On Researcher Bias in Software Engineering Experiments

Simone Romano[a,*], Davide Fucci[b], Giuseppe Scanniello[c], Maria Teresa Baldassarre[a], Burak Turhan[d], Natalia Juristo[e]

[a] *University of Bari, Bari, Italy*
[b] *Blekinge Institute of Technology, Karlskrona, Sweden*
[c] *University of Basilicata, Potenza, Italy*
[d] *University of Oulu, Oulu, Finland (and Monash University, Melbourne, Australia)*
[e] *Universidad Politécnica de Madrid, Madrid, Spain*

Abstract

Researcher bias occurs when researchers influence the results of an empirical study based on their expectations, either consciously or unconsciously. Researcher bias might be due to the use of *Questionable Research Practices* (*QRPs*). In research fields like medicine, *blinding* techniques have been applied to counteract researcher bias. In this paper, we present two studies to increase our body of knowledge on researcher bias in Software Engineering (SE) experiments, including: *(i)* QRPs potentially leading to researcher bias; *(ii)* causes behind researcher bias; and *(iii)* possible actions to counteract researcher bias with a focus on, but not limited to, blinding techniques. The former is an interview study, intended as an exploratory study, with nine experts of the empirical SE community. The latter is a quantitative survey with 51 respondents, who were experts of the above-mentioned community. The findings from the exploratory study represented the starting point to design the survey. In particular, we defined the questionnaire of this survey to support the findings from the exploratory study. From the interview study, it emerged that some QRPs (*e.g., post-hoc outlier criteria*) are acceptable in certain cases. Also, it appears that researcher bias is perceived in SE and, to counteract researcher bias, a number of solutions have been highlighted. For example, duplicating the data analysis in SE experiments or fostering open data policies in SE conferences/journals. The findings from the interview study are mostly confirmed by those from the survey, and allowed

*Corresponding Author

Email addresses: simone.romano@uniba.it (Simone Romano), davide.fucci@bth.se (Davide Fucci), giuseppe.scanniello@unibas.it (Giuseppe Scanniello), mariateresa.baldassarre@uniba.it (Maria Teresa Baldassarre), burak.turhan@oulu.fi (Burak Turhan), natalia@fi.upm.es (Natalia Juristo)

us to delineate recommendations to counteract researcher bias in SE experiments. Some recommendations are intended for SE researchers, while others are purposeful for the boards of SE research venues.

Keywords: Researcher bias, experimenter bias, survey, blinding

1. Introduction

In research, *bias* is defined as the combination of various design, data, analysis, and presentation factors tending to produce findings that should not be produced [1]. *Researcher bias*, or *experimenter bias*, occurs when the researcher (consciously or unconsciously) influences the results of an empirical study based on their expectations.

In some cases, researcher bias is due to the use of *Questionable Research Practices* (*QRPs*) to follow one's agenda and achieve specific expectations—*e.g.,* changing the procedure for excluding data after looking at the impact of doing so on the results. Another form of bias is *publication bias*, which occurs when studies are published based on their results—usually positive results are more likely to be published than negative ones [2].

To counteract researcher bias, according to established guidelines in Software Engineering (SE), researchers should disclaim their stance regarding an outcome. For example, Wohlin *et al.* [3] and Sjøberg and Bergersen [4] consider *experimenter expectancies* as a threat to validity in SE experiments.

In this paper, we present the results of two studies, an interview study [5] and a survey, to increase our body of knowledge about researcher bias in *human-* and *technology-oriented* SE experiments.[1] The interview study, intended as an exploratory study, aimed to gather the opinions of a group of experts about themes related to researcher bias in SE experiments. To collect data, we used semi-structured interviews. In particular, we interviewed nine experts of the empirical SE field. The interviews were concerned with: QRPs potentially leading to researcher bias, causes behind researcher bias, and possible actions to counteract it. Regarding the possible actions, we focused on (but not limited to) two *blinding* techniques, namely: *blind data extraction* and *blind data analysis*. The former consists of hiding some information (*e.g.,* treatment assignment) from the researchers who extract the data; while, the latter is the temporary and judicious removal of labels and/or alteration of values before someone analyzes the data [6]. Although extensively used in other research

[1]In human-oriented experiments, participants apply treatments to objects (or receive treatments), while in technology-oriented experiments, tools are usually applied to objects [3].

fields like medicine and physics [6, 7], SE researchers have used these techniques only in few occasions [8, 9].

The findings from the interview study represented the starting point to design our survey. In particular, we built a series of statements based on the findings from the interview study and then gathered, through a questionnaire, the level of agreement of experts in conducting SE experiments about these statements. The goal of the survey was to support the findings from the interview-based one. This methodological approach was inspired by past work in the SE research field (*e.g.,* [10, 11, 12]).

This paper extends the one by Romano *et al.* [5], presenting the findings from the interview study on researcher bias in SE experiments, as follows:

– It adds a new study, a survey with experts in the empirical SE field, aiming to support the findings from the interview study.

– It extends the discussion of the results by taking into account both interview study and survey.

Paper Structure. In Section 2, we summarize work related to ours. In Section 3, we present the design of both interview study and survey. The findings emerging from these two studies are shown in Section 4. In Section 5, we discuss the results, offering recommendations based on both studies, as well as possible limitations. Finally, we draw conclusions in Section 6.

2. Background

This section considers current relevant literature focusing on QRPs and researcher bias. We also illustrate some countermeasures adopted to deal with researcher bias, including blinding techniques.

2.1. Questionable Research Practices and Researcher Bias

Cases of QRPs, exploiting the grey area of what is considered acceptable, have been mounting in medicine, natural sciences, and psychology (*e.g.,* [13, 14]). As for the SE research field, Jørgensen *et al.* [15] documented the presence of researcher bias and publication bias in SE experiments. The authors conducted a quantitative questionnaire-based survey, with researchers from some SE sub-communities, comprising questions about QRPs potentially leading to researcher bias and publication bias. Three out of seven questions were on QRPs related to researcher bias, namely:

1. *Post-hoc hypotheses*—defined as reporting the results of one (or more) hypothesis tests where at least one of the hypotheses is formulated after looking at the data.

2. *Post-hoc outlier criteria*—defined as developing or changing the rules for excluding data (*e.g.,* outlier removal) after looking at the impact of doing so on the results.

3. *Flexible reporting of measures and analysis*—defined as using several variants of a measure or several tests and then reporting only the measures and tests that give the strongest results.

The authors gathered 34 responses and found that: *(i)* 67% of the respondents had followed the post-hoc hypotheses practice; *(ii)* 55% had followed the post-hoc outlier criteria practice; and *(iii)* 69% had followed the flexible reporting of measures and analysis practices. Jørgensen *et al.* [15] also built a model—based on 150 randomly-sampled SE experiments—to estimate the proportion of correct results at different levels of researcher bias and publication bias. The model suggests that both researcher bias and publication bias affect SE experiments since 52% of the statistically significant tests do not match a situation with no or low researcher bias and publication bias.

Shepperd *et al.* [16] in their meta-analysis of defect prediction techniques came to a conclusion similar to that by Jørgensen *et al.* [15]. The authors pointed out the presence of researcher bias in the studies included in the meta-analysis as the factor with the largest effect was the research group publishing the paper, while the effect of the prediction technique was small.

2.2. Countermeasures to Researchers Bias

Researchers have proposed solutions to counteract researcher bias (*e.g.,* [17, 18]). We can group these solutions into: *(i) rival theories*; *(ii) transparency*; and *(iii)* blinding. The first category consists of considering alternative or competing hypotheses with respect to the ones being tested in the study. The researcher should devise experiments that can explicitly distinguish competing hypotheses and, if possible, develop experiments that can distinguish between alternative theories. It is ideal that the researcher collaborates with a *team of rivals*—*i.e.,* other researchers that, while being skeptical about the hypotheses, collaborate towards developing alternative explanations.

Several approaches fall under the umbrella of the transparency category. The main example is *open science*—*i.e.,* the practice of sharing research data, computer code, and lab packages for public scrutiny so attempting to reproduce results. In research fields like medicine or psychology, transparency is also achieved through *pre-registration* (also known as *registered report*). It consists of submitting a study proposal presenting the study rationale and planning for peer review before conducting the study. Once the proposal is accepted, the researchers can conduct the study and submit a paper with the obtained results for a second round of revision. The paper cannot be rejected due to the study results (*e.g.,* negative results), while it can be rejected for other reasons (*e.g.,* deviations from the pre-registered analysis procedure) [19].

Finally, blinding (also known as *masking*) means concealing research design elements (*e.g.,* treatment assignment or research hypotheses) from individuals involved in an empirical study (*e.g.,* participants, data collectors, or data analysts) [20, 21]. Research fields like medicine and physics [6, 7] have been encouraging the use of blinding techniques to deal with research bias. As for the SE research field, Shepperd *et al.* [16] have fostered researchers to use blinding techniques in their studies. However, few researchers have applied blinding techniques in SE studies so far, namely: Fucci *et al.* [8] who used blind data extraction and analysis in a human-oriented experiment, and Sigweni and Shepperd [9] who applied blind data analysis in a technology-oriented experiment.

To explain how blind data extraction and analysis work, we refer to the experiment by Fucci *et al.* [8] as an example. The study goal was to assess the impact of Test-Driven Development (TDD) on *(i)* functional quality of developed programs, *(ii)* developers' productivity, and *(iii)* number of tests written. To that end, the experiment compared a *treatment group*—*i.e.,* a group of developers who applied TDD to implement some programs—to a *control group*—*i.e.,* a group of developers who implemented the same programs as the other group but by following Test-Last Development (TLD). Once the experiment was carried out, the raw dataset (*i.e.,* the programs implemented by the developers) was handed over to a researcher playing the role of data extractor. In particular, given the raw dataset, this researcher extracted the values of the metrics (*e.g.,* the PROD metric that quantified developers' productivity) so obtaining the dataset. The extraction of the metrics was done blindly because the data extractor was aware of neither the experimental goal, hypotheses, treatment assignment, nor design. Next, the dataset was forwarded to two data analysts who performed the analysis (both descriptive and inferential) blindly. This is because they worked on a sanitized dataset and did not know the experimental goal. To sanitize the dataset, the labels of the experimental groups were temporarily replaced (*e.g.,* the TDD group became the A group, while the TLD group became the B group) and the dependent variables were temporarily anonymized (*e.g.,* PROD was renamed as DV1). To correctly analyze the data, the analysts were provided with a minimal description of the dependent and independent variables (*e.g.,* DV1 is a dependent variable assuming values between 0 and 1), as well as the experimental design in which some information was adequately hidden (*e.g.,* the experimental groups were referred to as A and B). The hidden information was disclosed once the analysis was completed (*e.g.,* group A was actually the TDD group).

As mentioned-before, Sigweni and Shepperd [9] used blind data analysis in a technology-oriented experiment. In particular, they assessed four prediction methods for software effort estimation to demonstrate the practicality of blind data analysis in SE experiments. The analyst did not know the prediction methods to be assessed (*i.e.,* the name of the prediction methods was replaced). Moreover, any analysis was based on absolute residuals. The authors concluded that blind data analysis is a very practical technique that supports more objective analyses of experimental results.

3. Interview Study and Survey

In this section, we describe the design of both interview study and survey.

3.1. Protocol

For the first step of our research (*i.e.,* the interview study), we opted for interviews as a data collection means, rather than questionnaires, because: *(i)* they decrease the number of "don't know" and "no answers", as the interviewees can ask for clarifications if a question is not clear to them, and *(ii)* the interviewer can ask for clarifications/details if needed [3]. Also, such a data collection means fits the exploratory intention of our study.

We recruited researchers in our research network, who were experts in conducting (human- and technology-oriented) SE experiments. Nine researchers (also referred to as the interviewees, from here onward) were available to be interviewed either face-to-face or by phone. Each interview session involved the same interviewer (*i.e.,* the second author) and one interviewee at a time. At the beginning of the interview session, we obtained the consent of the interviewee for audio-recording the session. Also, we informed the interviewee that the gathered data would be treated confidentially. Each interview lasted between 50 and 75 minutes. We used semi-structured interviews [3]. That is, the questions listed in the interview script were not necessarily asked in order because, depending on how the conversation evolved, some questions were handled before others. Semi-structured interviews allow for improvisation and exploration of the investigated phenomenon. The interview script is roughly a checklist that the interviewer adopts to guide the discussion with the interviewee and make sure that relevant topics are covered [3]. In Figure 1, we show the interview script.

