Double-Blind Review in Software Engineering Venues
The Community's Perspective

Bacchelli, Alberto; Beller, Moritz

DOI
10.1109/ICSE-C.2017.49

Publication date
2017

Document Version
Accepted author manuscript

Published in
Proceedings - IEEE/ACM 39th International Conference on Software Engineering Companion, ICSE-C 2017

2017 IEEE/ACM 39th IEEE International Conference on Software Engineering Companion

Double-Blind Review in Software Engineering Venues: The Community's Perspective

Alberto Bacchelli
Delft University of Technology
a.bacchelli@tudelft.nl

Moritz Beller
Delft University of Technology
m.m.beller@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. However, the single-blind review mode currently used in most Software Engineering (SE) venues is susceptible to apparent and hidden biases, since reviewers know the identity of authors. We perform a study on the benefits and costs that are associated with introducing double-blind review in SE venues. We surveyed the SE community's opinion and interviewed experts on double-blind reviewing. Our results indicate that the costs, mostly logistic challenges and side effects, outnumber its 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 complex transition. While the surveyed community largely consents on the costs of DBR, only less than one-third disagree with a switch to DBR for SE journals, all SE conferences, and, in particular, ICSE; the analysis of a survey with authors of submitted papers at ICSE 2016 run by the program chairs of that edition corroborates our result.**

I. Introduction

Peer review is the practice that scientists use to evaluate research manuscripts and artifacts. The outcome of peer review is strongly connected to the advancement of scientific knowledge, researchers' careers [1], and funding decisions from governments and businesses [2].

In past decades, the peer review process has increasingly been called into question [2]. Researchers provided empirical evidence on the shortcomings of peer review, for example, related to reviewers' biases [3], low agreement among reviewers [4], and scarce fit to identify impactful ideas [5]–[7]. Recently, academics have started exploring how different modes of peer review can affect the quality of reviews and their outcome, such as adding monetary rewards for reviewers [8] and changing the number of reviewers [9]–[11].

One of the most recurring topics of debate on how to affect the quality of reviews concerns the visibility of those involved [2]. In particular, whether authors' identity should be visible to the anonymous reviewers, *i.e.*, single-blind review (SBR), or not, *i.e.*, double-blind review (DBR). In principle, the arguments in favor of DBR are predominantly motivated by considerations of fairness, backed up by studies that found that, when authors' identities are known, the evaluation is less objective and several (gender, nationality, language, *etc.*) biases play a role [12]. Arguments against DBR are that it unnecessarily hardens the writing and reviewing of manuscripts [12].

In the software engineering (SE) community, the traditional choice for most conferences and journals is to employ SBR. The International Conference on Software Engineering (ICSE), considered the flagship conference, makes no exception until the current edition. However, a letter to urge ICSE organizers to implement DBR was sent by Brun with the support of several researchers [13]. After one year of debate, ICSE is going to implement a lightweight DBR in 2018 [13].

However, deciding on a switch to DBR is all but 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 (an extensive literature survey is available [12]). 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.

Inspired by the effort by Brun [13], the upcoming switch of ICSE, and several other communities keen on reflecting on their review practice in their main technical tracks (*e.g.*, CSCW [14], CHI [15], medicine [16], economics [17], and neuroscience [18]), we conduct an investigation on DBR in the context of SE research, and ICSE in particular. ICSE covers a wide array of topics and has the largest impact and attendance of any SE conference. Therefore, the lessons learned from studying ICSE should be applicable to most SE venues. In fact, due to the enormous amount of work involved in organizing it, ICSE places the most rigorous constraints of any SE venue on changes to an established review process.

The target audience for this work includes: (1) SE researchers (both authors and reviewers), to be conscious of biases of SBR and challenges of DBR; (2) SE conference organizers, to weight benefits and costs of a switch and be aware of the community's perspective; (3) SE practitioners (sometimes critical of the impact of software engineering research [19]), to understand fairness and reliability of current/different selection processes behind (IC)SE papers; (4) funding agencies, to be aware of how the SE community is self-reflecting on its practices to maintain high scientific value and integrity.

