Delft University of Technology
Software Engineering Research Group
Technical Report Series

Double Blind Reviews in SE conferences: Practicability, Promises, and Perils

Moritz Beller, Alberto Bacchelli

Report TUD-SERG-2015-007

TUD-SERG-2015-007

Published, produced and distributed by:

Software Engineering Research Group
Department of Software Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Mekelweg 4
2628 CD Delft
The Netherlands

ISSN 1872-5392

Software Engineering Research Group Technical Reports:
`http://www.se.ewi.tudelft.nl/techreports/`

For more information about the Software Engineering Research Group:
`http://www.se.ewi.tudelft.nl/`

This paper is a tech-report of a work-in-progress paper.

Double Blind Reviews in SE conferences: Practicability, Promises and Perils

Moritz Beller, Alberto Bacchelli

Delft University of Technology,
The Netherlands
{m.m.beller, a.bacchelli}@tudelft.nl

ABSTRACT

The peer review process is central to the scientific method, the advancement and spread of research as well as crucial for individual careers. As such, a single blind review process is susceptible towards apparent and hidden biases in the reviewers, as they know the identity of the authors. In this paper, we perform an extensive study on the benefits and costs that would be associated with introducing double-blind reviews at large SE conferences. We do this by surveying the SE community's opinion, interviewing experts on double-blind reviewing and estimating the likelihood of reviewers being able to guess the author's names. Our results indicate that double-blind reviewing could be introduced in large SE conferences at lower-than-generally believed costs and that the majority of the SE community would be in favour of introducing it.

Categories and Subject Descriptors

D.2.5 [**Software Engineering**]

General Terms

Experimentation, Human Factors

Keywords

Reviewing, Single-blind Reviews, Double-Blind Reviews

1. INTRODUCTION

Peer review, *i.e.*, the practice that scientists use to evaluate research manuscripts and artifacts, is strongly connected to the advancement of scientific knowledge and researchers' careers [1]. For this reason, it is vital for scientific communities to continuously reflect on matters related to the quality this practice.

One of the most recurring topics of debate with respect to review quality, across many scientific disciplines, is the practice of double-blind review (DBR) as opposed to single-blind review (SBR). In SBR the reviewers are anonymous to the authors, but the authorship information is visible to the reviewers; in DBR, the authorship information is not visible. In principle, the arguments in favor of DBR are that reviewers are not influenced since they do not know any information about the authors of a manuscript, thus should produce fairer and higher quality reviews [2]; while arguments against are that it hardens the writing and reviewing of manuscripts [2].

In the software engineering (SE) community, the traditional choice for most conferences and journals is to employ SBR and ICSE (the International Conference on Software Engineering), considered the flagship conference, makes no exception.

However, a number of top conferences in computer science previously SBR, the related community of programming languages, and a number of conferences in SE subfields (*e.g.*, ISSTA and FASE) have switched, or are in the process of switching to some form of DBR. This trend raises the question whether the switch should also take place with other venues in SE and, in particular, with ICSE.

Deciding on such a switch is not trivial. Although previous work has demonstrated opportunities for bias due to the reviewers being able to clearly see who authored a submission, there is contrasting evidence on whether DBR has a significant impact in practice (see [2] for an extensive literature survey). Moreover, results found for other domains or venues might not be directly transferred to the general SE domain and ICSE, due to differences in: size of analyzed venues, artifacts produced, type and style of research, and (potential) perception and behavior of the community. In addition, most work on DBR effectiveness has been conducted on journals, in which a substantially different reviewing process takes place and Editor-in-Chief and associate editors serve for multiple years (as opposed to program chairs, who serve mostly for one year).

To better inform this decision and to make the SE community aware of issues around this topic, we conduct a study to reflect on double-blind in the context of SE research and ICSE in particular.

We set up our study as an exploratory investigation. We started without a priori hypotheses regarding whether and how DBR should be performed, with the aim of discovering the most important aspects to investigate. We started by surveying related literature and conferences that switched, and by interviewing 14 prominent members of the ICSE community about their perception on DBR and ICSE. From these, the overarching research question of our study emerged: *Are the benefits of DBR worth the costs?* We further refined it in more structured research questions and to answer these, we:(1) further analyzed interview data (2) interviewed 5 experts from DBR communities (3) created a tool to try to automatically infer authors from papers based on referenced citations and used it to try to guess authors of the papers in the ICSE proceedings of the last five years, (4) surveyed 282 researchers, 242 of which having SE as their main field.

Our results confirm that benefits of DBR are mostly related to increased fairness due to eliminating authorship visibility and its influence on reviewers' judgment. According to our participants, such influence can be seen as early as the bidding process (during which various participants reported to have been influenced in the choice of the papers they specified to review) and even during online and physical program committee discussions. Most survey respondents agree that the main benefit of DBR, in addition to reducing reviewers' bias, is an increase in the reputation of the conference switching to DBR. The costs (logistic challenges and side effects) of DBR outnumber the benefits and mostly regard difficulty for authors in blinding papers, for reviewers in understanding the increment with respect to previous work from the same authors, and for organizers to manage a very complex transition. Despite most participants agree with most costs of DBR, only less than one third of the respondents disagree with a switch to DBR for SE journals, all SE conferences, and, in particular, ICSE.

2. BACKGROUND

In this section, we first provide an extensive overview over literature on single- and double-blind reviewing. We conclude with an analysis of the state of practice of double reviewing.

2.1 Literature

While this paper constitutes the first study on double blind reviewing within the software engineering research domain, there is a large body of research on reviewing practices and biases in other fields.

Snodgrass provides an extensive overview on DBR in the context of the ACM SIGMOD conference, a premier forum for database research, which has introduced DBR in 2001 [2, 3]. He argues that the main benefit of DBR is increased fairness and groups it into actual fairness, i.e. an evaluation irrespective of personal relation, affiliation, popularity, gender, or seniority, and perceived fairness, i.e. a larger confidence of the community in the review process. Conversely, he lists several general costs of DBR, which we used as a basis for our software engineering-specific costs. He gives a balanced summary of previous studies demonstrating both beneficial and adverse effects of DBR on review quality, suggesting that quality of reviews might stay similar. Snodgrass describes several studies on the efficacy of blinding authors, demonstrating that even a light-weight blinding can successfully disguise the majority of authors from reviewers. His survey of the recommendations of scholarly societies shows that many suggest at least an optional DBR process, should authors so wish, and that DBR use has increased significantly. He concludes that DBR is still more prevalent in the social sciences than in computer science, despite its beneficial effect and the assumed low costs for a transition. As a result, Snodgrass maintains a document of frequently asked questions and answers regarding DBR [4].

A crucial factor for the success of DBR is that author identities are not too easy to infer. In the sub-field of particle physics, Hill and Provost could automatically identify authors 25% to 45% of the time [5]. We used an advanced machine-learning technique on a large corpus of SBR and DBR software engineering papers, revealing an author identity in only 20% of papers.

Some fields have conducted experiments and case studies with DBR [6]. Overall, the acceptance rate decreased, mainly affecting papers from near-top universities, leaving the rates for papers from both top universities and low-ranked universities unaffected. No significant effect was measured on the gender of authors. A similar study in the field of medicine found no effect on review quality or outcome [7]. Budden et al. showed that, after a venue introduced DBR, female authorship increased [8]. However, this was also true for other venues in the field and time period, which still employed SBR [9].

Outside the world of academic paper reviewing, both intentional and unintentional, conscious and unconscious, racial, gender and other biases have been extensively studied [10–12] and shown to exist, even in judges and physicians who reported they were unbiased [13, 14]. As two such examples, Rouse and Goldin found that when American symphony orchestras switched to blind auditions, the probability for a woman to advance to the next selection round increased by 50 percent [15]. Steinpreis et al. randomized names on otherwise identical academic resumes and found that supposedly-male applicants were hired more often than supposedly-female applicants [16]. We therefore conclude that, while the few studies on DBR report contrary results, we need more research into the effects of DBR and that it seems unreasonable to assume academic reviewing in particular to be free of hidden biases.