With the second step of our research (*i.e.,* the survey), we aimed to support the findings from the interview study by gathering the level of agreement of experts in conducting SE experiments about a series of statements we built upon the findings of the interview study. In other words, we aimed to apply a kind of *triangulation*[2] known as *methodological triangulation* [22]. Unlike the interview study, the questionnaire-based one is quantitative since it is informed by quantifiable data (*i.e.,* the level of agreement of experts in conducting SE experiments about some statements). We opted for questionnaires as a data collection means because it fits our research purpose—*i.e.,* validating the findings from a past exploratory investigation (*e.g.,* [12]). Moreover, questionnaires require less effort than interviews and can reach a larger part of the population [3].

We invited 317 empirical SE experts (or simply researchers, from here onwards) to fill in our (online) questionnaire. In particular, we invited researchers who had published papers in the ESEM[3] proceedings in the last three years. We (all authors of this paper) analyzed this list of empirical SE experts to validate and extend it. Each author added researchers (not included in this list) considered as an active researcher on topics related to empirical SE. We focused on ESEM because this conference can be considered the major forum for researchers acting in the context of empirical SE. It is worth mentioning that we did not invite the researchers who had taken part in the interview study because they would be clearly favorable towards the statements we built based on their opinions.

To ask SE experts to participate in the survey, we sent them an invitation letter via email (see Appendix A). The letter reported the objective of the survey, the due date to fill in the questionnaire, and the link to the online questionnaire. We also informed the invited researchers that they could freely share the questionnaire with other empirical SE experts. The invitation letter was sent on November 5th 2020. The survey was open for 20 days. We received 64 answers (response rate of 20%), of which 51 answers from respondents reporting to have carried out an experiment in the past. This resulted in a sample size (n) of 51. Each answer was unique (*i.e.,* the same researcher cannot send two answers) and anonymous.

[2]The procedure of combining two (or more) data sources, investigators, methodological approaches, theoretical perspectives, or analysis methods to increase confidence in study findings.

[3]International Symposium on Empirical Software Engineering and Measurement.

Hello {name}, thank you for agreeing to do this interview. With this study, I want to gather opinions of experts in the empirical SE community about researcher bias. Hence, I want to interview you as a member of said community, as well as a researcher who has been conducting experiments in SE. The gathered data will be handled confidentially and your name will not be exposed in the write-up of the study. Is there anything you would like to mention or ask before we begin?

Warm up:

1. What institution do you work for?
2. What is your job title?
3. What are your research interests?
4. For how many years have you been conducting research in empirical SE?
5. When was the last time you published a study reporting one or more experiments?

Experiments:

Walk me through your usual experimental process.

1. Can you summarize that experiment(s)?
2. Who was involved (researcher), and what was her role?
3. Can you elaborate on the threats to validity?

Questionable Research Practices:

Talking about conducting experiments, let's discuss the following practices (you are welcome to give examples):

1. What do you think about the practice of reporting the results of one or more hypothesis tests where at least one of the hypotheses is formulated after you have looked at the data?
2. What do you think about the practice of developing or changing the rules for whether to exclude data or not (*e.g.*,, outlier removal) after looking at the impact of doing so on the results?
3. What do you think about the practice of using several variants of a measure or several statistical tests and then using only the measures and tests that give the strongest results?

Researcher Bias:

It occurs when researchers, consciously or unconsciously, influence the results of a study based on their expectations.

1. Do you think that researcher bias is a problem in SE research? Why? If so, how widespread do you think this problem is?

A survey by Jørgensen *et al.* (published in 2015) reports that: 67% (of the surveyed researchers) had statistically tested and reported post-hoc hypotheses, 55% had developed/modified outlier criteria after looking at the impact of doing so on the results, and 69% had only reported the best among several measures or tests at least once. Much fewer of the participants (10-22%) admitted using each of these practices often.

2. What you think is causing such results and, in general, researcher bias?
3. How would you limit researcher bias? Are you aware of any technique or process that might help avoid or lessen researcher bias (not necessarily in SE)? Can you give me some examples (not necessary from SE)? Have you used any?

Blind Data Extraction:

A researcher (or more) transforms the raw dataset (*e.g.,* code bases) into the dataset to be analyzed without knowing some information like treatments, subjects, *etc.*

1. What are the main motivations for not using blind data extraction? Do you think some contexts are more/less suited for blind data extraction? To what extent do you believe SE research will benefit from using blind data extraction? Any specific context?
2. Do you think that SE experiments will benefit from the use of blind data extraction? Why?

Blind Data Analysis:

A researcher (or more) performs the data analysis on a dataset where labels (*e.g.,* references to treatments) have been temporarily and judiciously removed and/or the values have been temporarily and judiciously altered. So she does not know some information like treatment, dependent variable, etc.

1. Do you think that SE experiments will benefit from the use of blind data analysis? Why?
2. What are the main motivations for not using blind data analysis? Do you think some contexts are more/less suited for blind analysis? To what extent do you believe SE research will benefit from using blind data analysis? Any specific context?

Blind Data Extraction and Analysis:

1. Do you think the combination of blind data extraction and blind data analysis is enough to cope with researcher bias? Why?
2. Do you have any suggestion to ease the adoption of blind data extraction and analysis?

Wrap up:

1. Do you think you will use blind data extraction and analysis in the future?

Figure 1: Interview Script.

The questionnaire started with a *filter question*[4] in which we asked the researchers whether they had ever carried out an experiment (human- and/or technology-oriented). This is because our goal was to investigate researcher bias in SE experiments, and we were aware that some respondents could not be experts in conducting experiments (while regarding themselves as experts, for example, in conducting case studies). Respondents who had carried out at least an experiment in the past could continue with the questionnaire, while those who had never carried out an experiment ended the questionnaire immediately.

The first part of the questionnaire (*i.e., Demographics*) included demographic questions (*e.g.,* the academic position of the respondent or the research outlet where the respondent published her experiments) to better characterize the study context. To increase the response rate, the demographic questions were not mandatory as some respondents could not be willing to share some information such as the research outlet where the respondents published their experiments.

The remaining part of the questionnaire aimed to support the findings from the interview study. To that end, we built a series of statements based on the findings from the interview study. To keep the questionnaire at a reasonable length, we prioritized the statements extracted from the interview study by relevance and included in the questionnaire only those statements we deemed more relevant as suggested in the literature (*e.g.,* [23]). For each statement, respondents had to rate how much they agreed with that statement on a (Likert-type) scale from 1 (*i.e., "Strongly disagree"*) to 5 (*i.e., "Strongly agree"*). For example, one of the findings emerging from the interviews is that the post-hoc outlier criteria practice should be avoided because it potentially leads to researcher bias (see Section 4.2). Therefore, we asked the respondents their level of agreement with the following statement: *"The post-hoc outlier criteria practice should be avoided because it potentially leads to researcher bias."* As shown in Figure 2, we arranged these statements into three sections. The answers to these statements were mandatory.

To evaluate the comprehensibility of the questionnaire and reduce as much as possible sources of misunderstanding, we conducted a pilot with two junior researchers (who were not involved in this research and were not invited to participate in the actual survey). Based on pilot feedback, we made changes to improve the clarity of the questionnaire before the survey took place.

It is worth remarking that, from here onwards, we refer to the researchers/participants who took part in the interview study as the interviewees, while we refer to those who

[4]Filter questions are the ones that aim to avoid respondents answering questions that do not pertain to them.

Experiments and Questionable Research Practices

S1. In the experiments in which I took part as an experimenter, only one researcher usually performed the data analyses (*i.e.,* only one researcher played the data analyst role).

S2. The use of the post-hoc hypotheses practice does not lead to researchers bias as long as the researchers clearly report that these hypotheses are formulated in retrospect.

S3. The use of the post-hoc hypotheses practice does not lead to researcher bias as long as it is possible to ground such hypotheses on prior work.

S4. The post-hoc hypotheses practice could be a means to get new insight into the studied phenomenon, which researchers had not thought about when the study was planned.

S5. The post-hoc outlier criteria practice should be avoided because it potentially leads to researcher bias.

S6. The post-hoc outlier criteria practice does not lead to researcher bias as long as the researcher declares the use of this practice in the paper by providing the following information.

1. The analysis results before and after removing outliers.
2. The reasons behind the outlier removal.
3. An interpretation of the results (*e.g.,* why, after the outlier removal, a null hypothesis passes from non-rejected to rejected).

S7. If a statistical hypothesis test (e.g., paired t-test) revealed a significant difference that an equivalent test (*e.g.,* Wilcoxon signed-rank test) did not, that difference (estimated by using an effect size measure) would be probably negligible, so using a test rather than another one does not matter.

S8. The flexible reporting of measures practice leads to researcher bias.

Research Bias

S9. Researcher bias is present in SE experiments of the following kind:

1. Human-oriented experiments.
2. Technology-oriented experiments.

S10. Researcher bias affects the findings from experiments in the software engineering research field as much as other research fields (*e.g.,* medicine or psychology).

S11. When reviewing papers reporting SE experiments, I have suspected that authors bias the results.

S12. Researchers can unconsciously bias the results based on their expectations.

S13. Researcher bias is one of the reasons for inconsistent results among studies investigating the same constructs.

S14. The rejection of papers reporting negative/null results leads some researchers to bias the results (*e.g.,* transforming non-significant results into statistically significant ones).

S15. The pressure of publishing papers leads some researchers to (unconsciously or consciously) bias the results.

S16. The revision process of SE conferences/journals is focusing too much on the rigor of the empirical assessment rather than on the novelty of contributions.

S17. The use of pre-registration in SE conferences/journals can mitigate researcher bias.

S18. Fostering open data policies in SE conferences/journals can mitigate researcher bias.

S19. The use of duplicate data analysis can mitigate researcher bias.

S20. Increasing the awareness of SE researchers about researcher bias can mitigate it (*e.g.,* by means of panels on researcher bias in SE, an ethical code for SE warning researchers against this kind of bias, or papers on researcher bias in SE).

S21. Guidelines for reviewers of SE conferences/journals to instruct them not to judge papers on the basis of the study results (*i.e.,* positive/negative results) can mitigate researcher bias.

S22. Ad-hoc negative-results conference tracks and ad-hoc negative-results journal issues can mitigate researcher bias.

S23. Replicating experiments can mitigate researcher bias.

Blind Data Extraction and Analysis

S24. Blind data extraction is a useful technique to mitigate researcher bias.

S25. Blind data analysis is a useful technique to mitigate researcher bias.

S26. The combined use of blind data extraction and analysis is useful to mitigate researcher bias.

S27. To deal with researcher bias, in my next experiment I'm going to use the following technique:

1. Blind data extraction.
2. Blind data analysis.

Figure 2: Statements, arranged by section, we included in the questionnaire.

Table 1: Characterization of the interviewees.

ID	Institution region	Academic position	Main research interest	Experience as an experimenter	Last published experiment
R1	Southeastern Europe	Assistant professor	Defect prediction	5-10 (years)	< 6 months
R2	Northern Europe	Ph.D. student	Human and social aspects of SE	1-5 (years)	< 18 months
R3	Northern Europe	Full professor	Mining software repositories	11-20 (years)	< 6 months
R4	Northern America	Associate professor	Agile software development	11-20 (years)	< 6 months
R5	Central Europe	Assistant professor	Software maintenance and evolution	5-10 (years)	< 3 years
R6	Southern Europe	Associate professor	Software economics and metrics	11-20 (years)	< 1 year
R7	Southern Europe	Assistant professor	Project and process management	11-20 (years)	< 1 year
R8	Southern Europe	Full professor	Collaborative software development	> 20 (years)	< 18 months
R9	Southern Europe	Full professor	Software economics and metrics	11-20 (years)	< 6 months

took part in the survey as the respondents.

3.2. Participants

In Table 1, we report some information about the interviewees—this information was gathered through the *Warm-up* part of the interview (see Figure 1). To guarantee the anonymity of the interviewees, we refer to each of them through an ID (from R1 to R9). Each interviewee had experience in performing experiments and, at the time of the interview, had published at least one experiment in one of the following SE high-quality venues: ICSE,[5] EMSE,[6] TSE,[7] and/or TOSEM.[8] The participants were quite heterogeneous in terms of location of their institution, academic position,

[5]International Conference on Software Engineering.
[6]Empirical Software Engineering.
[7]Transaction on Software Engineering.
[8]Transaction on Software Engineering and Methodology.