978-1-5386-1589-8/17 $31.00 © 2017 IEEE
DOI 10.1109/ICSE-C.2017.49

We set up our study as an exploratory investigation. We started without *a priori* hypotheses whether and how DBR should be performed, with the aim of discovering the most important aspects to investigate. We first surveyed related literature and conferences that switched and interviewed 14 expert 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 refined it into sub-research questions, which we answer by further analyzing interview data, interviewing 5 experts from DBR communities, and surveying 282 researchers, 242 of which having SE as their main field.

Our results confirm that the benefits of DBR are mostly related to increased fairness due to eliminating authorship visibility and its influence on reviewers. According to our participants, such influences can be seen as early as in the bidding process (during which various participants reported to have been influenced in their choice of papers by author names) and even, albeit rarely, 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 of DBR, mostly logistic challenges and side effects, 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 complex transition. While participants largely consent on the costs of DBR, only less than one-third disagree with a switch to DBR for SE journals, all SE conferences, and, in particular, ICSE; the analysis of a survey [20] run by the program chairs of ICSE 2016 confirms this result's credibility.

II. Background

We first provide an introduction on literature on peer review, then on SBR vs. DBR. We conclude with an analysis of the state of practice of double reviewing.

A. Literature

Overall, the scholarly debate about the value of peer review covers different aspects of the process [2]. For example, parallel to the debate on the anonymity of authors is the discussion on the opportunity to disclose reviewers' identities [21]. In fact, there have been questions regarding the bias, negligence, and self-interest of reviewers [22], [23] that may be intensified by their anonymity [2]. On the other hand, the anonymity of reviewers is believed to increase their frankness, therefore the quality of their reviews [2], and to reduce cases of open rivalry.

The most common debate on peer review—focus of our work—regards the value of the anonymity of the authors to the reviewers [12]. 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 [12], [24]. 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 [25].

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 [26]. However further research [12] reported that reviewers that discover the authors of a paper from indirect clues while reading it are less influenced by authorship, than reviewers who see the names from the start.

Some fields have conducted experiments and case studies with DBR [17]. 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 [27]. Budden *et al.* showed that, after a venue introduced DBR, female authorship increased [28]. However, this was also true for other venues and time period, which still employed SBR [29]. Seeber and Bacchelli found that computer science venues using SBR display a lower rate of contributions from newcomers to the venue, in particular from newcomers otherwise experienced in publishing in other computer science conferences [30].

Outside the world of academic paper reviewing, both intentional and unintentional, conscious and unconscious, racial, gender and other biases have been extensively studied [31]–[33] and shown to exist, even in judges and physicians who reported they were unbiased [34], [35]. 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 [36]. Steinpreis *et al.* randomized names on otherwise identical academic resumes and found that supposedly-male applicants were hired more often than supposedly-female applicants [37]. Therefore, we conclude that more research is needed into the effects of DBR and that it seems unreasonable to assume academic reviewing to be free of hidden biases.

TABLE I
REVIEW MODE OF TOP-TIER COMPUTER SCIENCE VENUES.

Sub-Field	Venue	DBR?	Since
Artifical Intelligence	Expert Systems with Applications	No	
Comput. Linguistics	Meet. of the Assoc. for Comput. Linguistics	Yes	1993
Computer Graphics	Trans. on Graphics	No	
Comp. Hardw. Design	Journ. of Solid-State Circuits	No	
Computer Networks	Communications Magazine	No	
Computer Security	Symp. on Security and Privacy	Yes	2002
Computer Security	Symp. on Info., Comp. and Comm. Security	Yes	2010
Computer Security	Trans. on Inform. Forensics and Security	No	
Computer Vision	Conf. on Comp. Vision and Pattern Rec.	Yes	1985
Computing Systems	Trans. on Parallel and Distributed Systems	No	
Database Systems	Int'l. World Wide Web Conf.s	No	
Database Systems	Int'l. Conf. on Very Large Databases	No	
Human Comp. Interact.	Computer Human Interaction	Yes	2004
Medical Informatics	Journ. of Medical Internet Research	Opt.	
Medical Informatics	Journ. of the Amer. Medical Inform. Assoc.	No	
Robotics	Int'l. Conf. on Robotics and Automation	No	
Signal Processing	Trans. on Signal Processing	No	
Signal Processing	Trans. on Image Processing	Opt.	
Theoretical Comp. Sci.	Symp. on Theory of Computing	No	