2.2 Practice

Having gained an understanding of the general scientific evidence on double-blind reviewing, we compare the state of the practice of reviewing in the Software Engineering community to other Computer Science fields. We also report anecdotal expieriences that venues have made in switching to double-blind reviewing.

To assess where Software Engineering stands in comparison to other sub-fields of Computer Science, we first established which sub-fields are present in Computer Science, and what their top-tier venues, i.e. journals and conferences, are. We used the 15 Computer Science sub-fields suggested by Google Scholar [17] and selected the top venues based on their h5-indices: We considered a venue to be *top-tier,* if its h-index was $\geq 90\%$ of the highest index in this sub-field. To find out which reviewing mode a venue employs, we first studied the available call for papers and the front letters of previous proceedings, if they were accessible online. If these were not available or gave no clear indication of the reviewing mode and when a potential change to it, we contacted the the ex-program chairs or editors in chief of the prior editions of the venue.

Table 1 shows the 16 top-tier venues of the 14 sub-fields of Computer Science other than Software Engineering. Five of them apply DBR, two are optional depending on the authors' wishes and the remaining seven are single-blind. Venues typically switch to a double-blind review process during their evolution and do not revert back to SBR. Keith Price, program chair of CVPR 1985, summarized that "in all the debates about the [review] process, the number of papers selected was the issue, not whether double blind was good or bad (it was accepted as workable and good)."

Following this classification, in the field of Software Engineering, both the International Conference on Software Engineering (ICSE, h-index: 56) and the journal Transactions on Software Engineering (TSE, h-index: 52) are top-tier venues that notoriously do not employ double-blind reviewing. Some non-top-tier venues in Software Engineering recently switched to DBR, including the Symposium on Search Based Software Engineering (SBSSE) 2014 [18], the International Symposium on Software Testing and Analysis (ISSTA) 2016 [19], and the International Conference on Fundamental Approaches to Software Engineering (FASE) 2016 [20]. Contrary to this trend to switch to double-blind reviewing, the journal Empirical Software Engineering (EMSE) switched from DBR to SBR. Lionel Briand, EMSE's co-editor in chief since 2003, told us the reasons for this unique decision include that articles in EMSE are often extensions of previously published conference papers. To assess the delta to the conference paper, some reviewers of the conference paper are often also the reviewers of the extension. In these circumstances, DBR was not perceived as cost-effective.

Table 1: Review Mode of Top-Tier Computer Science Venues.

Sub-Field	Venue	Has DBR?	DBR since
Artifical Intelligence	Expert Systems with Applications	No	
Computational Linguistics	Meeting of the Association for Computational Linguistics	Yes	1993
Computer Graphics	Transactions on Graphics	No	
Computer Hardware Design	Journal of Solid-State Circuits	No	
Computer Networks	Communications Magazine	No	
Computer Security & Cryptography	Symposium on Security and Privacy	Yes	2002
Computer Security & Cryptography	Symposium on Information, Computer and Communications Security	Yes	2010
Computer Security & Cryptography	Transactions on Information Forensics and Security	No	
Computer Vision & Pattern Recognition	Conference on Computer Vision and Pattern Recognition	Yes	1985
Computing Systems	Transactions on Parallel and Distributed Systems	No	
Databases & Information Systems	International World Wide Web Conferences	No	
Databases & Information Systems	International Conference on Very Large Databases	No	
Human Computer Interaction	Computer Human Interaction	Yes	2004
Medical Informatics	Journal of Medical Internet Research	Opt.	
Medical Informatics	Journal of the American Medical Informatics Association	No	
Robotics	International Conference on Robotics and Automation	No	
Signal Processing	Transactions on Signal Processing	No	
Signal Processing	Transactions on Image Processing	Opt.	
Theoretical Computer Science	Symposium on Theory of Computing	No	

3. METHODOLOGY

We define the scope of our research, the data sources we use, and research questions and corresponding methodology.

3.1 Scoping

To scope our initial focus we tapped in the knowledge of experts, with different levels of seniority, from the ICSE community engaged in organizing the conference, participating in the program committee, and authoring papers themselves. To do so, we conducted a set of interviews with 13 researchers (described in Table 2) in the ICSE community. This allowed us not only to gather rich data for our study, but also to determine (and iteratively refine) the most compelling research questions to investigate.

The overarching theme emerged from the analysis of the interviews is the existence of an unclear trade-off between costs and benefits of switching to DBR. As one expert put it: "*in principle, double-blind review is a very good idea, who can disagree?* [...] [But] *given the additional overhead and cost, caused by the practical problems,* [DBR] *is only worth it if it has a large impact.*" [I2] With our study we aim at informing about this trade-off.

Table 2: Interviewed researchers

ID	Academic Rank	Community Service		h-index	sex
		ICSE PC	Steering C.		
I1	Full	✓		≥ 40	m
I2	Full	✓	✓(ICSE)	≥ 40	m
I3	Associate	✓		< 20	f
I4	Assistant	✓		20 – 40	m
I5	Full	✓		≥ 40	m
I6	Full	✓	✓(ICSE)	≥ 40	m
I7	Full	✓		≥ 40	m
I8	Full	✓	✓(ICSE)	20 – 40	f
I9	Full	✓	✓(ICSE)	≥ 40	f
I10	Associate	✓		< 20	m
I11	Assistant			< 20	m
I12	Associate	✓		≥ 40	m
I13	Associate			< 20	m
DB1	Full			20 – 40	m
DB2	Assistant			< 20	f
DB3	Full			20 – 40	f
DB4	Full		✓(PL conf.)	20 – 40	f
DB5	Full		✓(PL conf.)	≥ 40	f

3.2 Data sources

To investigate costs and benefits of a transition of ICSE to double-blind, we follow a mixed qualitative and quantitative approach [21], collecting and analyzing data from different sources for both triangulation and investigating different aspects relevant to our study: (1) a review of double-blind related literature, (2) an analysis of double-blind conferences, (3) 13 interviews with ICSE community members, [I1–13], (4) 5 interviews with members of communities employing DBR, [DB1–5], (5) a card sort on interview data and subsequent affinity diagramming, (6) an online survey to the SE community (particularly ICSE authors) with 281 respondents, and (7) a quantitative analysis of the relationship between authorship and cited references in technical papers in the ICSE and ASIACCS proceedings from 2010 to 2014.

Before diving into the methodological details (Section 3.4) of these steps, we outline our research questions and how the aforementioned sources and analyses concur to their answers. We refer to specific questions in our survey (publicly available [22]) using a [22.QX] notation, where X is the question ID.

3.3 Research questions and methods

We structure our investigation around three main research questions, organized in several sub-questions, for which we describe rationale and research method.

RQ1: What are the benefits of double-blind review?

Giving a definitive and comprehensive answer to this question has challenged researchers for several years. Our goal is to investigate it by looking at different aspects and relate it to the SE and ICSE community.

RQ1.1: How and in which stages of the review process could authorship visibility influence SE reviewers?
Rationale: The fundamental argument in favor of DBR is that it is fairer to the authors and the scientific progress, as the reviewers will judge a manuscript only on its scientific value without being in-

fluenced by extenuating circumstances (*e.g.*, the sex of the authors or their affiliations). In other words, the main expected benefit of DBR is that it eliminates the biases derived from knowing the authorship of a paper. Pinpointing the biases that could influence a reviewer in SBR is the first step in investigating on the value of DBR. In addition, even though authorship visibility may induce biases in the reviewers, their effect on the reviews might be mitigated by a number of factors (*e.g.*, submissions are evaluated by multiple reviewers who may have conflicting biases, resulting in a "fair" overall evaluation), thus resulting in a negligible impact in practice. For this reason, we also analyze in which steps of the process reviewers may be more visibly influenced by authorship information.
Research method: To answer this question, we first compile a list of biases that can potentially influence reviewers. To do so, we collect biases shown to potentially influence reviewers in other fields by analyzing double-blind related literature; then we discuss some of these during interviews and extract and group all the biases mentioned in our cards from [I1–13] to compile a list; finally we ask survey respondents how much, from their experience, they perceive that SE reviewers can be influenced by the listed biases (also allowing respondents to add any missing bias) [22.Q16] and to rank them by importance [22.Q17]. Subsequently, we investigate in which stages of the review process the biases may be stronger and more visible. We analyze the cards from [I1–13] to define the stages and highlight the potential influence of authorship visibility in there. Then, we complete it by explicitly ask survey respondents where they think influences of biases can be stronger for SE reviewers [22.Q18] and, from those with reviewer experience, in which stages they experience that they (or others) may have been influenced during the review process [22.Q20].