Table 2: Characterization of the respondents.

Characteristic	Values (Frequencies)
Institution county	Not provided (24), Brazil (5), Germany (4), Netherlands (3), Sweden (3), Canada (2), Spain (2), United States (2), Afghanistan (1), Australia (1), Estonia (1), Italy (1), Serbia (1), United Kingdom (1)
Academic position	Full professor (21), assistant professors (10), associate professor (10), Ph.D. student (4), post-doc (4), industry researcher (2)
Experience as an experimenter	11-20 years (17), 6-10 years (16), 1-5 years (11), > 20 years (7)
Last published experiment	< 6 months (28), < 3 years (23)
Kind of venue	Conference (32), journal (15), book chapter (1), others (2)
Kind of conducted experiments	Human-oriented experiment only (22), human- and technology-oriented experiment (17), technology-oriented experiment only (12)

main research interest, years of experience as an experimenter,[9] and date of the last published experiment. The interviewees were employed in academic institutions located in different regions throughout Europe and North America. At the time of the interview, three interviewees were full professors, two were associate professors, three were assistant professors, and one was a Ph.D. student. R8 (full professor in a Southern European institution) has more than 20 years of experience in conducting SE experiments and had published her last experiment less than 18 months before the interview. Other researchers (*e.g.,* R3, R4, R6, R7, and R9) had more than 10 years of experience in conducting SE experiments with their last experiment published less than one year before the interview. With the exception of R2 (the interviewee in a more junior position), the interviewees had more than five years of experience in conducting SE experiments. Only in one case (R5), the last experiment was published more than 18 months before the interview (but less the 3 years before the interview). The main research interest of the interviewees spanned across different sub-fields of SE, from human aspects to mining software repositories.

As for the respondents, we report some information about them—this information was gathered through the *Demographics* part of the questionnaire—in Table 2. As this table shows, most respondents (27) shared the location of the institution which

[9] We refer to an experimenter as a researcher conducting (or co-conducting) an experiment (human- or technology-oriented). To avoid misunderstandings, we made clear to the interviewees what we meant as an experimenter. Also, we made clear that we focused exclusively on experiments (*e.g.,* we were not interested in mining studies).

they worked for. These respondents worked for institutions located in 11 different countries. The most represented country was Brazil (with five responses). The respondents were, for the most part, senior researchers (21 full professors and 10 associate professors). Most respondents (40) had more than five years of experience in conducting experiments. More than half of the respondents (28) had published their last experiment less than six months before they filled in the questionnaire, while the remaining ones had published their last experiment within the last three years. The majority of the respondents usually published their experiments in conferences (32) and journals (15). As for the former, the most preferred venues were ESEM, ICSE, and ESEC/FSE.[10] As for the journals, the most preferred venues were EMSE, IST,[11] and TSE. The respondents have, for the majority, experience with human-oriented experiments only (22). Seventeen respondents have experience with both kinds of experiments, while 12 respondents have experience with technology-oriented experiments only.

3.3. Data Analysis

After transcribing the recordings of the interviews, we (*i.e.,* the first, third, and fourth authors) analyzed the transcripts by using a thematic analysis approach called template analysis, which is known to be flexible and fast [24]. Template analysis allows the investigators to develop a list of codes, each identifying a theme within the transcripts. The codes are arranged in a *template*—it usually is a hierarchical structure of codes—showing the relationships among themes, as defined by the investigators. In template analysis, the investigators start analyzing the transcripts by using an initial template. That is, they start attaching pre-defined codes, arranged in a template, to delimit portions of text in the transcripts related to the themes. As King [24] suggests, the best starting point for developing an initial template is the interview script. Accordingly, we developed our initial hierarchical template (see the non-bold text in Figure 3) from the interview script. As customary in template analysis, we revised the initial template during the analysis [24]. In particular, we renamed the second-level code *Presence of Researcher Bias* as *Presence of Researcher Bias and Clues* because we found portions of text about clues suggesting the presence of researcher bias. We concluded the analysis when any portion of text relevant to the goal of our interview study was coded and we agreed on the obtained template.

[10] Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering.

[11] Information and Software Technology.

- Experiment Planning
  - Researcher Roles
  - Threats to Validity
- Questionable Research Practices
  - Post-hoc Hypotheses
  - Post-Hoc Outlier Criteria
  - Flexible Reporting of Measures and Analyses
- Researcher Bias
  - Presence of Researcher Bias **And Clues**
  - Causes of Researcher Bias
  - Coping with Researcher Bias
    - Blind Data Extraction
      - Usefulness of Blind Data Extraction
      - Drawbacks of Blind Data Extraction
    - Blind Data Analysis
      - Usefulness of Blind Data Analysis
      - Drawbacks of Blind Data Analysis
    - Blind Data Extraction and Analysis
      - Effectiveness of Blind Data Extraction and Analysis
      - Fostering Blind Data Extraction and Analysis

Figure 3: Initial and final templates—we highlight in bold the text added to the initial template to obtain the final one.

To ease the thematic analysis of the transcripts, we used ATLAS.ti[12]—a tool for supporting qualitative data analyses, including template analysis.

As for the survey, we performed an exploratory data analysis of the answers. In particular, we visualized the results—*i.e.,* answers to the statements—by using stacked barplots. Each stacked barplot reported the absolute frequencies for each level of agreement about a statement.

4. Findings from the Interview Study and Survey

In this section, we present the findings emerging from the interview study according to the main themes identified by the first-level codes (*i.e., Experiment Planning*, *Questionable Research Practices*, and *Researcher Bias*) of the final template shown in Figure 3. We also support these findings by reporting excerpts of the related transcripts. We then triangulate these findings with those from the survey. In particular, we show the level of agreement of the survey respondents about the statements we

[12] atlasti.com.

built upon the findings from the interview study.

4.1. Experiment Planning

As Figure 3 shows, we defined two sub-themes within this main theme—namely, the roles of researchers in SE experiments and how they cope with threats to validity in their experiments.

Researcher Roles. It emerged from the interviews that, when conducting an experiment, there is a division of roles among the researchers involved in the experiment. Each researcher covers one or more roles (*e.g.,* one researcher is involved in the planning and execution of the experiment, another one extracts the metrics from the raw data, and so on). However, it seems that only one researcher takes care of data analysis (*i.e.,* one researcher plays the data analyst role). An excerpt from the interview with R6 follows:

> We [our research group] outlined the experiment design. The researchers from [other country] translated the experiment material into [other language] and carried out the experiment in [other country]. We then received the gathered data, some Excel files, and one of us executed the analysis.

As far as the survey results are concerned, most respondents (38), in their experience as experimenters, had more than one researcher involved in the data analysis (S1). In this case, we cannot support the finding from the interview study.

Threats to Validity. When we asked the interviewees to elaborate on the threats to validity, they provided a number of examples, but none of them mentioned researcher bias (accordingly, we could not define a corresponding statement in the questionnaire of the survey).

4.2. Questionable Research Practices

This theme includes three sub-themes (see Figure 3): the participants' perceptions of post-hoc hypotheses, post-hoc outlier criteria, and flexible reporting of measures and analyses (see Section 2.1).

Post-hoc Hypotheses. According to the interviewees, the post-hoc hypotheses practice should not lead to researcher bias as long as *(i)* the researchers clearly report that such hypotheses are formulated in retrospect, or *(ii)* it is possible to ground such hypotheses on prior work (thus, there is no need to make clear that such hypotheses are post-hoc). Regarding *(i)*, R5 said:

> In this case, first of all I am not sure we can talk about formulating hypotheses because you are already looking at the data of an experiment [...] In general, I don't think there is anything wrong with that if, and I think it is completely sound, if you explicitly say that it is an unexpected result when reporting this result. This is different from saying «we wanted to investigate this and we found that it is supported by the data.»

As for the point *(ii)*, R3 told us:

> Of course, there's the fact that, the hypothesis should be grounded on prior work. If you can ground something to solid prior work, then it doesn't really matter whether it was after the fact.

Furthermore, it seems that the post-hoc hypotheses practice could be a means to get new insights into the investigated phenomenon, which researchers had not thought about when the study was planned. On this matter, R4 said:

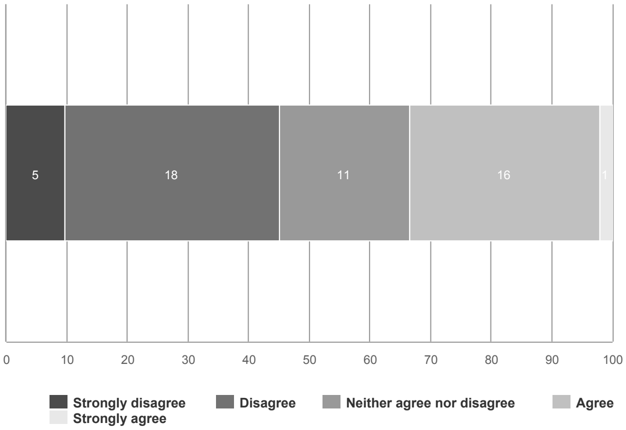
> It [a post-hoc hypothesis] emerged from the data and inevitably happens. When you look at the data, you may have, you may think of new insights that you haven't thought about because there is information that was not anticipated. [...] Sometimes there are research methodologies that don't even assume any questions, they are completely totally exploratory. So let's suppose that you have a set of questions, and you want to answer them first. After you answer those questions, then you see some other patterns in your data and then, in the next iteration, you formulate a set of other questions that maybe you can answer based on the same data. This is completely okay but it's not the same as fishing.

On the other hand, the majority of the respondents (23) believed that formulating post-hoc hypotheses leads to researcher bias even when they are disclosed as being formulated in retrospect in the reporting of the experiment (see Figure 4a). A higher number of respondents (27) believed that, even when grounded on prior work, post-hoc hypotheses still lead to researcher bias (see Figure 4b). However, most respondents (42) saw post-hoc hypotheses as a mean to get new insights into the phenomenon under study (see Figure 4c).

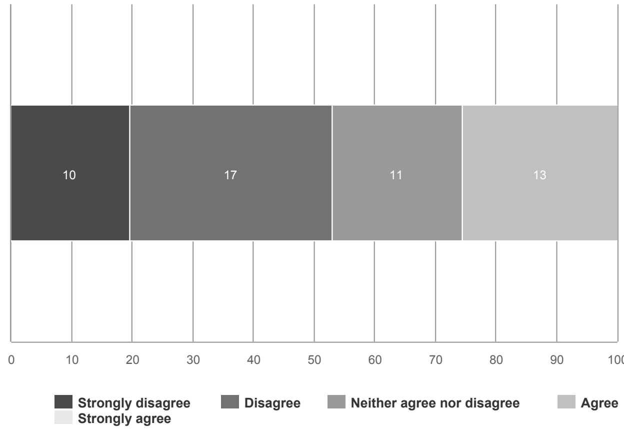
Post-hoc Outlier Criteria. The interviewees seemed to believe that this practice should be avoided because it potentially leads to researcher bias, though not necessarily. To this extent, R5 told us:

> Looking at the results and then removing outliers could sometimes be sensible, but I think the bias would be too strong.

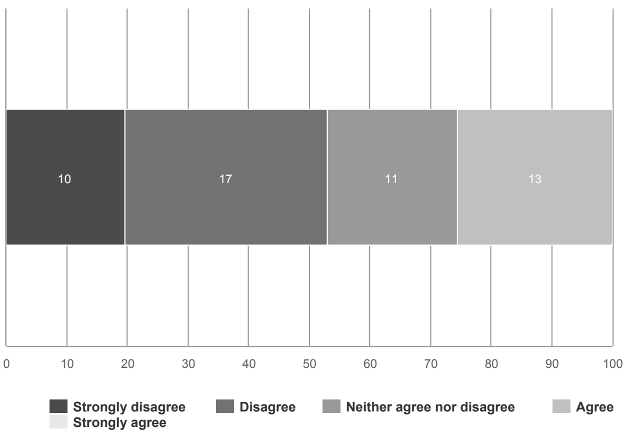
In case researchers apply the post-hoc outlier criteria practice, the interviewees agreed that they should declare the use of this practice in the paper by providing, for example, the following information: *(i)* the results before and after removing outliers;



(a) Agreement with S2 (*"The use of the post-hoc hypotheses practice does not lead to researchers bias as long as the researchers clearly report that these hypotheses are formulated in retrospect"*).