B. Practice

To establish which sub-fields are present in Computer Science and their top-tier venues, we used the 15 Computer Science sub-fields suggested by Google Scholar [38] 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. We extracted information on the peer review mode from the conference's websites (when this was not available, we contacted ex program-chairs/editors-in-chief of the prior editions of the venue).

Table I shows the 16 top-tier venues of the 14 sub-fields of Computer Science other than Software Engineering. 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 Intl. 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 SBSSE 2014 [39], ISSTA 2016 [40], and FASE 2016 [41]. Contrary to this trend to switch to DBR, 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 EMSE articles are often extensions of conference papers. Reviewers of the extension are often the same of the conference version, thus DBR was not perceived as cost-effective.

III. METHODOLOGY

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

A. Scoping

To scope our initial focus we tapped into the knowledge of experts from the ICSE community. We targeted ex-chairs, Program Committee (PC) and Steering committee (SC) members. We used 'snowball' sampling starting with randomly selecting people from the ICSE SC and we conducted interviews with 13 of them (Table II). This allowed us not only to gather rich data for our study, but also to let iteratively emerge 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.

B. Data Sources

To investigate costs and benefits of a transition of ICSE to double-blind, we follow a mixed qualitative and quantitative approach [42]. To triangulate and investigate different aspects, we collect and analyze data from different sources: (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, and (6) an online survey to the SE community (particularly ICSE authors) with 281 respondents.

C. Research Questions And Methods

We structure our investigation around three main research questions, organized in several sub-questions.[1]

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

We investigate this question by looking at different aspects and relate it to the SE community.

[1]We refer to specific questions in our survey (publicly available [43]) using a [43.Q*X*] notation, where *X* is the question ID.

TABLE II
INTERVIEWED RESEARCHERS

ID	Academic Rank	Community Service: ICSE PC	Community Service: Steering C.	h-index	sex
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

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 influenced by extenuating circumstances (*e.g.*, the sex of the authors). 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.*, different reviewers 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.

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 DBR 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 (allowing respondents to add any missing bias) [43.Q16] and to rank them by importance [43.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. We complete it by asking survey respondents where they think influences of biases can be stronger for SE reviewers [43.Q18] and, from those with reviewer experience, in which stages of the review process they perceived that they/others may have been influenced [43.Q20].

RQ1.2: Can DBR bring benefits other than fairness?

Rationale: Previous literature reports potential benefits ascribed to DBR in addition to increased fairness [12]. 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 ([43.Q22–24]), with space to add missing ones.

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

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 [12]; 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. We ask survey respondents how much they agree that these costs derive from DBR ([43.Q22–24]), with space to add missing ones.

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

We investigate this aspect through two research questions.

RQ3.1: How does the ICSE community perceive DBR?

Rationale: We aim to understand which kind of value the SE community 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 applies to the SE community.

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.*, [43.Q16]). For example, we ask respondents whether they have ever though if one of their paper was accepted/rejected due to authorship visibility [43.Q14,15], what the strength of reviewers' biases may be [43.Q16], how much the final score and decision of a review may be influenced by authorship [43.Q18], whether they experienced biases as reviewers [43.Q20], and consequences of a switch to DBR [43.Q22–24]. Finally, we ask whether they would like ICSE to DBR [43.Q34,38], as well as other SE conferences [43.Q37,41] and journals [43.Q36,40].

RQ3.2: Which costs would the community pay for DBR?

Rationale: The cost of logistical challenges related to DBR are mostly to be paid in additional time. 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.