RQ1.2: Can DBR bring benefits not related to increased fairness?
Rationale: Previous literature reports potential benefits ascribed to DBR in addition to increased fairness [2]. To have a more comprehensive picture of DBR, we investigate them in our context.
Research method: We answer this research question compiling a comprehensive list of potential additional benefits, not related to fairness, from literature. Then, we add benefits addressed on our cards from [I1–13] and [DB1–5] and merge them with the list from literature. Finally, we ask survey respondents how much they agree that these benefits derive from DBR (across different questions [22.Q22–24]), with space to add any missing ones.

RQ2: What are the costs of double-blind review?

Similarly to RQ1, we split our analysis to reflect on this question in a number of sub research questions that investigate different aspects related to the costs, challenges, or drawbacks of DBR.

RQ2.1: How easy is it to guess the authorship of a manuscript?
Rationale: "Any benefits ascribed to double-blind reviewing assume that the blinding of the submitted manuscript has been successful" [2]. The most problematic drawback of DBR would be easily discovering the characterizing author(s) of a manuscript, despite the mechanisms put in place to achieve anonymity.
Research method: We analyze this aspect by from two angles. On the one hand we ask survey respondents how frequently they believe they could correctly guess at least one characterizing author (*e.g.*, a senior author or a researcher known for a certain topic) of a blinded paper, with a 6-level Likert-scale [22.Q21] and how much they agree with the sentence "Reviewers will not be affected by double-blind as they can deduce authors" [22.Q24]. On the other hand, we build a tool, AUTHORINFERENCER (Section 3.4), to check whether and how correctly one could take an ICSE paper and identify its authors using only the citations it includes. We use it on the papers published in the technical research track of ICSE from 2010 to 2014 and compare the results with papers published in the technical research track of ASIACCS (Symposium on Information, Computer and Communications Security, see Table 1) in the same years. This way, we are able to give an indication whether it is easier to guess the authors of single-blind or double-blind papers, and whether there exist substantial differences in the way in which both refer to related work.

RQ2.2: What are the challenges of switching to DBR?
Rationale: The transition to DBR requires to handle a number of steps and changes to various practices for organizers, reviewers, and authors. Moreover, in addition to clear steps that have to be completed when switching and managing a DBR conference, other unintended side-effects can raise the costs of a switch decision. Pinpointing the challenges that have to be handled in the transition and when DBR is in place is key in reflecting on the value of DBR.
Research method: To ensure our list of costs (challenges and drawbacks) is complete, we start our investigation with a literature study on costs of DBR [2]; then, we extract and group costs addressed on our cards from [I1–13] and merge them with our set of costs from literature. As experts on DB are more aware of the actual costs, we triage our preliminary set of costs with the answers from [DB1–5]. Then, we merge highly related costs. Similarly to the analysis of potential benefits of DBR, we complete our set of costs by explicitly asking survey respondents how much they agree that these costs derive from DBR (across different questions [22.Q22–24]), with space to add any missing ones.

RQ3: What is the opinion of the community on DBR?

We investigate this aspect answering two research questions.

RQ3.1: How does the ICSE community perceive DBR?
Rationale: We aim to understand which kind of value the SE community, particularly the ICSE one, gives to the topic of DBR. Emerging from the analysis of cards from [I1-13], one of the additional potential benefits of adopting DBR is an increased perception of the scientific value (due to increased fairness) of the conference that switches, regardless of whether the other benefits have a significant tangible effect. We investigate whether this would be the case for the SE community, particularly the ICSE one.
Research method: The answer to this question is captured from a number of survey questions, which in some cases we also use to answer other questions (*e.g.*, [22.Q16]). For example, we ask respondents whether they have ever though if one of their paper was accepted/rejected due to authorship visibility [22.Q14,15], what the strength of reviewers' biases may be [22.Q16], how much the final score and decision of a review may be influenced by authorship [22.Q18], whether they experienced biases in the role of reviewer [22.Q20], and consequences of a switch to DBR [22.Q22–24]. Finally, we explicitly ask whether they would like ICSE to DBR [22.Q34,38], as well as other SE conferences [22.Q37,41] and SE journals [22.Q36,40].

RQ3.2: Up to which costs is the community willing to pay for DBR?
Rationale: The cost of logistical challenges related to DBR are mostly to be paid in additional time for running the process. These can be one-time costs for the transition or repeated costs to keep the DBR mechanism working. From the interviews to ICSE members, the notion that program chairs would have to pay the highest costs of DBR emerges. This is not confirmed by the experts on DB (cards from [D1–5]), rather they report time costs for DBR to be shared among all community members, mostly reviewers and authors. We investigate up to which time costs the ICSE community is willing to invest as authors and reviewers to tackle the challenges of DBR.

Research method: We investigate this question by explicitly asking survey respondents whether they would be willing to invest time as authors [22.Q26] and as reviewers [22.Q30] to make DBR review possible. If not, we ask the reason, otherwise we additionally ask how much time the would devote to additional (*e.g.*, learning how to write/review a DBR paper [22.Q28,32]) or more expensive (*e.g.*, declaring conflicts of interest [22.Q33]) tasks.

3.4 Research method details

Having gained an understanding of the complete picture of the research questions and methods used to answer them, we zoom-in on the methodological details.

Interviews with ICSE and DBR experts. We first conducted a series of interviews with experts from the ICSE community each taking 25-60 minutes (average 36). We contacted people from the ICSE community who have served in the steering, program, and/or organizing committee, and who possibly had experience as program chair. To increase chances that people would be available for the interview, we contacted people we knew through our professional networks and possibly expressed firm views on DBR in the past. We started interviewing a small set of people and expanded progressively as more findings emerged, until—with 13 interviews—we reached a *saturation* point [23]: New interviewees were providing insights very similar to the earlier ones.

Subsequently, we interviewed experts from communities employing DBR, each for 35-45 minutes (average 40). In this case, we selected people that had contributed to the switch of conference(s) to DBR, moved from SBR communities to ones already using DBR for several years, and/or had extensive experience with publishing in DBR-only communities.

Each meeting was in the form of a *semi-structured* interview [24]. This form of interviews makes use of an *interview guide* that contains general groupings of topics and questions rather than a predetermined exact set and order of questions. They are often used in an exploratory context to "find out what is happening *[and]* to seek new insights" [25]. The guideline was iteratively refined after each interview, in particular when we were receiving very similar answers. We conducted most interviews (15) online. With consent, assuring the participants of anonymity, we recorded and transcribed the audio. Afterwards, we analyzed the transcripts and split them into smaller *coherent units* (*i.e.*, blocks expressing a single concept), for subsequent analysis.

Card sort on interviews. To analyze our interview data, we created 811 cards from the transcribed coherent units. Each card included: the context (*e.g.*, last question asked by the interviewer), the interviewee's name, the unit content, and an ID for later reference. Two authors together did a *card sort* [26] to extract salient themes. Card sorting is a sorting technique that is widely used in information architecture to create mental models and derive taxonomies from input data. In practice, card sort participants read each card and progressively sort them into meaningful groups with a descriptive title. After macro categories were discovered, we reanalyzed their cards to obtain a finer-grained categorization. Finally, we organized the categories using *affinity diagramming* [27], a technique that allows large numbers of ideas to be sorted into groups for review and analysis [28]. We used it to generate an overview of the topics that emerged from the card sort to connect the related concepts and derive the main themes.