(b) Agreement with S3 (*"The use of the post-hoc hypotheses practice does not lead to researcher bias as long as it is possible to ground such hypotheses on prior work"*).

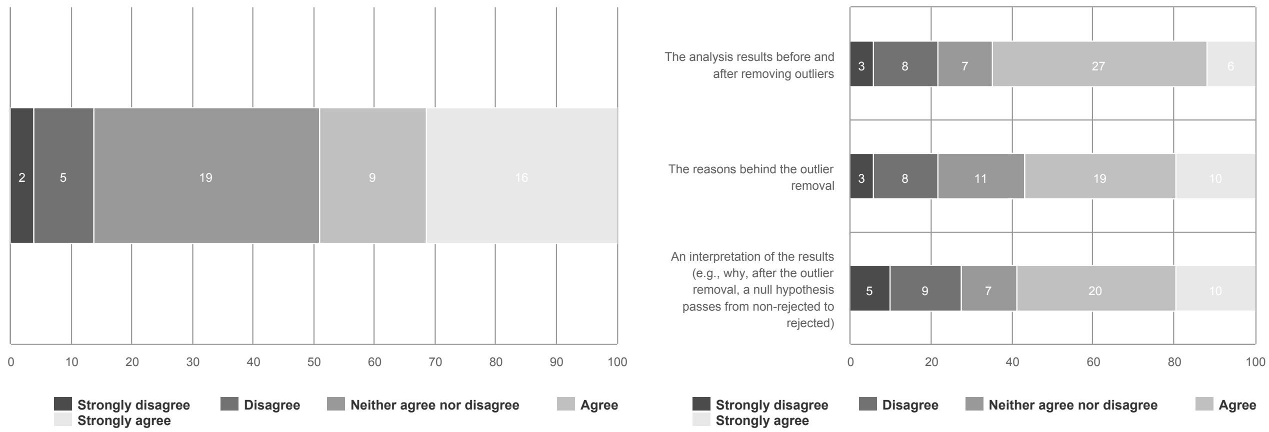


(c) Agreement with S4 (*"The post-hoc hypotheses practice could be a means to get new insight into the studied phenomenon, which researchers had not thought about when the study was planned"*).

Figure 4: Results regarding the post-hoc hypotheses practice.

(ii) the reasons behind the outlier removal; and *(iii)* an interpretation of the results (*e.g.*, why, after the outlier removal, a null hypothesis passes from non-rejected to rejected). On this matter, we report R4's comment:

> As long as you declare the results and you present maybe both of them [before and after the



(a) Agreement with S5 (*"The post-hoc outlier criteria practice should be avoided because it potentially leads to researcher bias"*)

(b) Agreement with S6 (*"The post-hoc outlier criteria practice does not lead to researcher bias as long as the researcher declares the use of this practice in the paper by providing the following information."*).

Figure 5: Results regarding the post-hoc outlier criteria practice.

> outlier removal], depending on how other factors influence your interpretation. Maybe there are other things that you discovered during your data analysis that justifies that decision. But as long as you declare them, I mean that is one of the purposes of the peer review, the reviewers can also decide which one is, whether that decision was sensible or not.

As for the survey, the majority of the respondents (25) agreed that the post-hoc outlier criteria practice leads to researcher bias. However, 19 of them neither agreed nor disagreed with the statement reported in Figure 5a. Nevertheless, the respondents believed that disclosing additional information regarding the outlier removal does not lead to researcher bias (see Figure 5b). In particular, the majority believed that what needs to be reported is: the results with and without the outliers (33); the reasons for different results once outliers are removed (30); and the reasons behind the outlier removal (29).

Flexible Reporting of Measures and Analysis. Based on interviewees' experience, when researchers can choose among equivalent statistical hypothesis tests (*e.g.,* paired t-test and Wilcoxon signed-rank test), the results (*i.e.,* p-values) are not so different. R8's thought on this point follows:

> It's true that there are a lot of statistical hypothesis tests and there are a lot of variants as well, when using statistical packages we are spoilt for choice, but in my experience they don't vary so much.

Furthermore, according to R3, if a statistical hypothesis test revealed a significant difference (*e.g.,* p-value slightly less than $\alpha = 0.05$) that an equivalent test did not (*e.g.,* p-value greater than $\alpha = 0.05$), that difference would be probably negligible. In other words, the effect size would show the true impact of that difference, so having or not a significant difference would not matter:

> It [using a statistical hypothesis test or an equivalent one] doesn't really impact the results very much. It's a very very tiny difference, at least what I have seen. It doesn't change from .04 to .0004, or something. I mean you might, if you again use this magical threshold of .05, then it might matter. But if you report the effect sizes, then it really doesn't. The effect sizes sort of reveal the true impact.

As for the practice of using several variants of a measure and then reporting only the variants that give the strongest results, it is perceived as a bad practice. The researchers should discuss any variant of that measure in the paper. In this respect, R4 said:

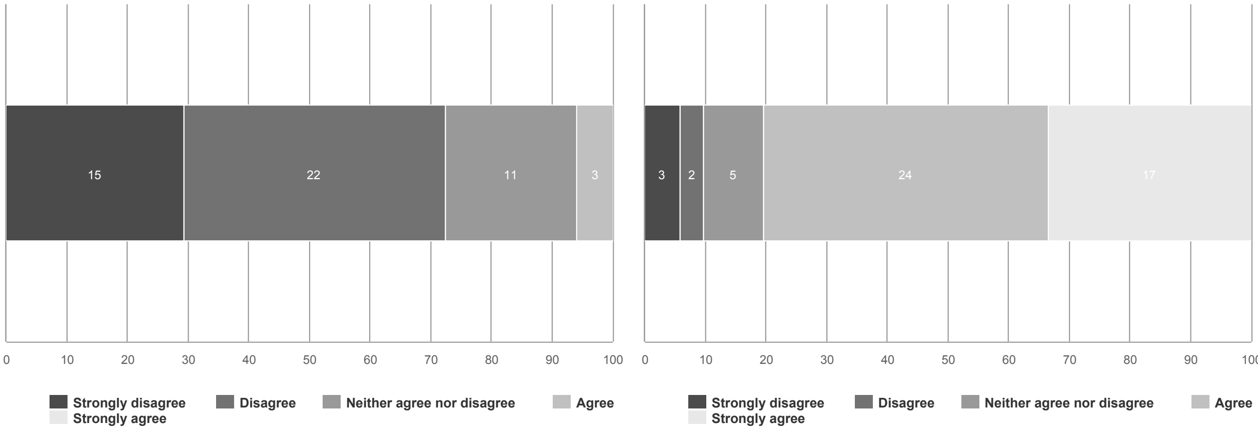
> Yeah I think that is a no, in general. If you've done [flexible reporting of measures], there needs to be a discussion of how your attempt to triangulate the results with different measures failed. That should be part of the discussion and it's part of the validity threats that you have.

As for the respondents, most of them (37) disagreed that reporting the results of a statistical test, rather than those of an equivalent one, does not matter because the difference (estimated by using an effect size measure) would be probably negligible (see Figure 6a). On the other hand, the majority of respondents (41) agreed that the flexible reporting of measures practice leads to researchers bias (see Figure 6b).

4.3. Researcher Bias

This theme has three sub-themes (see Figure 3): the presence of researcher bias in experiments and clues suggesting such a presence; causes of researcher bias; and strategies to cope with researcher bias.

Presence of Researcher Bias and Clues. From the interviews, it emerged that researcher bias affects the SE community. Although the interviewees did not have proofs about the presence of researcher bias in SE, they pointed out four clues suggesting its presence: *(i)* researcher bias affects any community (*e.g.,* medicine or psychology); *(ii)* when reviewing papers, it is not rare to suspect authors biasing the results; *(iii)* whoever could unconsciously bias the results based on her expectations; and *(iv)* there are sometimes inconsistent results among studies investigating the same constructs. On the points *(i)* and *(ii)*, R4 stated:



(a) Agreement with S7 (*"If a statistical hypothesis test (e.g., paired t-test) revealed a significant difference that an equivalent test (e.g., Wilcoxon signed-rank test) did not, that difference (estimated by using an effect size measure) would be probably negligible, so using a test rather than another one does not matter").*

(b) Agreement with S8 (*"The flexible reporting of measures practice leads to researcher bias"*).

Figure 6: Results regarding the flexible reporting of measures and analysis practice.

> I think it [researcher bias] must be happening because it's probably happening in every community. But I'm not sure. I mean I think, in terms of my review work, when things are suspicious, it's usually obvious and it's usually not just from one reviewer picking on them, rather, multiple reviewers do and it's only because, the researchers actually let it be understood in the paper.

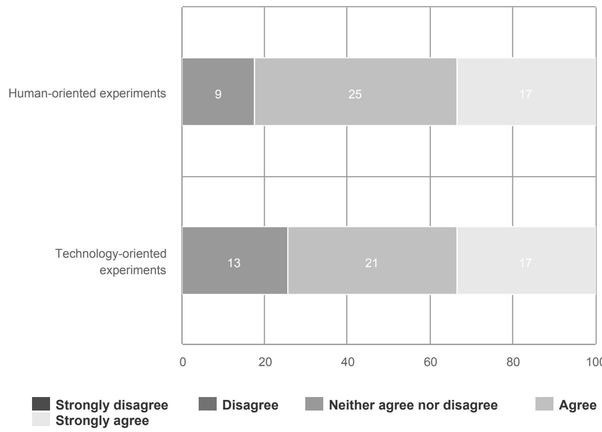
As for the point *(iii)*, R3's thought follows:

> I guess everyone that does experiments is somehow biased because you know that negative results cannot be published and it probably, sort of unconsciously, alters your actions.

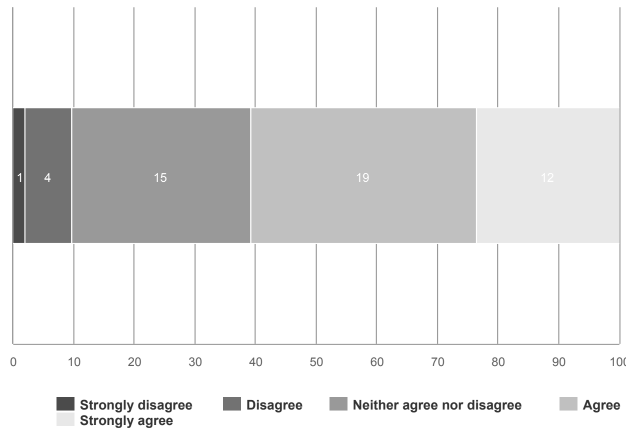
On the last point, R8 said:

> That is, if I see that a given result isn't confirmed [by another study], then it is a clue of researcher bias.

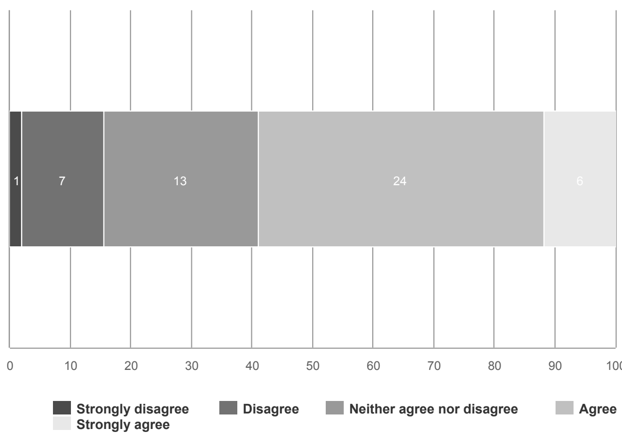
These findings seem to be confirmed in the survey. In particular, as shown in Figure 7a, the presence of researcher bias in SE experiments appears to be independent of the experiment kind; and its presence seems to be perceived as widespread as in other research fields (see Figure 7b). From their experience as reviewers of SE experiments, the majority of respondents (30) had suspected that researchers bias the results of their experiments (see Figure 7c). Also, most respondents (43) agreed that researchers can unconsciously bias the results based on their expectations (see



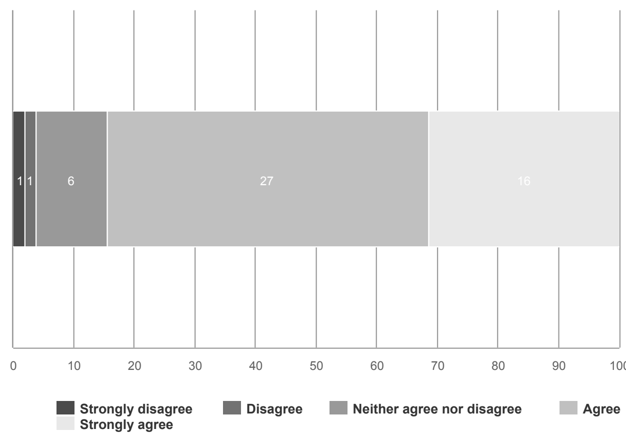
(a) Agreement with S9 (*"Researcher bias is present in SE experiments of the following kind"*.)