Research method: We ask survey respondents whether they would be willing to invest time as authors [43.Q26] and as reviewers [43.Q30] to make DBR review possible. If not, we ask the reason, otherwise we additionally ask how much time they would devote to additional (*e.g.*, learning how to write/review a DBR paper [43.Q28,32]) or more expensive tasks (*e.g.*, declaring conflicts of interest [43.Q33]).

D. Methodological Details

Having gained an understanding of our research questions and methods, 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, using 'snowball' sampling, until—with 13 interviews—we reached a *saturation* point [44]: New interviews were providing insights very similar to the previous.

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 a *semi-structured* interview [45]. This form of interviews uses an *interview guide* that contains general groupings of topics and questions rather than a predetermined exact set and order of questions. 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, then we analyzed the transcripts and split them into *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* [46] 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 re-analyzed their cards to obtain a finer-grained categorization. Finally, we organized the categories using *affinity diagramming* [47], a technique that allows large numbers of ideas to be sorted into groups for review and analysis [48]. 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 the concepts that emerged from the previous phases in the context of the whole (IC)SE community, we created an online survey [43]. For the design of the survey, we followed Patten's guidebook on questionnaire research [49] and Kitchenham and Pfleeger's guidelines for personal opinion surveys [50]. The survey was anonymous and offered the chance to enter a raffle for a 50 Euro gift [51].

To verify clarity and appropriateness of our questions, as well as discuss redundant or missing elements, we run a pilot survey with 10 respondents from our target population.

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. We 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 [52]) completed the survey from the email link. The survey ran in August 2016, before the ICSE 2016 deadline. In total, we collected with 282 complete responses, discarding from the analysis an additional set of responses (163) that did not reach the 'submit' page.

Survey respondents. The 282 participants in our survey hold diverse academic position: 21% Ph.D. students, 14% postdoctoral researchers and the three different professor levels (assistant, associate, and full) account for 20%, 16%, 21% of responses. 18% respondents reported to (also) work in industry. People from 31 countries responded (26% working in the US). 29% of respondents are native English speakers. 69% of participants were Caucasian and 84% male. 242/281 participated at least once in the last 5 ICSEs (median of 2).

IV. Limitations and threats to the validity

We designed our study to analyze different aspects of DBR, from different angles. While we have endeavored to uncover and report benefits, costs, and community perception of DBR, limitations may exist. Especially with the qualitative aspects, gauging the validity of findings is difficult [53]. 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 an 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, the 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 [54] (*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.*, [43.Q22–24]) and we interleaved challenges and benefits. Whenever we decided not to randomize the elements, we ordered the questions based on the natural sequence of actions (*e.g.*, steps in the review process) to help respondents recall and understand the context. Social desirability bias [55] 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 not have 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 an email invitation to participate in our survey to authors of previous ICSE papers. This way, we reduced the effect that by *e.g.*, just sharing the survey on Twitter, we could reach only like-minded researchers in our own network.

Self-selection bias. 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. To assess the existence and strength of this bias, we compared our results with the results of the survey that the program co-chairs of ICSE 2016 sent to all authors of the submitted (accepted and rejected papers) [20]. In that survey, one question asked: "ICSE should use double-blind." We compare it with the results of our similar question: Both surveys report the same direction (*i.e.*, most participants want to adopt DBR), with our results being moderately less strongly in favor of DBR (46% vs. 63%).

V. Results

In this section, we present the answers to our questions.

RQ1: Benefits of Double-Blind Reviewing

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 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.* [56], we shorten the items wrt. what presented in the survey, the precise wording of each question, is given in [43]. Respondents associated a perceived strength to each influence [43.Q16] and ranked the top 3 [43.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.