Survey. To validate, extend, and put in the context of the whole SE and ICSE community the concepts that emerged from the previous phases, we created an online survey [22]. For the design of the survey, we followed Patten's guidebook on questionnaire research [29] and Kitchenham and Pfleeger's guidelines for personal opinion surveys [30]. The survey was anonymous and offered the chance to enter a raffle for a 50 Euro gift to increase response rate [31].

To verify clarity and appropriateness of our questions, measure the time to complete the survey, and discuss redundant or missing elements, we run a pilot survey with 10 respondents from our target population. Whenever reasonable, we integrated their feedback.

The final survey comprised 41 questions, mostly closed with multiple choice answers, grouped in 10 pages and was shared with the target population in two phases. In the first phase, we advertised the survey through research collaborations via personal emails, Twitter, and Facebook (particularly on the group 'Software Engineering Research Community' with more than 4.000 members). In the second phase, to receive a maximally unbiased list of participants to our survey that best represents the general ICSE community's opinion on DBR, we sent an email invitation to participate in our survey to authors of previous ICSE papers. In particular, we used the AUTHORINFERENCER tool to extract the email addresses of authors of full research papers from ICSE 2014 to 2010 proceedings. After data cleaning, removal of duplicates and people we already contacted, we sent 848 personal invitation emails to complete the survey. We received 147 responses stating that the message could not be delivered. From the remaining 701, 122 recipients (17.4%, typical response rate of online surveys in software engineering [32]) completed the survey from the email link. In total, we collected with 282 complete responses, while we discarded from the analysis 163 responses in which the respondent did not reach the 'submit' page.

Survey respondents. The 282 participants in our survey were diverse in their academic position with 21% PhD students, 14% postdoctoral researchers and the three different professor levels (assistant, associate, and full) accounting for 20%, 16%, 21% of responses. Of the respondents, 18% reported to (also) work in industry. In total, people from 31 countries responded, with most working in the US (26%). 29% of respondents are native English speakers. 69% of participants were Caucasian and 84% was male.

Inferring paper authors. We developed AUTHORINFERENCER [33] as an objective benchmark to showcase how easy or difficult it might be for a human to guess the authors of a blinded paper. The AUTHORINFERENCER extracts features of a paper, supplied to it in PDF format, that a human reviewer would likely also – perhaps unconsciously – observe, like the number of citations for an author, the number of times a citation is referred, when in the text the citation occurs for the first time and how old the cited work is. Our initial assumption for the extraction of these features was that recent work by a similar set of authors which is referred more often throughout the paper has a high chance of being written by the same authors. By comparing an extracted reference to the set of authors supplied in the metadata of the PDF, the AUTHORINFERENCER can automatically classify possible authors. We use this classification to learn the extracted features of our corpus of all ICSE and ASIACCS (Symposium on Information, Computer and Communications Security, see Table 1) technical research track publications from 2010 to 2014 with a machine-learner. This simple method is fully paper-contained and does not need external information sources. Naturally, the AUTHORINFERENCER is agnostic about certain sub-domains of research that are occupied by one person, and a distinctive writing style of authors or additional information obtained through other side-channels. Moreover, the AUTHORINFERENCER can automatically extract the email addresses of a PDF paper and match them to an author entry when using the –extract-email-addresses option.

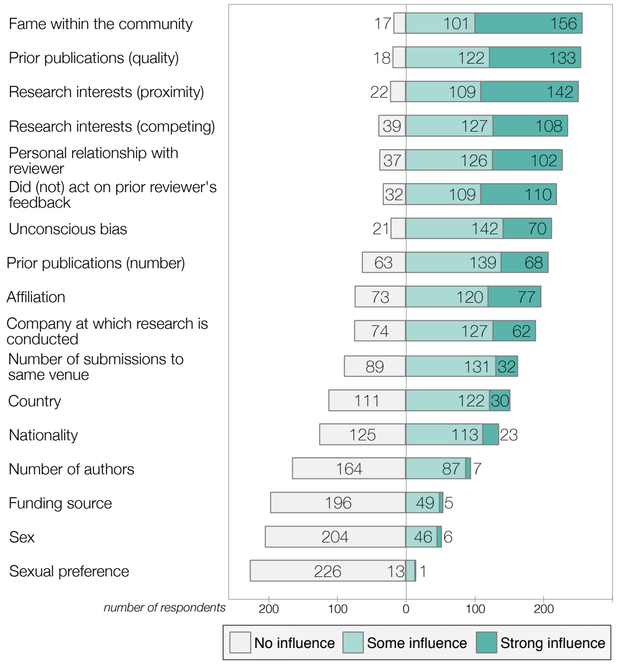

Figure 1: Authors' features potentially influencing reviewer's judgment, according to survey respondents [22.Q16,17]

4. RESULTS

In this section, we present the results to our research questions, using a combination of the research techniques from Section 3.

RQ1: Benefits of Double-Blind Reviewing

We start establishing the potential benefits of DBR, in general by first collecting them from previous literature and card sort on interview data, then asking survey respondents their opinion and experience in that regard. We begin detailing factors related to authorship visibility that can influence reviewers' judgment and where these can play a more visible role (RQ1.1). After we detail benefits, other than increased fairness, that could derive from DBR (RQ1.2).

RQ1.1: Authorship visibility bias, which & where

Rows in Figure 1 list the complete of authors' features that have the potential to influence reviewer's judgment, according to our literature survey and interviewees. In this figure and similar ones, we show the individual results through stacked barcharts for Likert-scale, as suggested by Robbins *et al.* [34], we shorten the items wrt. what presented in the survey, the precise wording of each question, is given in [22]. Respondents associated a perceived strength to each influence [22.Q16] and ranked the top 3 [22.Q17]. The former is used to sort the elements in the figure, the latter is corresponds almost perfectly (each time an influence is ranked 1,2,3 by a respondent, it is assigned a score of 3,2,1, respectively. The final ranking is done summing the scores), so we omit it. The absolute majority of respondents find most of these influences (13 out of 17) to have at least 'some influence' on reviewers' judgment, with authors' fame within the community, quality of prior publications, and proximity of research interests with reviewers ranked as top 3.

A number of previous studies reported that gender/sex of authors and sexual preference can bias reviewers' judgment [], yet these are not deemed as influencer by most of our respondents. Nevertheless, when we take reported sex of the respondents into account, we find a significant relationship (p < 0.01, assessed using the χ^2 with $df = 1$) of weak/moderate strength ($\phi = 0.2$) between it and the influence (s)he associates to author sex on reviewer's judgment. With an odds ratio of 3.5 [35], female respondents are 3.5 times more likely to report that sex has at least some influence on reviewer's judgment (42% of female respondents) than males (17% of male respondents). Interestingly, *all* female interviewees reported that they never felt being judged differently because of their sex.

In the open fields, 25 respondents mentioned reasons that influence reviewers' judgments that are not related to authorship visibility (*e.g.*, quality of the research and presentation) and 9 mentioned authorship visibility related biases, which could be referred to those mentioned in the list in Figure 1, such as "affiliation of the author to some of competing groups," "research institute," and "revenge from previous reject when the [roles where inverted]." This suggests that the list of influential aspects is likely to be complete.

From interviews we identified five situations that can be influenced by authorship visibility bias:(1) when reviewers indicate which papers are preferred for review, *i.e.*, bidding ("*I think the bias* [...] *already starts in the bidding phase*" [I9]) (2) the order in which reviews are done ("*names do matter* [because they change] *the order in which I review*" [I3]); (3) the initial expectations towards the submission ("*if a paper comes from respected authors, I have higher expectations.*" [I7]); (4) the thoroughness with which reviewers conduct a review ("*I just do a more thorough work on names that I don't know, which gives more benefit of doubt to the big guys.*" [I3]); (5) and decision ("[during a meeting] *this other person said: "I actually know the work, it's better than what they described. I think it should be published, and you will accept it anyway because they will fix it for camera ready"*." [DB3]).