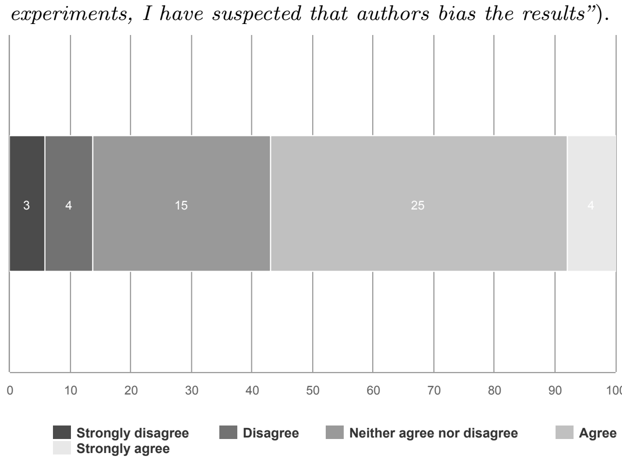
(b) Agreement with S10 (*"Researcher bias affects the findings from experiments in the software engineering research field as much as other research fields (e.g., medicine or psychology)"*).



(c) Agreement with S11 (*"When reviewing papers reporting SE experiments, I have suspected that authors bias the results"*).



(d) Agreement with S12 (*"Researchers can unconsciously bias the results based on their expectations"*).



(e) Agreement with S13 (*"Researcher bias is one of the reasons for inconsistent results among studies investigating the same constructs."*).

Figure 7: Results regarding the presence of researcher bias and clues.

Figure 7d). Finally, most respondents (29) agreed that researcher bias is one of the reasons for inconsistent results among similar studies—*i.e.,* studies addressing the same constructs (see Figure 7e).

Causes of Researcher Bias. Three causes of researcher bias emerged from the interviews. First, interviewees believed that *negative-results papers are usually rejected*. This would lead researchers to bias their results (*e.g.,* transforming non-significant results into statistically significant ones). R2 said:

> I think the main reason to that [researcher bias] is there is no acceptance for reporting the negative results. You are a researcher and your responsibility is just to explore the phenomenon, whether it is in favor of your hypothesis or it's against your hypothesis you should report it, but I've personally felt like there is no in general acceptance for that.

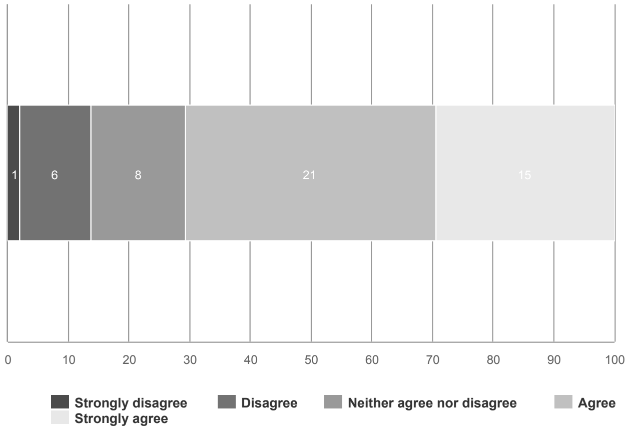
Second, the *pressure of publishing papers* can lead researchers to (unconsciously or consciously) bias the results. R5 said:

> Especially young researchers, for example Ph.D. students, that carry out and are therefore responsible for the experiment, may tend to have high expectations on what they have developed or towards the hypothesis being verified, to the point that, even unconsciously, they may tend to guide the experiment towards a certain expected result. I am quite confident to say that, although not always, this occurs especially with novice experimenters that are more eager for publications and may therefore be led to experimenter bias.

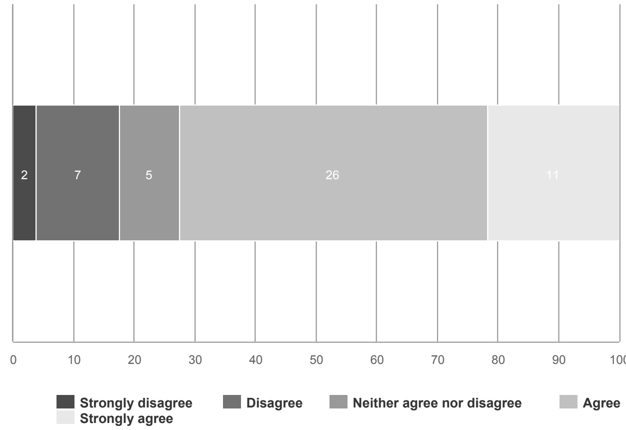
Third, it seems that *revision processes of SE conferences/journals are focusing too much on the empirical assessment*, rather than on the contributions of the ideas to the body of knowledge. Thus, researchers would be led to bias their studies by making the results more publishable. R5 told us:

> I think that the main problem of several review processes is that they are highly based on the empirical aspect and much less on the novelty of the ideas. So in spite of you propose an interesting and novel idea that several other researchers can build on, if the experimental results are not strong enough you are likely to receive a comment like "okay nice idea but ...". On the other hand, if a study is empirically perfect, from the point of view of the design and results, but has very limited novelty, it's difficult that it will be rejected.

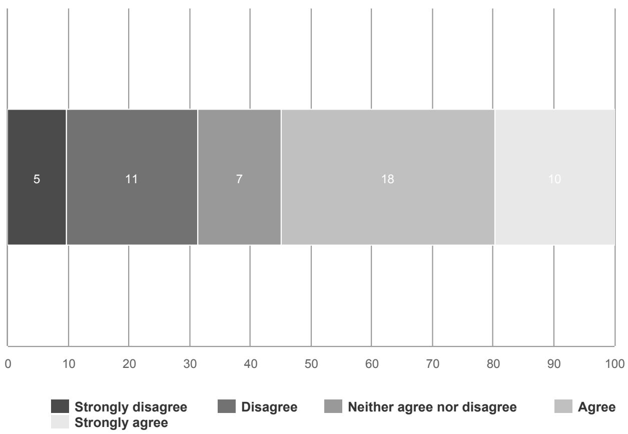
The three causes of researcher bias identified from the interview study were all endorsed by the larger part (between 28 and 37) of the respondents. In fact, for each cause the greater part of the respondents either strongly agreed or agreed (Figure 8). This finding is slightly less pronounced on the statement concerned the revision processes of SE conferences/journals (S16).



(a) Agreement with S14 (*"The rejection of papers reporting negative/null results leads some researchers to bias the results (e.g., transforming non-significant results into statistically significant ones)"*).



(b) Agreement with S15 (*"The pressure of publishing papers leads some researchers to (unconsciously or consciously) bias the results"*).



(c) Agreement with item S16 (*"The revision process of SE conferences/journals is focusing too much on the rigor of the empirical assessment rather than on the novelty of contributions"*).

Figure 8: Results regarding the causes behind researcher bias.

Coping with Researcher Bias. The interviewees suggested seven strategies to cope with researcher bias. First, the use of *pre-registration* in SE conferences/journals (see Section 2.2). This should prevent negative-results papers from being rejected. Moreover, pre-registration increases both credibility of study results and study replicability [19]. Accordingly, researchers should be less prone to bias the results of their studies. In this respect, R5 said:

> Personally, I have an idea. It doesn't relate to the experimental design, rather to a discipline. It consists of having dedicated tracks of a conference or sections of a journal where authors don't submit the results of an experiment, but the experiment they plan to carry out.

Second, fostering *open data policies* in SE conferences/journals. This means not only making the gathered data publicly available, but also the analysis scripts of the study. Such open data policies should allow reviewers (and any other researcher) to repeat the data analysis of that study so attributing credibility to study outcomes and increasing the replicability of the study. Therefore, researchers should be discouraged from biasing their studies. An excerpt from the interview with R1 follows:

> Another thing could be publishing all the analyses together with the data. But then that implies during the review process that, as a reviewer, I have to go and take a look at the analysis as well.

Third, *duplicate data analysis*. That is, two researchers analyze the same data with their own scripts without interacting with one another. Then they exchange the scripts and data to cross-check them. Finally, the results of the data analysis are compared. R5 told about this kind of data analysis (she was using at the time of the interview), which should mitigate the unconscious bias of researchers involved in the data analysis.

> The only thing I do, from about three years, is that data is always analyzed independently by two researchers. Next, they exchange the scripts and cross-check them. They exchange the data and cross-check them as well. Finally, they compare their conclusions.

Fourth, *means for increasing the awareness* of researcher bias in SE. For example, panels on researcher bias in SE, an ethical code for the SE research field to warn researchers against this kind of bias, or papers on researcher bias in SE studies. Therefore, by increasing the awareness of researcher bias, researchers should be warned against this kind of bias. On this matter, R6 said:

> Fostering panels and discussions on this [researcher bias], conducting surveys and studies, like the one you are conducting, to understand the status of the community.

Fifth, *guidelines for reviewers* in SE conferences/journals. These guidelines should instruct the reviewers not to judge papers based on the study results (*i.e.,* positive/negative results). As a consequence, researchers would bias the study results less because having a paper reporting positive/negative results would be equally valid. On this point, R4 said:

> Perhaps review guidelines may also help, in the sense that you instruct the reviewers, specifically not to bias their reviews only if the results are favorable to the hypothesis of the researchers.

Sixth, *ad-hoc research tracks* in SE conferences (or ad-hoc issues in SE journals). For example, specific tracks for papers reporting negative results or specific tracks for studies having a not so strong empirical assessment. Such a kind of track should lead researchers not to bias their results to have more publishable results. On this point, R7 said:

> Having various publication-levels where non-rigorous studies carried out by research groups or companies can be published in prestigious journals.

Seventh, *replicated experiments* because the more the results of a study are confirmed by replications, the lower the likelihood of researcher bias is. R8's thought follows:

> I trust when the results are confirmed by more studies carried out by researchers that are not co-authors. I don't think only one paper is enough. I don't confide in the results of only one paper. Of course, this doesn't mean that single studies are conducted incorrectly or are error-prone, it simply impacts on generalizability.

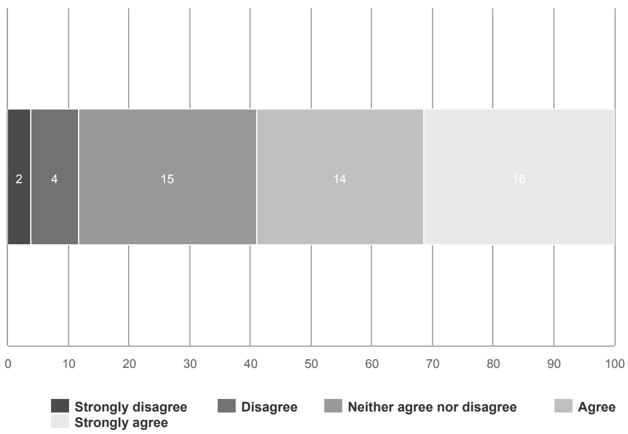
The majority of respondents (between 43 and 44) agreed that actions based on experiment replication (see Figure 9g), as well as actions regarding data analysis (see Figure 9c) and sharing of experimental material (see Figure 9b), can mitigate researcher bias in SE experiments. A lower number of respondents (between 29 and 39) agreed that actions targeting community efforts can mitigate researcher bias. Among these actions, there are initiatives to increase the awareness about researcher bias (see Figure 9d), peer-review guidelines (see Figure 9e), and initiatives within conference and journal steering groups to set up experiment pre-registration (see Figure 9a) and negative-results tracks and special issues (see Figure 9f).

Besides the above-mentioned strategies to cope with researches bias, we asked the interviewees their thoughts on two further strategies, blind data extraction and blind data analysis, used alone and together. In the following subsections, we report the findings concerning the sub-themes for blind data extraction, blind data analysis, and both these strategies. We also triangulate these findings with those from the survey.