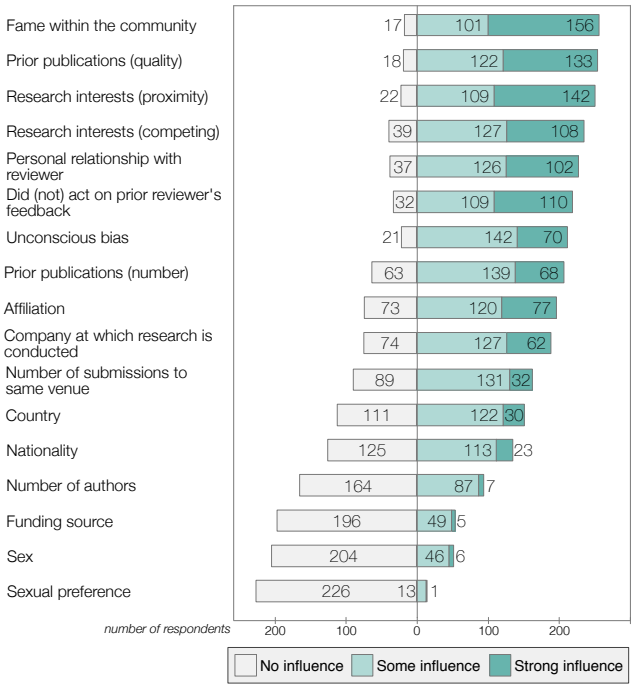

Fig. 1. Characteristics of authors that may influence reviewer's judgment, according to survey respondents [43.Q16,17]

A number of previous studies reported that gender/sex of authors and sexual preference can bias reviewers' judgment [12], 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 [57], female respondents are 3.5 times more likely to report that sex has at least some influence (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 [43.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 how often they have been personally influenced or have seen the possible influence of authorship visibility bias [43.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.

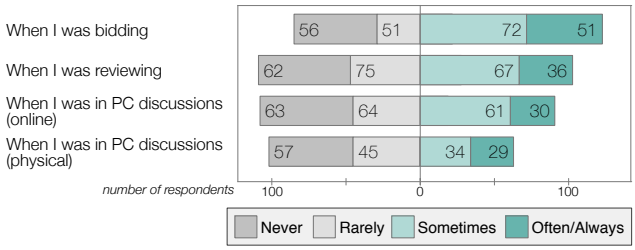

Fig. 2. When respondents with reviewer experience think authorship visibility played a role, by frequency [43.Q20]

RQ1.2: Other DBR benefits 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 [43.Q23], for reviewers [43.Q24], and for the community and conference [43.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.

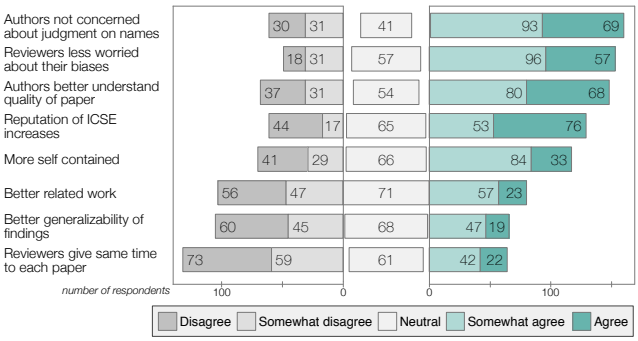

Fig. 3. Additional benefits potentially deriving from DBR [43.Q22–24], by survey respondents' agreement

Only 18 respondents (6%) included 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) ($\phi = 0.3$, $p \ll 0.001$ with χ^2 with $df = 1$).

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. 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 ([43.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