In the survey, we asked all respondents to indicate how much they think these aspects are influenced by SBR [22.Q18]: All aspects were deemed to receive at least "some influence" by the absolute majority of the respondents. Reviewer's expectations ranked first and bidding behavior second, closely. Respondents with reviewers' experience were asked to how often they have been personally influenced or have seen the possible influence of authorship visibility bias [22.Q20]; results are presented in Figure 2. We note that the first ranked situation is bidding, where the majority of reviewers felt they at least "sometimes" influenced.

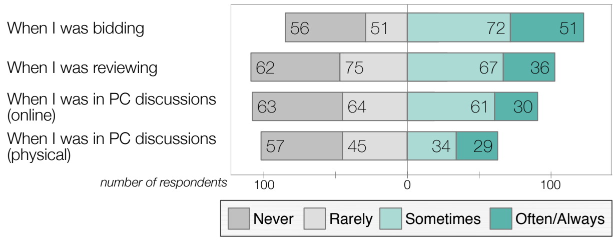

Figure 2: When respondents with reviewer experience think authorship visibility played a role, by frequency [22.Q20]

RQ1.2: DBR benefits other than more fairness

In addition to reducing biases caused by authorship visibility, interviewers reported other benefits deriving from DBR. We list those across three set of questions (we split in consequences for authors [22.Q23], for reviewers [22.Q24], and for the community and conference [22.Q22]) and we ask survey respondents how much they agree that these benefits will derive from DBR, with a 5-level Likert-scale. We also leave space for additional consequences. To reduce bias the potential benefits are interleaved with challenges and side-effects (RQ2.2). Figure 3 details the results.

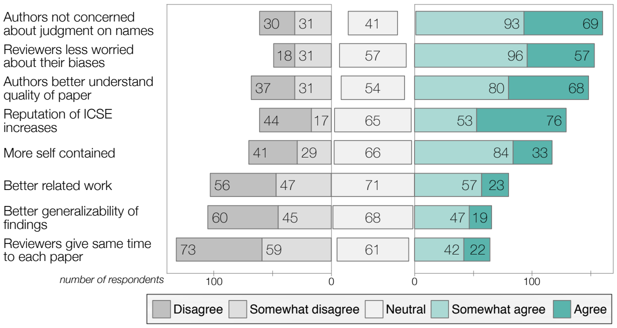

Figure 3: Additional benefits potentially deriving from DBR [22.Q22–24], by survey respondents' agreement

Only 18 respondents (6%) included any additional consequence; all regarding negative side-effects of DBR, but one, which can be considered both ways: "Reviewers will be more definitive about which topics they bid on for review." Results show that, in general, respondents, agree with benefits close to increased fairness, such as "authors will not be concerned with being accepted/rejected due to their identities." As one of our interviewees put it: "*To start with* [DBR] *would simply create a bigger amount of perceived fairness.*" [I7]. However, they are skeptical about more indirect benefits, such as those related to a change the writing style: "papers' related work quality will improve" or "papers will be more self-contained." The indirect benefit with which the majority of respondents agree is that the "reputation of ICSE" will increase. One of our interviewees was strongly supporting this: "[my] *positive attitude to DBR is not because I think the outcome will be very much improved, it's because of the perception we'll have.* [...] *If there was only one reason I would do it for this.*" [I2] In particular, non tenured academics (*i.e.*, assistant professors, post-docs, Ph.D. students, *etc.*) are 3.9 times more likely to agree with this benefit, than tenured ones (*i.e.*, associate and full professors) ($\phi = 0.3$, $p < 0.001$ with χ^2 with $df = 1$). Further research can be conducted to investigate the underlying reason.

RQ2: Costs of Double-Blind Reviewing

Having established the potential benefits of DBR, the question stands which *costs* would be associated with such a fundamental process change, mainly *whether* DBR can easily be circumvented because it might be trivial to guess the authors (RQ2.1). After answering this question, we see which costs and challenges would be associated with switching ICSE to a DBR process, *i.e.*, whether the expected effects are worth these costs (RQ2.2), and which *unintended negative effects* might arise.

RQ2.1: Inferring the blinded authors

For RQ2.1, we obtained all technical papers from ICSE 2014 to 2010 from the ACM Digital Library, resulting in 1,005 PDFs. Similarly, we obtained all 257 technical research papers from ASIACCS in the same period. We chose to compare ICSE against ASIACCS because it is the top-tier conference that most recently switched to double-blind reviewing (see Table 1), and its paper format is similar to ICSE, making it possible to analyze it with the AUTHORINFERENCER. Extracting features from both sets of papers resulted in 37,502 author-feature-tuples from 1,003 parseable ICSE papers (99.8%), and 10,447 from 236 ASIACCS papers (91.8%).

Following the usual machine learning experimentation process [36], we subsequently applied different machine learning algorithms on the data with the aim to predict whether a given possible author of a cited paper, based on the extracted features, is a real author of the paper. Due to a large skew of the attributes `true` and `false` in the predicted class `isAuthor` of ∼1:10, we need a learner that is especially good for the few true authors. Even with ZeroR, a naïve machine learner that always predicts the majority attribute (`false`), we achieve an accuracy of 91.8% on the ICSE data (94.7% for ASIACCS). We received the best improvement over this baseline with a simple decision-table-based learner [37]. When applying it using ten-fold-cross validation on our dataset, we received an overall accuracy of 96.3% (98.0%). For the true authors, figures are not as promising: We receive a precision of 91.8% (99.7%) and a recall of 60.5% (61.5%). The low recall for this class means that only 60% of our guesses for true authors are correct.

However, there does not seem to be a difference in terms of the way the single-blind ICSE papers and the double-blind ASIACCS refer to related work: When applying the model trained on ICSE to ASIACSS papers, we receive an accuracy of 97.8% (-0.2 percentage points). Guessing authors via references cannot succeed for 138 ICSE papers (13.8%) and 20 ASIACCS papers (8.4%) out of methodological reasons, because none of the authors appear in a reference. In total, we were able to guess at least one author for 213 ICSE (21.2%) and 42 ASIACCS papers (17.7%).

Our results show: 1) Our approach can unblind a minority (∼20%) of papers. ∼10% are immune to trying to infer the authors from references. 2) Double-blind and single-blind papers do not seem to differ substantially in the way they referr to previous work. 3) A small pilot study among the three researchers with 9 papers unveiled similar author detection capabilities than the one measured here. We conjecture that humans are worse at structurally gathering all available features than our automated approach, but compensate for this through features we did not measure, *e.g.*, writing style and by knowing the field. 4) There is an important difference in people's behavior when reviewers are directly exposed to author names versus them having an assumption about the authorship seven pages into the paper.

RQ2.2: Challenges and side-effects of DBR

Figure 4 shows the individual costs (challenges and side-effects) that can be a (potential) consequence of DBR, according to our interviewees and the analysis of guidelines from other conferences that made the switch. Costs are ranked by the agreement of the survey respondents ([22.Q22–24]).

We notice that the cardinality of costs we collected (31) greatly exceeds that of benefits, even when considering single influences generated by authorship visibility. Moreover, the absolute majority of respondents mostly agrees ('somewhat agree' and 'agree' answers combined) with 13 of them. These costs mostly regard organizers and authors: The former are supposed to check submissions for *DB compliance* [C1], will have great responsibility during the transition [C5], will have to write extensive *guidelines* [C7] and *educate* the community [C9]; the latter are supposed to have *difficulties in blinding* submissions—especially when building on previous work [C2] or presenting tools [C3]—and additional material (*e.g.*, source code, data, and figures) [C4], not only during the transition period [C11], but also once the DBR process is well established [C10]. Moreover, respondents agree that more *famous authors* will have more difficulties in blinding their identities [C12] and that all authors have to spend time *learning* how to write a DB