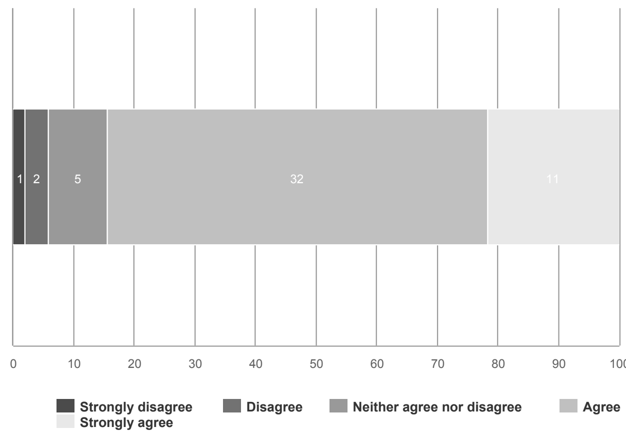
4.3.1. Blind Data Extraction

Two sub-themes were defined for this theme (see Figure 3): usefulness and drawbacks of blind data extraction in SE experiments.

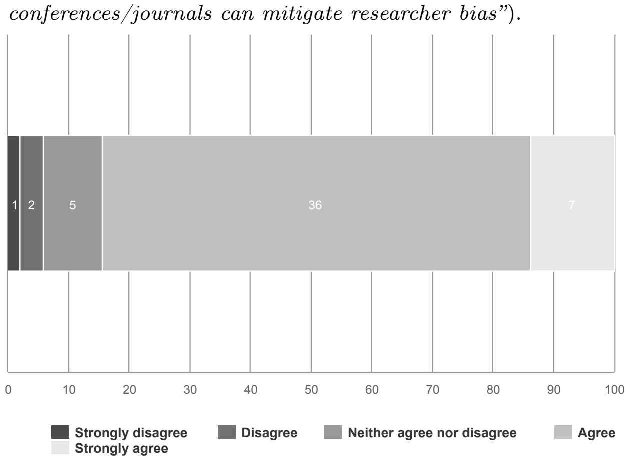
Usefulness of Blind Data Extraction in SE. It emerged from the interviews that blind data extraction could be a useful technique to mitigate researcher bias



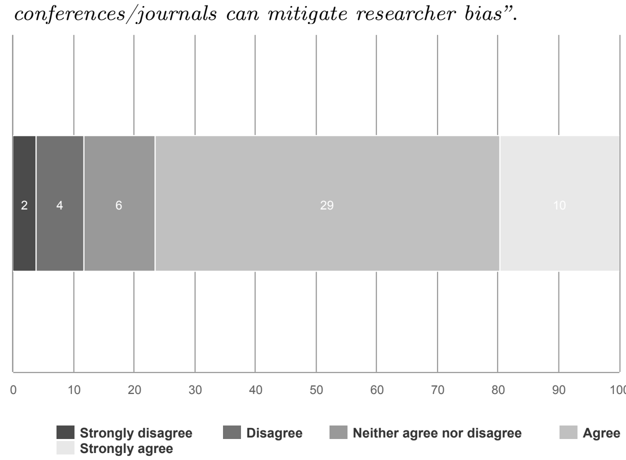
(a) Agreement with S17 (*"The use of pre-registration in SE conferences/journals can mitigate researcher bias"*).



(b) Agreement with S18 (*"Fostering open data policies in SE conferences/journals can mitigate researcher bias"*.

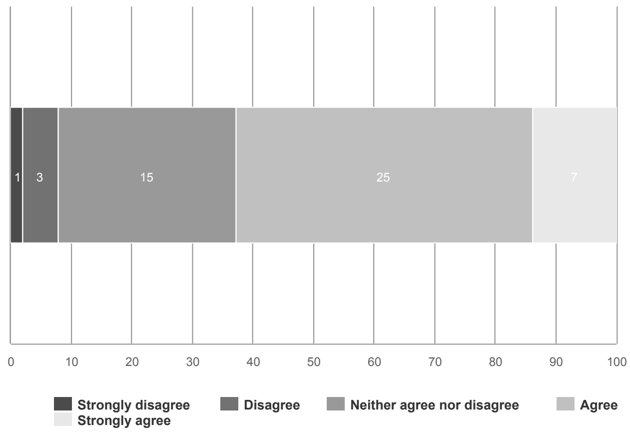


(c) Agreement with S19 (*"The use of duplicate data analysis can mitigate researcher bias"*.

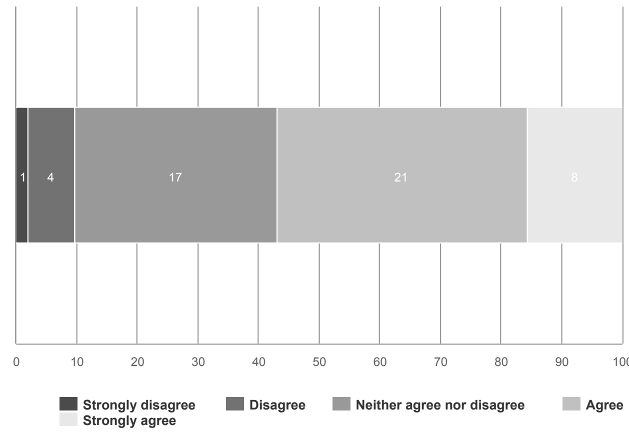


(d) Agreement with S20 (*"Increasing the awareness of SE researchers about researcher bias can mitigate it (e.g., by means of panels on researcher bias in SE, an ethical code for SE warning researchers against this kind of bias, or papers on researcher bias in SE)"*).

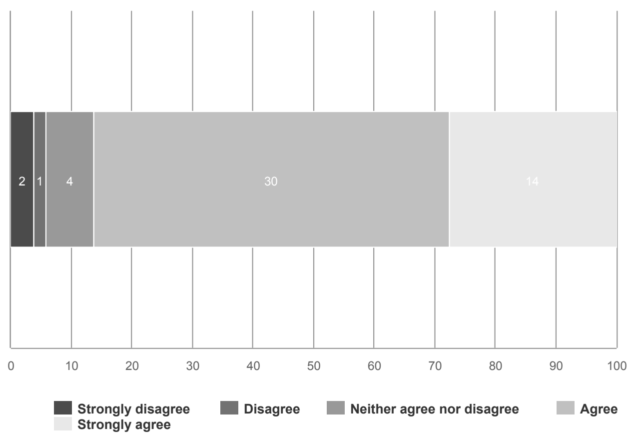
Figure 9: Results regarding the actions to counteract researcher bias (the figure continues in the next page).



(e) Agreement with S21 (*"Guidelines for reviewers of SE conferences/journals to instruct them not to judge papers on the basis of the study results (i.e., positive/negative results) can mitigate researcher bias"*).



(f) Agreement with S22 (*"Ad-hoc negative-results conference tracks and ad-hoc negative-results journal issues can mitigate researcher bias"*).



(g) Agreement with S23 (*"Replicating experiments can mitigate researcher bias"*).

Figure 9: Results regarding the actions to counteract researcher bias (the figure continues in the previous page).

because, even when extracting the metrics, a researcher could favor a given treatment based on her expectations. In other words, if the data extractor (*i.e.,* the person who is responsible for extracting the metrics from the raw dataset) is aware of research design elements (*e.g.,* treatment assignment), then the likelihood of influencing the results towards a given treatment is higher. This is why having blinded extractors would lessen the likelihood of influencing the results. On this point, R3 said:

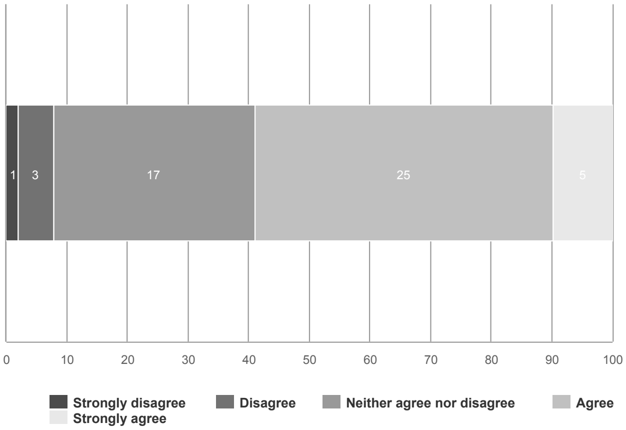
> Yeah, I think it [blind data extraction] sounds like a good idea. I believe that they [the researchers] may apply bad practices of statistical analysis but actually I believe more that one does it, consciously or unconsciously, while they code the data, or do it even before running the experiments because the researcher knows what treatment is and what the control is. I think that's a good idea that labels are removed and someone else transforms the data.

As far as the survey is concerned, the majority of the respondents (30) agreed that blind data extraction can mitigate researcher bias, whereas only a few (four) disagreed with such a statement (see Figure 10a).

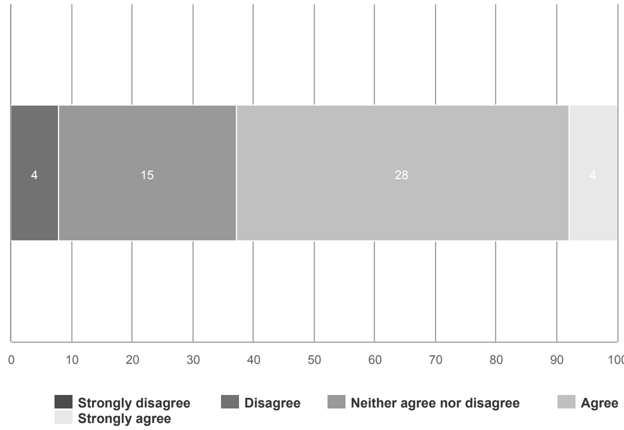
Drawbacks of Blind Data Extraction. As for the drawbacks of blind data extraction, the interviewees pointed out that the implementation of blind data extraction requires at least two people: an individual (*i.e.,* the study executor) responsible for executing the experiment and another individual (*i.e.,* the data extractor) with the necessary skills to extract the metrics from the raw dataset. The latter has to be blinded to research design elements. This seems to be little feasible when both study executor and data extractor belong to the same research group—guessing or finding out about hidden information (*e.g.,* research hypotheses) would be more likely when both executor and extractor belong to the same research group. Therefore, to implement blind data extraction, it is preferable to have: *(i)* a research collaboration between two research groups where the experimenter and the extractor are not part of the same group; or *(ii)* an external expert that takes care of the metric extraction. In this respect, R8 stated:

> I think it [blind data extraction]'s complicated. In many cases it's you and your Ph.D. student, do you really think that your student isn't aware of who did certain things? [...] Maybe it can work in a joint experiment where you have a large group of people collaborating from various independent research groups. On the other hand, within the same group it is applicable in theory because you have several researchers involved, however it becomes an "open secret" as everyone is aware of what is going on. How much would it work within the same group?

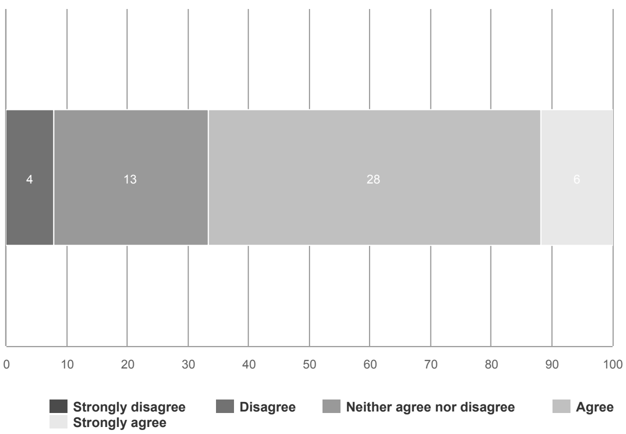
It is worth noting that R5 had already used blind data extraction. In particular, she (and her colleagues) had involved some experts to extract metrics from a raw dataset:



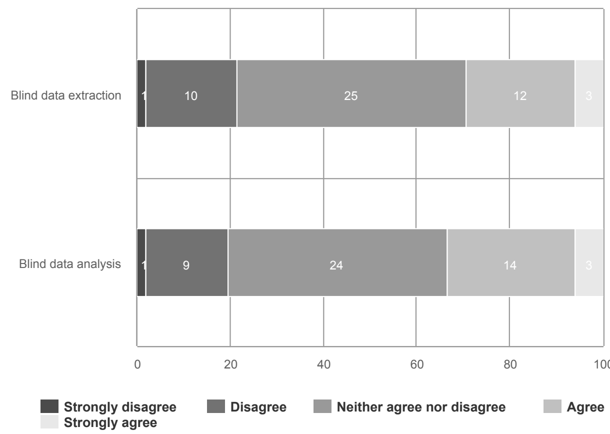
(a) Agreement with with S24 (*"Blind data extraction is a useful technique to mitigate researcher bias."*).