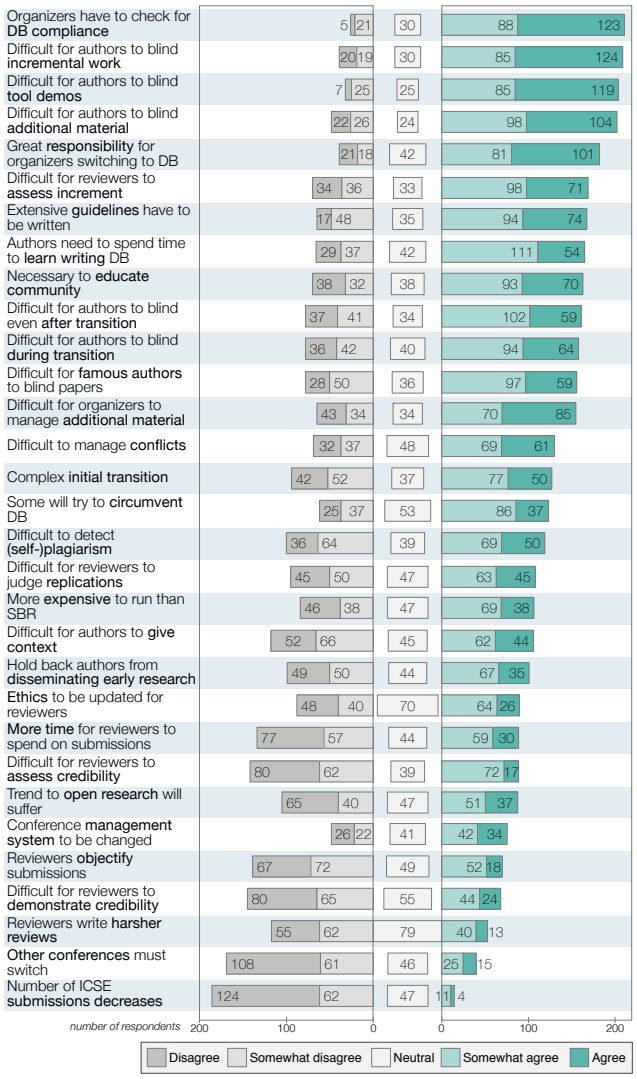

Fig. 4. Challenges and side-effects of DBR, sorted by survey respondents' agreement [43.Q22–24]

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 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 on 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 the 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 an 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 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 double-blind SE venues

Having established benefits and costs of DBR, the question stands if the SE community believes in DBR (RQ3.1) and up to which time costs it is willing to invest in DBR (RQ3.2).

RQ3.1: The community perception of DBR.

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 into 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%).

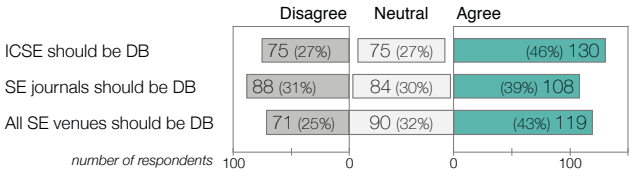

Fig. 5. Respondents on switch to double-blind review

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 impossible to maintain anonymity. Results from the survey run by program co-chairs of ICSE 2016 are in line with our findings: Among all the authors of submissions to ICSE 2016, the trend is towards DBR: "63% of the total authors polled were in favour, only 15% against and 24% neutral." [20] Most of our respondents (55) who agreed on a switch for ICSE would like it to be in 2017, followed by 2016 (47), and 2018 (13) [43.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 ($p \ll 0.001$, using multinomial logistic regression and controlling for sex, main research field is SE, the number of publications, the 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) ($\phi = 0.25$, $p \ll 0.001$ with χ^2 with $df = 1$).

RQ3.2: Willingness to invest 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 [43.Q26]. Among the respondents who reported having reviewer experience with SBR venues in SE (240), 162 (68%) declare to be willing to invest time *as reviewers* to make DBR possible [43.Q30]. Interestingly, even respondents who would disagree with a switch to DBR report to be available to invest more time as authors (25 respondents) or reviewers (23 respondents) to make DBR work.

Both authors and reviewers are willing to invest up to a median of 4 hours to learn to write/review DB papers if necessary [43.Q28,32]. Authors report to be ready to invest up to a median of 2 hours per submission to make it DB compliant [43.Q29] and reviewers report [43.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 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 DB is 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.

VI. Discussion

In this section, we are discussing our research findings.