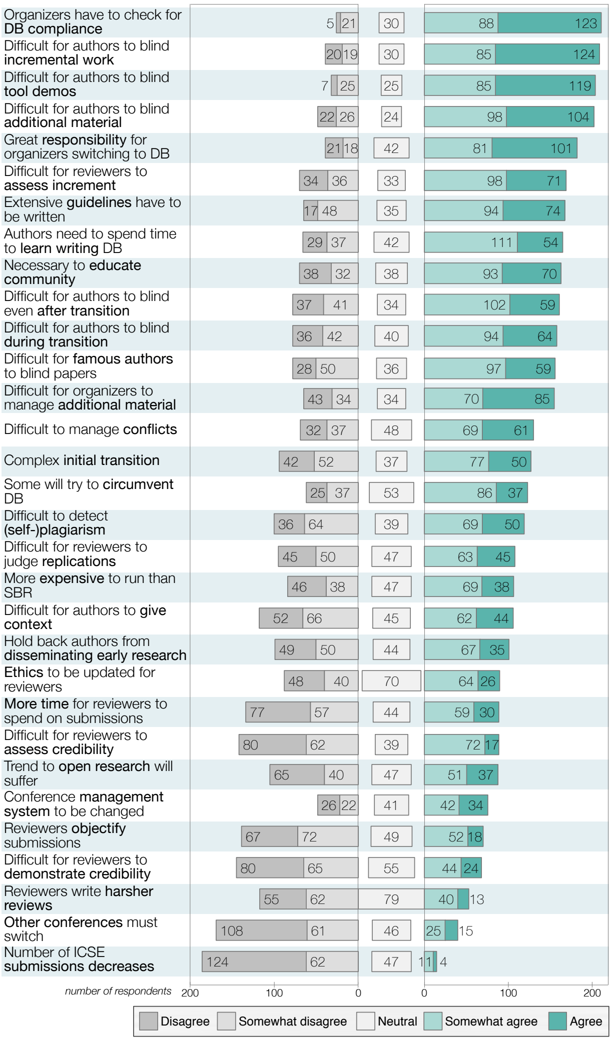

Figure 4: Challenges and side-effects of DBR, sorted by survey respondents' agreement [22.Q22–24]

paper [C8]. The only cost for reviewers on which the absolute majority of respondents agree is the difficulty in *assessing the increment* of submission with respect to previous work from the same authors [C6].

Among the least agreed costs of a switch to DBR, we find those related to demonstrating and assessing work's *credibility* [C28,24]. On this, an interviewee stated: "*I feel like that* [making the names and, thus research background, visible] *gives you a little bit more credibility.*" [I10] Other interviewees stated that they give more benefit of doubt to people they know have done good work in the past, especially on fixes for the camera-ready version. Moreover, respondents do not agree that other SE conference would need to switch to make DBR work [C30], in contrast to our interviewees who were concerned with how resubmitting a paper rejected from a SB conference to a DB conference would make the blinding ineffective, given the overlap of program committee members (*e.g.*, between ICSE and FSE). Finally, the least agreed cost (186 respondents mostly disagreeing) is a decrease in ICSE submissions in case of a switch [C31].

Most of the additional costs (specified in total by 18 respondents) were more specific descriptions of items listed in our questions; among the others we find that some respondents are concerned with a loss in submissions' quality: "authors can submit low-quality papers without a loss in reputation because their identity is blinded."

RQ3: The community on a double-blind ICSE

Having established benefits and costs of DBR, the question stands whether the SE community believes it would be valuable to have DBR at ICSE (RQ3.1). Moreover, we investigate up to which time costs community members are willing to invest in DBR (RQ3.2).

RQ3.1: The community perception of DBR

In the previous research questions we reported the community perception on a set of factors related to DBR. In general, they agree with the majority of the challenges, but also find influences on reviewer's judgment due to authorship visibility realistic and to have a tangible effect on different moment of the review process. However, this does not inform us on whether SE community members perceive that benefits outweigh costs or *vice versa*.

To assess this, we ask a set of direct questions on whether ICSE, SE journals, and other SE venues should switch to DBR or remain SBR. To avoid bias due to the formulation of this important set of questions, we randomly split the respondents in two groups: One had to answer questions in the form of "Do you think that [ICSE/SE journals/all SE venues] should employ double-blind review?", the other group received questions in the form "Do you think that [ICSE/SE journals/all SE venues] should remain single-blind?" Leaving out neutrals, neither formulation made responders more likely to want to switch or stay ($\chi^2 = 0.64$, $\phi = -0.04$), so Figure 5 reports results aggregated on a single formulation. This set of questions received the highest proportion of answers from the 282 respondents who completed the survey. For example, in Figure 3 "Better generalizability of findings in papers" received 239 (85%) answers, while the switch question received 280 (99%).

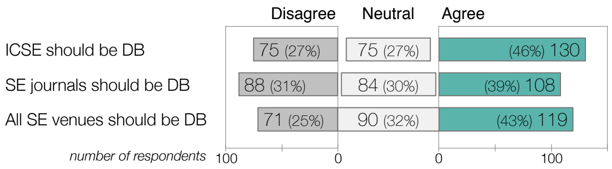

Figure 5: Respondents on switch to double-blind review

Although most respondents agreed with most of the challenging consequences of a switch to DBR, only less than one third of the respondents think that ICSE, SE journals, and other SE venues should remain SBR. The difference between disagreement and agreement is larger on ICSE and conferences, with 130 respondents (46%) agreeing that ICSE should employ DBR. Those agreeing with DBR for conferences but not for journals commented that this was due to the fact that often journal papers are extensions of conference ones, thus it would be almost impossible to maintain authors' anonymity. Most respondents (55) who agreed on a switch for ICSE would like 2017 to be the year for the switch, followed by 2016 (47), and 2018 (13) [22.Q35,39], thus indicating the desire for a rapid change.

Investigating whether characteristics of respondents' relate with the willingness to switch, we find that academic position is significantly related (p < 0.001, using multinomial logistic regression and controlling for sex, main research field is SE, number of publications, number of times at ICSE, and overall occupation). Leaving out neutrals, non tenured academics are 2.96 times more likely to agree with a switch than tenured ones (*i.e.*, associate and full professors) ($\phi = 0.25$, p < 0.001 with χ^2 with $df = 1$).

RQ3.2: Willingness to invest in time for DBR

Among all our respondents, 210 (74%) declare to be willing to invest time *as authors* to make DBR possible, in addition to the time they already put authoring submissions [22.Q26]. Among the respondents who reported to have reviewer experience with SBR venues in SE (240), 162 (68%) declare to be willing to invest time *as reviewers* to make DBR possible, in addition to the time they normally spend reviewing submissions [22.Q30]. Interestingly, even respondents who would disagree with a switch to DBR report that would be available anyway to invest more time as authors (25 respondents) or reviewers (23 respondents) to make DBR work.

In particular, both authors and reviewers are willing to invest up to a median of 4 hours to learn to write/review DB papers if necessary [22.Q28,32]. Authors report to be ready to invest up to a median of 2 hours per submission to make it DB compliant [22.Q29] and reviewers report [22.Q33] a median of up to 15 additional minutes to check if a submission is DB compliant and a median of up to 20-30 minutes per submission for other activities, such as detecting (self-)plagiarism and understanding the increment with respect to previous work.

Respondents not willing to spend additional time for DBR had the chance to motivate their choice with an open field. By analyzing their explanations, we read that respondents not willing to spend additional time as authors motivate their choice with how easy it is to guess the authors (especially due to reviewers bidding on papers on their topic), how time consuming is DB for authors (especially when additional material has also to be masked), or how difficult it is to explain the paper without clear references to previous work. Reviewers not willing to spend additional time motivate their choice explaining that they do not see the need for DBR, that they see reviewing as a substantial time investment and increasing it would be not sustainable, or that they do not think reviewing time should be impacted by the DBR process.

5. DISCUSSION

In this section, we are discussing our research findings.

Too easy to guess the authors. We found that a number of respondents are concerned with the real effectiveness of DBR, because they deem to easy to identify the main authors of a submission, just by looking at the references. Although our quantitative evaluation confirmed that it is possible to guess authors from references, this does not happen for all authors and it does not work in the majority of the case. Moreover, previous research in other fields [2] reported that reviewers that discover the authors of a paper from indirect clues while reading it are less influenced by authorship, than reviewers who can clearly see the names from the start.