(b) Agreement with with S25 (*"Blind data analysis is a useful technique to mitigate researcher bias."*).



(c) Agreement with S26 (*The combined use of blind data extraction and analysis is useful to mitigate researcher bias*").



(d) Agreement with S27 (*To deal with researcher bias, in my next experiment I'm going to use the following technique."*).

Figure 10: Results regarding blind data extraction and analysis.

> Well now that you have mentioned it [blind data extraction], we actually have done it on two papers in the past that I had forgotten about. What we did was to gather the artifacts produced by the participants and then give all to external people who evaluated the artifacts. [...] Yes, I think this is surely useful.

4.3.2. Blind Data Analysis

Two sub-themes were defined for this theme (see Figure 3): usefulness and drawbacks of blind data analysis in SE experiments.

Usefulness of Blind Data Analysis. According to the interviewees, blind data analysis is a useful technique to mitigate researcher bias. This is because a blinded analyst (*i.e.,* an analyst unaware of research design elements) would perform the data analysis more objectively than an analyst aware of research design elements. On this matter, R7 said:

> It can be a means for a more objective analysis because it's human to be inclined to one's proposals and expectations. This can be thus an involuntary contribution, either positive or negative, that a researcher provides.

As for the respondents, the majority of them (32) agreed that blind data analysis is a useful technique to mitigate researcher bias. Only a few (four) disagreed with such a statement (see Figure 10b).

Drawbacks of Blind Data Analysis. Similarly to blind data extraction, the drawback of blind data analysis is that at least two researchers are needed—the former conducts the study and sanitized the dataset, while the latter performs the data analysis on the sanitized dataset. Moreover, it is preferable (as for blind data extraction) that the researchers do not belong to the same research group. In this respect, R8 said:

> It's similar to blind data extraction. That is, if you are conducting a joint experiment, you can apply blind data analysis.

4.3.3. Blind Data Extraction and Analysis.

We defined three sub-themes for this theme: effectiveness of blind data analysis and extraction in coping with researcher bias, strategies to foster the adoption of blind data analysis and extraction in SE experiments, and intention to use blind data analysis and extraction.

Effectiveness of Blind Data Extraction and Analysis. From the interview study, it emerged that researcher bias could arise even if blind data extraction and analysis are applied together. That is, using both blind data analysis and extraction is considered a way to mitigate researcher bias (rather than a way to remove it). In fact, researcher bias could arise not only during the metric extraction and analysis phases but also during the execution of the experiment itself. Below, we report R3's answer when we asked if the combination of blind data extraction and blind data analysis was enough to cope with researcher bias:

> Most likely not. Like I said previously, the step before where you set up and where you run the experiment also introduces some [bias].

The respondents found that the combined use of blind data extraction and analysis can be considered an appropriate technique to mitigate researcher bias (see Figure 10c). The majority of the respondents (34) agreed that blind data analysis is a useful technique to mitigate researcher bias, while four disagreed.

Fostering Blind Data Extraction and Analysis. The interviewees suggested a number of strategies to ease the adoption of blind data extraction and analysis in SE. The first strategy is a *policy* for conferences/journals similar to the double-blind peer-review one. That is, this policy would consist of requiring that any submitted experiment to that conference/journal had to use blind data extraction and analysis. However, this strategy is not always feasible, as the same interviewees observed, due to the following reasons: *(i)* the reviewers cannot make sure the authors of a paper have really used blind data extraction and analysis; *(ii)* researchers, who are not involved in research collaborations, would be harmed by this policy; and *(iii)* empirical evidence on the effectiveness of blind data extraction and analysis in SE studies is necessary to foster conferences/journals to adopt this policy. Regarding the point *(i)*, R1 said:

> For example, how can I understand if someone does a blind data analysis or not? I cannot.

On the point *(ii)*, R8 said:

> In most cases, you have a [research] group that works independently... it does not involve several units, or you have a group made up of Ph.D. student and supervisor. In this case, how do you distinguish the roles and introduce any blinding in the process?

As for the last point, R4 said:

> The conference committees won't do it [that policy] without any evidence that it's gonna be effective, just because it sounds like a good idea. Then, if there is enough evidence that it's a good idea, then maybe some conferences will start using it [that policy].

The second strategy to foster the use of blind data extraction and analysis is a *third-party service provider* that takes care of metric extraction and data analysis blindly. For example, the researchers conduct the experiment and, when needed, sanitize the raw dataset (*e.g.,* it removes any label to the treatments). Then they submit the raw dataset to this service provider, which extracts the metrics and then analyzes the data. After analyzing the data, the service provider sends the results to the researches. In this respect, R5 said:

> An example could be an online service for data analysis where each participant, at the end of the [experimental] task, uploads its data on that platform and then someone else performs the data analysis. So who carries out the experiment does not interact with or manipulate the data, rather only acknowledges the results of the analysis. Clearly, this is costly and not easy to be realized.

This strategy also has its drawbacks. As pointed out by R5, it is not easy to realize such a system. Also, the researchers should trust the service provider as well as the people that perform blindly the data extraction and analysis. Furthermore, it would most likely introduce extra costs. The third strategy consists of a *guideline* for applying blind data extraction and analysis in SE. R6 told us:

> Someone should try to give guidelines on how to put them [blind data extraction and analysis] in practice.

Finally, *empirical evidence* on the effectiveness of blind data extraction and analysis in SE would foster the adoption of these blind techniques. In this respect, R4 said:

> It would be nice if there could be some pilots or meta-studies that demonstrate how blind analysis and extraction change the results in either way, in favor or against the researcher's hypothesis.

Intention to Use Blind Data Extraction and Analysis. All interviewees stated they would take into account blind data extraction and analysis for their experiments. For example, R8 stated:

> If I have to participate in a large joint experiment between several research groups, I can take this into account when assigning the roles, why not! Instead of doing everything myself.

When we asked whether the respondents would use blind data extraction and/or analysis in their next experiment (see Figure 10d), the majority of the respondents were on the fence (25 for blind data extraction, 24 for blind data analysis). A lower number of respondents was willing to use blind data extraction (15) and blind data analysis (17) in their next experiment, while 11 respondents would not use blind data extraction and 10 will not use blind data analysis in their next experiment.

5. Discussion

In this section, we first discuss the results from both studies we presented in this paper and then the limitations of these studies.

5.1. Overall Discussion

Studies on researcher bias and its mitigation have a longstanding tradition in the natural and medical sciences. For example, physicists employ sophisticated blinding techniques to their data tailored to specific types of investigation [25]; medical researchers use double-blind randomized clinical trials as the standard way to avoid bias [26]. In the SE research field, the discourse on QRPs and RB mitigation started to appear in 2014–2015 in the work by Jørgensen *et al.* [15] and Shepperd *et al.* [16, 27]. In this section, we present the recommendations of our research. Some recommendations are intended for SE researchers while others are intended for the boards of SE research outlets. These recommendations are based on an *introspection* within our SE community and represent a first step towards the level of sophistication and awareness observed in other research fields.

The results of both interview study and survey support those by Jørgensen *et al.* [15] and Shepperd *et al.* [16]—*i.e.,* researcher bias affects SE experiments. According to the respondents, the different kinds of experiments (*i.e.,* human- and technology-oriented) seem to be equally affected by researcher bias. Also, it seems to be widely accepted that researcher bias is an unconscious phenomenon that needs to be addressed to improve the generation and solidification of scientific knowledge, and to avoid a methodological crisis (*i.e.,* the impossibility to reproduce experimental results [28]).

According to the interviewees, the formulation of post-hoc hypotheses should not be considered a QRP as long as the researcher explicitly mentions their use or it is possible to ground such hypotheses on prior work. On the contrary, the majority of the respondents consider post-hoc hypotheses to lead to researcher bias even when such hypotheses are disclosed and grounded on the literature. However, from both interview study and survey, it seems that this practice can be used to gain new insights into the investigated phenomenon (*e.g.,* for further studies). Based on these results, we can delineate the following recommendation:

❧ *Research hypotheses, generated after looking at the results of a study, need to be carefully disclosed by researchers. The investigation of such hypotheses can be the subject of follow-up studies.*

This recommendation is also inline with those Jørgensen *et al.* [15] delineated for SE researchers. In particular, the authors wrote: "make it clear whether a hypothesis was stated in advance or derived after looking at the data (exploratory hypothesis to be tested in follow-up studies)."

According to the results from both interview study and survey, the post-hoc outlier removal practice is not always questionable. It is considered acceptable if the researchers provide the results after and before the outlier removal, justify the outlier removal, and discuss the causes behind possible differences. Existing guidelines for evaluating SE experiments (*e.g.,* [29]) require authors to provide a clear outlier dropout analysis, which is particularly relevant for researchers interested in integrating the results of similar experiments (*e.g.,* meta-analysis). Accordingly, we can draw the following recommendation:

❧ *Researchers should have dedicated sections to report why and how outliers are removed, and how the results are impacted. Make the results (and possibly the dataset), before the outlier removal, available.*

Although we observed that the post-hoc outlier removal practice is not always considered questionable, the results from both studies suggest avoiding the use of this practice. In other words, researchers should still define the inclusion/exclusion outlier criteria in advance [15]. However, if a researcher faces a situation in which the use of the post-hoc outlier removal practice is reasonable, she should follow the above-mentioned recommendation.

The flexible reporting of measures is strongly perceived to lead to researcher bias in both studies. We make our the recommendation by Jørgensen *et al.* [15] to report on all measures and extend it as follows:

❧ *Researchers should disclose all measures in the paper and share the results for the measures they cannot include in the paper (e.g., for space reasons) by using an appendix or a replication package.*

Both interviewees and respondents saw the potential of blinding (both when extracting and analyzing data) and, to some extent, were favorable to use it. Although useful for mitigating researcher bias, blind data extraction and analysis do not solve the problem. In fact, as the interviewees suggested, blind data extraction and analysis are more effective when the key roles (*e.g.,* study executor and data extractor) are taken up by people that do not belong to the same research group. Our recommendation follows:

❧ *Researchers should consider blind data extraction and analysis especially if they can involve external experts, or collaborate with other research groups to have external researchers, who take care of blind data extraction and analysis.*

Involving external experts or collaborating with other research groups is not always possible. A simple form of blind data analysis can be achieved within the same research group by relabelling the experimental groups with non-identifying terms to hide the actual treatments from the data analyst [6, 8]. To mitigate researcher bias, the interviewees suggested to use duplicate data analysis—*i.e.,* asking two or more people to analyze the data independently. This approach was largely endorsed by the respondents. Also, according to the respondents, more researchers are usually involved when analyzing the data, so making duplicate data analysis a feasible solution. Duplicate data analysis can be easily extended to data extraction, and can be applied in alternative (or in conjunction) with blind data extraction and analysis. Our recommendation follows:

❧ *Researchers should consider simpler forms of blinding possibly together with duplicate data extraction and analysis if they cannot involve external experts or external researchers in the process of data extraction and analysis.*

The interviewees suggested other strategies to mitigate researcher bias. A large part of respondents considered open data policies to be effective in mitigating researcher bias. Publicly-available datasets and analysis scripts foster external replications, which can help us understand how large is the role that researchers play when attempting at replicating experimental results. In the SE research, there seems to be a shortage of replication studies. A 2005 literature survey of 103 controlled experiments published in leading SE journals [30] reported that only 18% were replications. Out of these, the experimental results tend to be confirmed when the same team of researchers attempts to replicate the results. For example, this was the case for six out of the seven experiments categorized as differentiated replications. The lack of result replicability is usually attributed to the variations in the contextual factors of the experiments (*e.g.,* programming language, participants' experience) [31]. However, to the best of our knowledge, only few studies directly attribute the different results to the fact that other researchers carried out the replication [16]. Two other recommended strategies to mitigate researcher bias, both largely supported by the respondents, are: *(i)* experiment protocol pre-registration and *(ii)* negative-results conference tracks and journal issues. We can thus delineate the following recommendation:

❧ *Editorial and program boards should explicitly promote and reward open data policies. When possible, they should establish pre-registration and negative-results tracks and special issues to limit publishing results hampered by researcher bias.*

According to the interviewees, researcher bias could be triggered by specific reviewers' behaviors. The respondents largely agreed that such behaviors are the reviewers tendency to reject negative-results papers and to focus too much on empirical assessment at the expenses of novel contributions to the body of knowledge. These behaviors, combined with the pressure to publish (perceived by the large majority of the respondents), lead researchers to bias their results to make them more publishable. We can thus delineate the following recommendation:

❧ *Editorial and program boards should instruct reviewers to not judge the quality of a submission based on its results, either positive or negative. For submissions reporting interesting findings but with weak empirical assessment, boards should consider ad-hoc shepherding initiatives.*

In several research fields, researcher bias seems to be the leading cause of a methodological crisis (*e.g.,* [32, 33]). The sample of the empirical SE community we surveyed largely considered it to be the case also in the SE research field. We are concerned that the practitioners and the general public will consider the SE research field less credible due to the impact of researcher bias on the validity of SE research inquiries. Therefore, our last recommendation is:

❧ *The SE research community needs to raise awareness on researcher bias, the problems it can cause, as well as initiatives for limiting it. This can be accomplished, for example, with special conference panels and town hall meetings.*

Some of our recommendations have been already applied in fields where experiments with different degrees of control are the predominant research approach (*e.g.,* medicine [26]). The forensic sciences employ a technique called *sequential unmasking* [34]. Similar to data blinding, the approach aims at minimizing the influence of information (such as a suspect profile) when analyzing DNA collected from evidence. The approach also proposes a separation of tasks between individuals familiar with case information and the analyst from whom domain-irrelevant information is masked.