"*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 [58], 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 are available 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 at 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 have 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 SE 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) SE organizers would have to ensure that at least hard conflicts, like former Ph.D. student-supervisor author-reviewer tuples, do not occur. 4) The conference management system that venues use 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) Venues' responsibility would include providing extensive guidelines to enable DB 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 excellently blinded paper and a set of concise DB guidelines typically no longer than two 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 distinguishable writing style, or because they have coined a certain area of research. We 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 the case of acceptance, an additional shepherding 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 [59], most ICSE submissions are rejected, and authors will submit rejected material to other venues, for example ESEC/FSE. It is questionable which benefits DBR would bring to the whole community if a potential ESEC/FSE reviewer 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 to tackle the problems in understanding the increment wrt. previous work, but it would make DB problematic for resubmissions.

On the value of the community's perceptions. Measuring perceptions of actors in a community is crucial as perceptions drive behavior [60]. In our research, knowing what participants perceive as the most relevant challenges of DBR is a fundamental indication of (1) what has been hindering DBR adoption so far and (2) what must be addressed with the utmost care (*e.g.*, with proper guidelines) should a transition take place. This regardless of whether these challenges are factually more problematic, *e.g.*, as determined by DBR veterans. Similarly, when respondents agree that they perceive that the "reputation of ICSE increases" with DBR, this situation is regardless real in its consequences [60].

Points for further research. Two findings of our study were particularly interesting to us. First, gender is known to create bias in the related literature [12] and it is also feared to create it within our survey participants. Nevertheless, the female experts that we interviewed reported to never have experienced such a bias. This could be due to the success of the specific people we interviewed or social pressure might have lead female experts not to report on perceived gender bias. Nevertheless, one interesting path of future research would be to investigate what caused the perception of the interviewed experts and where the distrust of the SE community on gender bias is originated from. Second, full professors reported being more skeptical to a switch to DBR. A further study could be designed to verify to what extent this is due to an unwillingness to change and traditionalism and to what extent this is the expression of an experienced insight that DBR is not a good solution to ensure good review quality.

VII. Conclusion

We investigated DBR 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 SE venues. It is our hope that the insights we have discovered lead to an informed decision on whether SE venues should remain single-blind or should switch to double-blind and how.

We provide a publicly available replication package [61] with (i) questionnaire, (ii) answers, and (iii) analysis scripts.

Acknowledgments

We owe our gratitude to the participants in our interviews and our survey. Thank you so much for your inspiring input! Moreover, we 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 Felienne Hermans for her initial help on this research. We thank Ivan Beschastnikh and Daniel Rozenberg for their contribution to disseminating our survey. Last but not least, we thank the reviewers of the previous version of this manuscript for their valuable feedback and suggestions.

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] M. Weicher, "Peer review and secrecy in the "information age"," *Proceedings of the American Society for Information Science and Technology*, vol. 45, no. 1, pp. 1–12, 2008.

[3] J. S. Armstrong, "Peer review for journals: Evidence on quality control, fairness, and innovation," *Science and engineering ethics*, vol. 3, no. 1, pp. 63–84, 1997.

[4] L. Bornmann and H.-D. Daniel, "The luck of the referee draw: the effect of exchanging reviews," *Learned publishing*, vol. 22, no. 2, pp. 117–125, 2009.

[5] J. Campanario, "Rejecting and resisting nobel class discoveries: accounts by nobel laureates," *Scientometrics*, vol. 81, no. 2, pp. 549–565, 2009.

[6] J. Chen and J. A. Konstan, "Conference paper selectivity and impact," *Communications of the ACM*, vol. 53, no. 6, pp. 79–83, 2010.

[7] K. Siler, K. Lee, and L. Bero, "Measuring the effectiveness of scientific gatekeeping," *Proceedings of the National Academy of Sciences*, vol. 112, no. 2, pp. 360–365, 2015.

[8] F. Squazzoni, G. Bravo, and K. Takács, "Does incentive provision increase the quality of peer review? an experimental study," *Research Policy*, vol. 42, no. 1, pp. 287–294, 2013.

[9] P. J. Roebber and D. M. Schultz, "Peer review, program officers and science funding," *PLoS One*, vol. 6, no. 4, p. e18680, 2011.

