA and PLDI reviewers were asked if the use of citations revealed the authors. Of the reviews with guesses, 37% (11% of all reviews) and 44% (11% of all reviews) said they did, respectively. The ASE reviewers were asked what informed their guesses. The answers were guessing based on paper topic (75 responses); obvious unblinding via reference to previous work, dataset, or source code (31); having previously reviewed or read a draft (21); or having seen a talk (3). The results suggest that some deanonymization may be unavoidable. Some reviewers discovered GitHub repositories or project websites while searching for related work to inform their reviews. Some submissions represented clear extensions of or indicated close familiarity with the authors' prior work. However, there also exist straightforward opportunities to improve anonymization. For example, community familiarity with anonymization, consistent norms, and clear guidelines could address the incidence of direct unblinding. However, multiple times at the PC meetings, the PC chairs heard a PC member remark about having been sure another PC member was a paper author, but being wrong. Reviewers may be overconfident, and sometimes wrong, when they think they know an author through indirect unblinding.

PC Chairs' Observations

After completing the process, the PC chairs of all three conferences reflected on the successes and challenges of double-blind review. All PC chairs were strongly supportive of continuing to use double-blind review in the future. All felt that double-blind review mitigated effects of (subconscious) bias, which is the primary goal of using double-blind review. Some PC members also felt so, indicating anecdotally that they were more confident their reviews and decisions had less bias. One PC member remarked that double-blind review is liberating, since it allows for evaluation without concern about the impact on the careers of people they know personally.

All PC chairs have arguments in support of their respective decisions on the timing of revealing the authors (that is, after review submission, before PC meeting, or only for accepted papers). The PLDI PC chair advocated strongly for full double-blind, which enables rejected papers to be anonymously resubmitted to other double-blind venues with common reviewers, addressing one cause of deanonymization. The ASE PC chairs observed that in a couple of cases, revealing author identities helped to better understand a paper's contribution and value. The PLDI PC chair revealed author identities on request, when deemed absolutely necessary to assess the paper. This happened extremely rarely, and could provide the benefit observed by the ASE PC chairs without sacrificing other benefits. That said, one PC member remarked that one benefit of serving on a PC is learning who is working on what; full anonymization eliminates learning the who, though still allows learning the what.

Overall, none of the PC chairs felt the extra administrative burden imposed by double-blind review was large. The ASE PC chairs recruited two review process chairs to assist, and all felt the effort required was reasonable. The OOPSLA PC chair noted the level of effort required to implement double-blind review, including the management of conflicts of interest, was not high. He observed that it was critical to provide clear guidance to the authors on how to anonymize papers (for example, http://2016.splashcon.org/track/splash-2016-oopsla#FAQ-on-Double-Blind-Reviewing). PLDI allowed authors to either anonymize artifacts (such as source code) or to submit non-anonymized versions to the PC chair, who distributed to reviewers when appropriate, on demand. The PC chair reported this presented only a trivial additional administrative burden.

All PC chairs were strongly supportive of continuing to use double-blind review in the future.

The primary source of additional administration in double-blind review is conflict of interest management. This task is simplified by conference management software that straightforwardly allows authors and reviewers to declare conflicts based on names and affiliations, and chairs to quickly cross-check declared conflicts. ASE PC chairs worked with the CyberChairPro maintainer to support this task. Neither ASE nor OOPSLA observed unanticipated conflicts discovered when author identities were revealed. The PLDI PC chair managed conflicts of interest more creatively, creating a script that validated author-declared conflicts by emailing PC members lists of potentially conflicted authors mixed with a random selection of other authors, and asking the PC member to identify conflicts. The PC chair examined asymmetrically declared conflicts and contacted authors regarding their reasoning. This identified erroneous conflicts in rare instances. None of the PC chairs found identifying conflicts overly burdensome. The PLDI PC chair reiterated that the burden of full double-blind reviewing is well worth maintaining the process integrity throughout the entire process, and for future resubmissions.

Conclusion

Data from ASE 2016, OOPSLA 2016, and PLDI 2016 suggest that, while anonymization is imperfect, it is fairly effective. The PC chairs of all three conferences strongly support the continued use of double-blind review, find it effective at mitigating (both conscious and subconscious) bias in reviewing, and judge the extra administrative burden to be relatively minor and well worth the benefits. Technological advances and the now-developed author instructions reduce the burden. Having a dedicated organizational position to support double-blind review can also help. The ASE and OOPSLA PC chairs point out some benefits of revealing author identities midprocess, while the PLDI PC chair argues some of those benefits can be preserved in a full double-blind review process that only reveals the author identities of accepted papers, while providing significant additional benefits, such as mitigating bias throughout the entire process and preserving author anonymity for rejected paper resubmissions. C

References

1. Budden, et al. Double-blind review favours increased representation of female authors. *Trends in Ecology and Evolution 23*, 1 (Jan. 2008), 4–6.
2. Gastroenterology, Bethesda, MD, USA. U.S. and non-U.S. submissions: An analysis of reviewer bias. *JAMA 280*, 3 (July 1998), 246–247.
3. Moss-Racusin, C.A. et al. Science faculty's subtle gender biases favor male students. *PNAS 109*, 41 (Apr. 2014), 16474–16479.
4. Roberts, S.G. and Verhoef, T. Double-blind reviewing at EvoLang 11 reveals gender bias. *J. of Language Evolution 1*, 2 (Feb. 2016), 163–167.
5. Snodgrass, R. Single-versus double-blind reviewing: An analysis of the literature. *SIGMOD Record 35*, 3 (May 2006), 8–21.
6. Tomkins, A., Zhang, M., and Heavlin, W.D. Single versus double-blind reviewing at WSDM 2017. CoRR, abs/1702.00502, 2017.

Claire Le Goues (clegoues@cs.cmu.edu) is an Assistant Professor in the School of Computer Science at Carnegie Mellon University, Pittsburgh, PA, USA.

Yuriy Brun (brun@cs.umass.edu) is an Associate Professor in the College of Information and Computer Sciences at the University of Massachusetts Amherst, Amherst, MA, USA.

Sven Apel (apel@uni-passau.de) is a Professor of Computer Science and Chair of Software Engineering in the Department of Informatics and Mathematics at the University of Passau.

Emery Berger (emery@cs.umass.edu) is a Professor in the College of Information and Computer Sciences at the University of Massachusetts Amherst, Amherst, MA, USA.

Sarfraz Khurshid (khurshid@ece.utexas.edu) is a Professor in Electrical and Computer Engineering at the University of Texas at Austin, Austin, TX, USA.

Yannis Smaragdakis (smaragd@di.uoa.gr) is a Professor in the Department of Informatics at the University of Athens.

Kathryn McKinley suggested an early draft of the reviewer questions used by OOPSLA and PLDI. This work is supported in part by the National Science Foundation under grants CCF-1319688, CCF-1453474, CCF-1563797, CCF-1564162, and CNS-1239498, and by the German Research Foundation (AP 206/6).