"Conflicts, conflicts, conflicts." [I7] Most interviewees among ICSE experts were concerned with the difficulty of managing conflicts in a DBR setting and they found it impracticable to ask program chairs to handle them. As found in guidelines for DB conferences and as DB experts explained, conflicts could be declared by authors (with the risk, though, that some authors declare conflicts with some reviewers for extraneous reasons) or by reviewers. In the latter case, it is advisable to help the conflict declaration by mining previous publication information, for example from the DBLP archive [38], and automatize part of the process [1]. Moreover, to eliminate the possibility that reviewers could infer authors of papers from the list of authors they had to look at for checking conflicts, it is advisable to add names of people in the community that did not submit a paper. Finally, the 162 reviewers who are willing to invest time in DBR declared to are willing to spend an additional 20 minutes for conflict management.

Checking for DB compliance. If omitted, papers are not checked for DB compliance reviewers risk receiving an unblinded paper and wasting their review efforts because they find out who the authors are. There are various levels of thoroughness at which this check could be performed by the Program Chairs, and different possible reactions: The possibilities range from a 10 second check to identify whether there are no names on the paper, to a more thorough check of the content of the paper and how it refers to previous work, which could take as much as 30 minutes per paper. In this case, reviewers who responded to our survey declared to be willing to cover for this time. Another solution for the organizers might be to blind the submissions themselves. In a pilot study among the authors, we successfully blinded published ICSE papers within one to three hours per paper. A very related corollary of DBR is that it is harder for reviewers to detect self-plagiarism because it is unknown who the authors are. Hence, this check could be incorporated in a level where the authors are still known. However, this solution seems only feasible if the number of submissions is very limited or diluted in time.

Bidding. We found that authorship visibility bias can be present as early as the bidding phase. This can be a problem because papers by unknown authors might not receive bids and thus having researchers not expert on the topic to review their papers. Interestingly, with the conference management systems used by a number of SE conferences, the conflict declaration phase and the bidding phase are merged. This means that a reviewer, even if (s)he wanted to avoid looking at names when deciding on which papers to bid, would not be actually able to do it. A simple solution to this issue would be to clearly separate the two phases. After this, blinding the bidding phase would be mostly cost-free and would remove authorship visibility bias in the initial stage of the review process.

A great responsibility. There is wide agreement among survey respondents that an initial transition to DBR is going to be complex: 1) The decision to go double-blind should be well-founded with an emphasis on the idiosyncrasies of the ICSE community. We hope to have significantly reduced this cost with this paper. 2) Similar to organizer's responsibility not to leak the identities of reviewers, they now have to protect the author's, too. This high responsibility – both in terms of fighting accidental errors as well as targeted attempts to circumvent DBR rules in an effort to profit from the DBR process– calls for a smaller dry-run phase before ICSE, best established in a less high-risk setting. 3) ICSE organizers would have to ensure that at least hard conflicts, like former PhD student-supervisor author-reviewer tuples do not occur. 4) The conference management system that ICSE uses for orchestrating the review process needs to support DBR. In particular, this means to support a declaration of conflicts phase that is based on author names (possibly mixed with authors who submitted to previous editions to make guessing of authors harder), and not displaying author names together with the submission for reviewers, but pertaining this information for, for example, program chairs. 5) ICSE's responsibility would include providing extensive guidelines to enable double-blind submissions, educating the whole community.

Learning to do DB research. How much learning effort authors require to be able to write a double-blind paper? Having studied existing guidelines of double-blind conferences, we conjecture that reading one excelently blinded paper and a set of concise DB guidelines typically no longer than two A4 pages suffices to get authors started to blind their paper within one work day. Many respondents agreed that it will be harder for famous authors to conceal their identity, for example because they have a disguising writing style, or because they have coined a certain area of research. It is important to note that a blinded paper does not have to be resistant against any imaginable attempt to conceal the author's real identities. Instead, a code of conduct for reviewers has to be established not to make such attempts. Moreover, a large part of the benefits of DBR stems from the fact that there is no immediate association with author names, allowing reviewers to have a neutral, unbiased start on a paper. One sub-challenge of this is that even after the initial transition, DBR will be more expensive for both authors and conference organizers. Experts in DBR asserted us that there is no difference when writing the paper, except for having to blind additional material. However, removing author's name from additional material is no different and in most cases even easier than anonymizing data sets when publicly shared now. Another solution could be that additional material is not accessible to reviewers at the time of submission, and in case of acceptance, an additional shepperding phase ascertains that authors did share their data, as promised. Both solutions are established in conferences.

A community switch? With an average acceptance rate of 17.4% from 2010 to 2014 [39], most ICSE submissions are rejected, and authors will submit rejected material to other venues, for example ESEC/FSE. It is questionable which benefit DBR would bring to the whole community, if a potential ESEC/FSE reviewer then sees the unblinded version of the ICSE paper. However, this would be no regression from the *status quo*. As such, only a minority agreed to this challenge, and most respondents believe that a DBR ICSE alone would be very effective. A lightweight double-blind process (where the names of the authors are disclosed as soon as a reviewer submits a review) would help tackling the problems in understanding the increment wrt. previous work, but it would make DB more problematic for resubmissions of rejected papers.

6. LIMITATIONS

We designed our study to analyze different aspects of the double-blind subject, from different angles. While we have endeavored to uncover and report benefits, costs, and community perception of DBR, limitations may exist. Especially, regarding the qualitative aspects, gauging the validity of findings is a difficult undertaking [40]. We describe the steps we took to increase confidence and validity.

To achieve a comprehensive view of DBR, we triangulated by collecting and comparing results from multiple sources. For example, we not only analyzed the guidelines of conferences using DBR, but we also interviewed experts who participated to the switch. By starting with exploratory interviews of a smaller set of representative ICSE members (13) followed by open coding to extract themes, we identified core questions that we addressed to DBR experts (5) and the larger SE audience via online survey (282 complete responses). The questions of the survey were validated through (i) consultation with colleagues expert in qualitative research, (ii) a formal pilot run, and (iii) several mini-runs of the survey.

Internal validity – Credibility. We used card sorting to classify the interview data and coding to classify responses in open-ended questions. The coding process is known to lead to increased processing and categorization capacity at the loss of accuracy of the original response. Moreover, result of card sorting could differ depending on the participants. To alleviate this issue, we conducted peer card sorting, where two authors participated and discussed together each card and its placement. Question-order effect [41] (*e.g.*, one question could have provided context for the next one) may lead the respondents to a specific answer. To mitigate this bias, we randomized the elements of most questions in which respondents had to express their opinion in a Likert-scale (*e.g.*, [22.Q22–24]) and we interleaved challenges and benefits. Whenever we decided not to randomize the elements, we decided to order the questions based on the natural sequence of actions (*e.g.*, steps in the review process) to help respondents recall and understand the context of the questions asked. Social desirability bias [42] (*i.e.*, a respondent's possible tendency to appear in a positive light, such as by showing they are fair or rational) may have influenced the answers of both interviewees and survey respondents. To mitigate this issue, we informed participants that the responses would have been anonymous and evaluated in a statistical form. In addition, we ensured interview participants that we would have not shared the transcripts without their written permission.

Generalizability – Transferability. Our interviewees may not be representative of the *average* ICSE community members, because we selected more expert people. To increase the generalizability of our findings, we tested them with the larger SE community. We sent survey invitations not only through our professional networks, which may suffer from convenience bias and be not be representative of the whole community, but we also sent email invitation to participate in our survey to authors of previous ICSE papers. This way, we hope to have minimized the effect that by e.g. just sharing the survey on Twitter, we could reach only like-minded researchers in our own network.

Nevertheless, our survey responses may suffer from a self-selection or voluntary response bias: People who volunteered to respond may have strong opinions on DBR and a potential switch may have decided to invest time in our survey. This bias could affect our sample in both direction: We may have a sample of respondents that is on average either more in favor or against the switch to DBR. Thus, we may argue that if the study is repeated sampling respondents differently (*e.g.*, asking all ICSE participants to fill the survey when registering), the results may be different, especially in terms of number of people who are neutral to the switch.