[10] F. Bianchi and F. Squazzoni, "Is three better than one? simulating the effect of reviewer selection and behavior on the quality and efficiency of peer review," in *2015 Winter Simulation Conference (WSC)*, pp. 4081–4089, IEEE, 2015.

[11] R. R. Snell, "Menage a quoi? optimal number of peer reviewers," *PloS one*, vol. 10, no. 4, p. e0120838, 2015.

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

[13] Y. Brun, "A case for double-blind reviewing in software engineering." https://people.cs.umass.edu/~brun/doubleblind.html, Apr 2015.

[14] M. Jacovi, V. Soroka, G. Gilboa-Freedman, S. Ur, E. Shahar, and N. Marmasse, "The chasms of cscw: a citation graph analysis of the cscw conference," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pp. 289–298, ACM, 2006.

[15] J. Kaye, "Some statistical analyses of chi," in *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pp. 2585–2594, ACM, 2009.

[16] J. R. Gilbert, E. S. Williams, and G. D. Lundberg, "Is there gender bias in jama's peer review process?," *Jama*, vol. 272, no. 2, pp. 139–142, 1994.

[17] 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.

[18] R. Walker and P. Rocha da Silva, "Emerging trends in peer review-a survey.," *Frontiers in neuroscience*, vol. 9, pp. 169–169, 2014.

[19] D. Lo, N. Nagappan, and T. Zimmermann, "How practitioners perceive the relevance of software engineering research," in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, pp. 415–425, ACM, 2015.

[20] L. W. Willem Visser, "Insights and Lessons Learned from Analyzing ICSE 2016 Survey and Review Data." http://2016.icse.cs.txstate.edu/static/downloads/2016ReviewProcess.pdf, May 2016.

[21] R. G. Bachand and P. P. Sawallis, "Accuracy in the identification of scholarly and peer-reviewed journals and the peer-review process across disciplines," *The Serials Librarian*, vol. 45, no. 2, pp. 39–59, 2003.

[22] J. M. Campanario, "Peer review for journals as it stands today—part 1," *Science Communication*, vol. 19, no. 3, pp. 181–211, 1998.

[23] J. M. Campanario, "Peer review for journals as it stands today—part 2," *Science Communication*, vol. 19, no. 4, pp. 277–306, 1998.

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

[25] R. Snodgrass, "Frequently-Asked Questions About Double-Blind Reviewing." http://tods.acm.org/editorials/doubleblindfaq.pdf.

[26] 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.

[27] 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.

[28] 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.

[29] T. J. Webb, B. O'Hara, and R. P. Freckleton, "Does double-blind review benefit female authors?," *Heredity*, vol. 77, pp. 282–291, 2008.

[30] M. Seeber and A. Bacchelli, "Does single blind peer review hinder newcomers?," *Scientometrics*, forthcoming.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

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

[39] C. L. Goues, "SSBSE with double blind." https://www.cs.cmu.edu/~clegoues/double-blind.html.

[40] "ISSTA'16 Call for Papers." http://issta2016.cispa.saarland/?page_id=37.

[41] "19th International Conference on Fundamental Approaches to Software Engineering (FASE)." http://www.etaps.org/index.php/2016/fase.

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

[43] M. B. Alberto Bacchelli, "Double-blind review in software engineering venues – Survey." http://sback.it/dblind/survey.html.

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

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

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

[47] 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.

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

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

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

[51] 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.

[52] T. Punter, M. Ciolkowski, B. Freimut, and I. John, "Conducting online surveys in software engineering," in *International Symposium on Empirical Software Engineering*, ISESE'03, pp. 80–88, IEEE, 2003.

[53] N. Golafshani, "Understanding reliability and validity in qualitative research," *The qualitative report*, vol. 8, no. 4, pp. 597–607, 2003.

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

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

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

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

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

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

[60] W. I. Thomas, "The methodology of behavior study," in *The child in America: Behavior problems and programs*, ch. 13, pp. 553–576, A. A. Knopf, 1928.

[61] A. Bacchelli and M. Beller, "Support package for current submission." http://sback.it/dblind.