Finally, we have mostly targeted ICSE authors, but the SE community also includes subfields that should probably have a say on how the flagship conference is run. To alleviate this bias, we publicized our survey also on Facebook and Twitter.

7. CONCLUSION

We investigated double-blind review in the context of SE conferences, particularly with the ICSE community. We identified benefits and costs of DBR, and gathered opinions of SE researchers about this topic, in particular with respect to adopting it for the ICSE conference. It is our hope that the insights we have discovered lead to an informed decision on whether ICSE should remain single-blind or should switch to double-blind and how.

We provide a publicly available replication package with: (i) printable questionnaire, (ii) survey answers, (iii) survey analysis scripts, and (iv) source code for the tools we developed [33].

Personal Statement. During the execution of this study, we discovered many varied opinions on DBR and changed our own opinion many times. In the end, we believe that a light-weight DBR process might bring most of the benefits of DBR with little of its drawbacks.

Acknowledgments

We owe our gratitude to the participants in our interviews and our survey. Thank you so much for your inspiring input! Moreover, we would like to thank Charles Dyer for climbing his desk to access ACL proceedings back to 1977 and Marcel Ackermann for unbureaucratically lifting DBLP access restrictions. We thank Ivan Beschastnikh and Daniel Rozenberg for their contribution to disseminating our survey.

8. REFERENCES

[1] K. S. McKinley, “Editorial: More on improving reviewing quality with double-blind reviewing, external review committees, author response, and in person program committee meetings.” `http://www.cs.utexas.edu/users/mckinley/notes/blind-revised-2015.html`, June 2015. Accessed 2015/08/17.

[2] R. Snodgrass, “Single-versus double-blind reviewing: an analysis of the literature,” *ACM Sigmod Record*, vol. 35, no. 3, pp. 8–21, 2006.

[3] R. Snodgrass, “Editorial: Single-versus double-blind reviewing,” *ACM Transactions on Database Systems (TODS)*, vol. 32, no. 1, p. 1, 2007.

[4] R. Snodgrass, “Frequently-Asked Questions About Double-Blind Reviewing.” `http://tods.acm.org/editorials/doubleblindfaq.pdf`.

[5] S. Hill and F. Provost, “The myth of the double-blind review?: author identification using only citations,” *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 179–184, 2003.

[6] R. M. Blank, “The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review,” *The American Economic Review*, vol. 81, no. 5, pp. pp. 1041–1067, 1991.

[7] van Rooyen S, G. F, E. S, S. R, and B. N, “Effect of blinding and unmasking on the quality of peer review: A randomized trial,” *JAMA*, vol. 280, no. 3, pp. 234–237, 1998.

[8] A. E. Budden, T. Tregenza, L. W. Aarssen, J. Koricheva, R. Leimu, and C. J. Lortie, “Double-blind review favours increased representation of female authors,” *Trends in ecology & evolution*, vol. 23, no. 1, pp. 4–6, 2008.

[9] T. J. Webb, B. OâĂŹHara, and R. P. Freckleton, “Does double-blind review benefit female authors?,” *Heredity*, vol. 77, pp. 282–291, 2008.

[10] D. M. Amodio, E. Harmon-Jones, P. G. Devine, J. J. Curtin, S. L. Hartley, and A. E. Covert, “Neural signals for the detection of unintentional race bias,” *Psychological Science*, vol. 15, no. 2, pp. 88–93, 2004.

[11] P. G. Devine, E. A. Plant, D. M. Amodio, E. Harmon-Jones, and S. L. Vance, “The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice.,” *Journal of personality and social psychology*, vol. 82, no. 5, p. 835, 2002.

[12] H. N. Garb, “Race bias, social class bias, and gender bias in clinical judgment,” *Clinical Psychology: Science and Practice*, vol. 4, no. 2, pp. 99–120, 1997.

[13] H. Zeisel, “Race bias in the administration of the death penalty: The florida experience,” *Harv. L. Rev.*, vol. 95, p. 456, 1981.

[14] A. R. Green, D. R. Carney, D. J. Pallin, L. H. Ngo, K. L. Raymond, L. I. Iezzoni, and M. R. Banaji, “Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients,” *Journal of general internal medicine*, vol. 22, no. 9, pp. 1231–1238, 2007.

[15] C. Goldin and C. Rouse, “Orchestrating impartiality: The impact of "blind" auditions on female musicians,” *American Economic Review*, vol. 90, no. 4, pp. 715–741, 2000.

[16] R. Steinpreis, K. Anders, and D. Ritzke, “The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study,” *Sex Roles*, vol. 41, no. 7-8, pp. 509–528, 1999.

[17] G. S. Metrics, “Top publications - Engineering & Computer Science.” `https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng`.

[18] C. L. Goues, “SSBSE with double blind.” `https://www.cs.cmu.edu/~clegoues/double-blind.html`.

[19] “ISSTA’16 Call for Papers.” `http://issta2016.cispa.saarland/?page_id=37`.

[20] “19th International Conference on Fundamental Approaches to Software Engineering (FASE).” `http://www.etaps.org/index.php/2016/fase`.

[21] J. W. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications, 3rd ed., 2009.

[22] TBD, “Double-blind review and software engineering research.” `somefigsharelink`.

[23] B. Glaser, *Doing Grounded Theory: Issues and Discussions*. Sociology Press, 1998.

[24] B. Taylor and T. Lindlof, *Qualitative communication research methods*. Sage Publications, Incorporated, 2010.

[25] R. Weiss, *Learning from strangers: The art and method of qualitative interview studies*. Simon and Schuster, 1995.

[26] D. Spencer, “Card sorting: a definitive guide.” http://boxesandarrows.com/card-sorting-a-definitive-guide/, April 2004.

[27] B. Martin and B. Hanington, *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*. Rockport Publishers, 2012.

[28] J. E. Shade and S. J. Janis, *Improving Performance Through Statistical Thinking*. Mcgraw-Hill, 2000.

[29] M. L. Patten, *Questionnaire Research: A Practical Guide*. Pyrczak Pub., 2011.

[30] B. Kitchenham and S. Pfleeger, “Personal opinion surveys,” *Guide to Advanced Empirical Software Engineering*, pp. 63–92, 2008.

[31] P. Tyagi, “The effects of appeals, anonymity, and feedback on mail survey response patterns from salespeople,” *Journal of the Academy of Marketing Science*, vol. 17, no. 3, pp. 235–241, 1989.

[32] T. Punter, M. Ciolkowski, B. Freimut, and I. John, “Conducting on-line surveys in software engineering,” in *International Symposium on Empirical Software Engineering*, ISESE’03, pp. 80–88, IEEE, 2003.

[33] A. Bacchelli and M. Beller, “Support package for current submission.” `http://sback.it/dblind`.

[34] N. B. Robbins and R. M. Heiberger, “Plotting likert and other rating scales,” in *Proceedings of the 2011 Joint Statistical Meeting*, 2011.

[35] J. M. Bland and D. G. Altman, “The odds ratio,” *Bmj*, vol. 320, no. 7247, p. 1468, 2000.

[36] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[37] R. Kohavi, “The power of decision tables,” in *Machine Learning: ECML-95*, pp. 174–189, Springer, 1995.

[38] M. Ley, “Dblp: some lessons learned,” *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1493–1500, 2009.

[39] T. Xie, “Software engineering conferences (statistics).” `hhttp://taoxie.cs.illinois.edu/seconferences.htm`. Accessed 2015/08/17.

[40] N. Golafshani, “Understanding reliability and validity in qualitative research,” *The qualitative report*, vol. 8, no. 4,

pp. 597–607, 2003.

[41] L. Sigelaman, “Question-order effects on presidential popularity,” *Public Opinion Quarterly*, vol. 45, no. 2, pp. 199–207, 1981.

[42] A. Furnham, “Response bias, social desirability and dissimulation,” *Personality and Individual Differences*, vol. 7, no. 3, pp. 385 – 400, 1986.

TUD-SERG-2015-007
ISSN 1872-5392