Fields focusing on collecting and analyzing qualitative data have developed other ways to address researcher bias, such as "Interview the interviewers" [35]. This approach allows the interviewer to identify a priori assumptions about the participants by becoming one of them and being interviewed by a third-party who does not have any specific expectations on the answers (*e.g.,* a colleague not involved in the study). The interviewer records the interview and compares it with the script, self-reflecting on the parts that were included or left out. In the social sciences, there are two recommended approaches to do so, *journaling* [36] and *inter-personal recalling* [37]. Similar forms of self-reflection and peer-review are recommended as ways to reduce researcher bias in fields, such as anthropology, which make extensive use of ethnographies as research methods [38].

5.2. Limitations

The response rate (20%) of the survey might imply that only motivated researchers took part in the survey. This might have affected the results of the survey; however, motivated researchers are more likely to answer truthfully.

We left the online questionnaire open only for 20 days. This might have affected the response rate of the survey and thus the results. Despite we included in the questionnaire only the statements we deemed more relevant as suggested in the literature (*e.g.,* [23]), the number of statements in the questionnaire might have had an effect on the response rate. On the other hand, reducing further the number of statements included in the questionnaire would have affected our capability of triangulating the results from the two studies.

The sampling method used in the interview study, as well as the one used in the survey, might have affected the results. Both interviewees and respondents might not have answered truthfully because scarcely motivated or afraid of being judged. To mitigate this threat in the interview study, the participation in the study was voluntary—volunteers are generally more motivated [3]—and we informed the interviewees that the gathered data would be treated confidentially. As for the survey, the answers to the questionnaire were anonymous.

Respondents of questionnaires might have difficulty comprehending statements or questions (*e.g.,* because ambiguous, not clear, or not well formulated). To mitigate such a threat, we conducted a pilot study with two junior researchers. The use of unfamiliar terms in questionnaires might negatively influence questionnaire comprehensibility as well. We mitigated such a threat by including in the questionnaire explanations of terms that could be unfamiliar to the respondents.

Investigators might unconsciously influence the results based on their expectations. We mitigated such a threat by involving more than one author in the analyses of the data from the interview study and survey (*i.e.,* we applied *investigator triangulation* [22]).

Finally, since the recommendations delineated in Section 5.1 are based on evidence collected from interviewees and respondents within the SE community, we cannot claim they will apply to other research fields.

6. Conclusion

In this paper, we investigate researcher bias in SE experiments, including: *(i)* QRPs potentially leading to researcher bias; *(ii)* causes behind researcher bias; and *(iii)* possible actions to counteract researcher bias with a focus on, but not limited to, blind data extraction and analysis. To pursue such an objective, we adopted a two-step methodological approach comprising a qualitative interview study followed by a quantitative survey. The interview study is intended as an exploratory study. The findings from this survey represented the starting point to design the survey, which we conducted to support the findings from the interview study. The findings from the interview study are mostly confirmed by those from the survey—*e.g.,* the post-hoc outlier removal practice is not always questionable for both interviewees and respondents. In few cases, the findings from the interview study are not confirmed—*e.g.,* the interviewees did not find questionable the formulation of post-hoc hypotheses, while the respondents did. Both interviewees and respondents perceived the presence of researcher bias in se experiments. Therefore, researcher bias cannot be underestimated. To counteract it, we delineated a series of recommendations; some of them are intended for se researchers, while others are purposeful for the boards of SE research venues.

Acknowledgment

The authors would like to thank both interviewees and respondents for their participation in the studies presented in this paper.

Appendix A. Invitation letter to the survey

References

[1] J. Ioannidis, Why most published research findings are false, PLOS Med (2005).

[2] K. Dwan, D. G. Altman, J. A. Arnaiz, Systematic review of the empirical evidence of study publication bias and outcome reporting bias, PloS one 3 (2008) e3081.

[3] C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson, B. Regnell, A. Wessln, Experimentation in Software Engineering, Springer, 2012.

[4] D. Sjøberg, G. Bergersen, Construct Validity in Software Engineering (2021).

[5] S. Romano, D. Fucci, G. Scanniello, M. T. Baldassarre, B. Turhan, N. Juristo, Researcher bias in software engineering experiments: a qualitative investigation, in: Proceedings of EUROMICRO Conference on Software Engineering and Advanced Applications, IEEE, 2020, pp. 276–283.

Dear colleague,
you are receiving this email as you are an active researcher on topics related to empirical software engineering (ESE).

In our previous work (https://arxiv.org/abs/2008.12528), we conducted an explorative, qualitative study to investigate researchers' bias [*] in software engineering (se) experiments. We have now planned a survey as we want to validate the statements obtained in our previous work within the ESE community at large.

As so, we are reaching out to you as an expert in such community and ask if you could participate in our survey which is available at: https://ww2.unipark.de/uc/rbse/.

Please feel free to forward this survey to other researchers with experience on the topic. The link will be available until November 25th 2020.

If you have any questions, don't hesitate to contact us.

Thank you in advance for participating in this survey!

Sincerely yours,
Maria Teresa Baldassarre, Davide Fucci, Natalia Juristo, Simone Romano, Giuseppe Scanniello, Burak Turhan.

****

* Researcher bias occurs when researchers influence the results of an empirical study based on their expectations [1]. It might be due to the use of questionable research practices (e.g., the exclusion of data that are inconsistent with a theoretical hypothesis). In research fields like medicine, different techniques have been applied to counteract researchers' bias.
[1] Sackett, D.L. Bias in Analytic Research. Journal of Chronic Diseases, 1979; 32: 51-63.
****

Figure A.11: Invitation letter to the survey.

[6] R. MacCoun, S. Perlmutter, Blind analysis: hide results to seek the truth, Nature 526 (2015) 187–189.

[7] P. J. Karanicolas, F. Farrokhyar, M. Bhandari, Blinding: Who, what, when, why, how?, Canadian Journal of Surgery. 53 (2010) 345–348.

[8] D. Fucci, G. Scanniello, S. Romano, M. Shepperd, B. Sigweni, F. Uyaguari, B. Turhan, N. Juristo, M. Oivo, An external replication on the effects of test-driven development using a multi-site blind analysis approach, in: Proceedings of International Symposium on Empirical Software Engineering and Measurement, ACM, 2016, pp. 3:1–3:10.

[9] B. Sigweni, M. Shepperd, Using blind analysis for software engineering experiments, in: Proceedings of International Conference on Evaluation and Assessment in Software Engineering, ACM, 2015, pp. 32:1–32:6.

[10] R. Francese, C. Gravino, M. Risi, G. Scanniello, G. Tortora, Mobile app development and management: Results from a qualitative investigation, in: Proceedings of International Conference on Mobile Software Engineering and Systems, IEEE, 2017, p. 133–143.

[11] E. Murphy-Hill, T. Zimmermann, N. Nagappan, Cowboys, ankle sprains, and keepers of quality: How is video game development different from software development?, in: Proceedings of International Conference on Software Engineering, ACM, 2014, p. 1–11.

[12] D. Falessi, N. Juristo, C. Wohlin, B. Turhan, J. Münch, A. Jedlitschka, M. Oivo, Empirical software engineering experts on the use of students and professionals in experiments, Empirical Software Engineering 23 (2018) 452–489.

[13] D. Fanelli, R. Costas, J. P. A. Ioannidis, Meta-assessment of bias in science, Proceedings of the National Academy of Sciences 114 (2017) 3714–3719.

[14] L. John, G. Loewenstein, D. Prelec, Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling, Psychological Science (2012) 524–532.

[15] M. Jørgensen, T. Dybå, K. Liestøl, D. I. Sjøberg, Incorrect results in software engineering experiments: How to improve research practices, Journal of Systems and Software 116 (2016) 133–145.

[16] M. Shepperd, D. Bowes, T. Hall, Researcher bias: The use of machine learning in software defect prediction, IEEE Transactions on Software Engineering 40 (2014) 603–616.

[17] C. J. Pannucci, E. G. Wilkins, Identifying and avoiding bias in research, Plastic and reconstructive surgery 126 (2010) 619–625.

[18] R. Nuzzo, Fooling ourselves, Nature 526 (2015) 182–185.

[19] B. Nosek, D. Lakens, Registered reports: a method to increase the credibility of published results, Social Psychology 45 (2014) 137–141.

[20] L. Miller, M. Stewart, The blind leading the blind: Use and misuse of blinding in randomized controlled trials, Contemporary Clinical Trials 32 (2011) 240–243.

[21] S. J. Page, A. C. Persch, Recruitment, retention, and blinding in clinical trials, American Journal of Occupational Therapy 67 (2013) 154–161.

[22] V. Thurmond, The point of triangulation, Journal of Nursing Scholarship 33 (2001) 253–258.

[23] M. Kasunic, Designing an Effective Survey, Software Engineering Institute, 2005.

[24] N. King, Using templates in the thematic analysis of text, in: C. Cassell, G. Symon (Eds.), Essential Guide to Qualitative Methods in Organizational Research, Sage, 2004, pp. 256–270.

[25] A. Roodman, Blind analysis in particle physics, Statistical Problems in Particle Physics, Astrophysics, and Cosmology (2003) 166.

[26] R. Doll, Controlled trials: the 1948 watershed, Bmj 317 (1998) 1217–1220.

[27] M. Shepperd, How do i know whether to trust a research result?, IEEE Software 32 (2015) 106–109.

[28] H. Pashler, E. Wagenmakers, Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?, Perspectives on Psychological Science 7 (2012) 528–530.

[29] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, J. Rosenberg, Preliminary guidelines for empirical research in software engineering, IEEE Transactions on Software Engineering 28 (2002) 721–734.

[30] D. Sjoeberg, J. Hannay, O. Hansen, V. Kampenes, A. Karahasanovic, N.-K. Liborg, A. Rekdal, A survey of controlled experiments in software engineering, IEEE Transactions on Software Engineering 31 (2005) 733–753.

[31] N. Juristo, S. Vegas, The role of non-exact replications in software engineering experiments, Empirical Software Engineering 16 (2011) 295–324.

[32] J. P. A. Ioannidis, T. D. Stanley, H. Doucouliagos, The Power of Bias in Economics Research, The Economic Journal 127 (2017) F236–F265.

[33] M. F. Dacrema, P. Cremonesi, D. Jannach, Are we really making much progress? a worrying analysis of recent neural recommendation approaches, in: Proceedings of Conference on Recommender Systems, ACM, 2019, p. 101–109.

[34] D. E. Krane, S. Ford, J. R. Gilder, K. Inman, A. Jamieson, R. Koppl, I. L. Kornfield, D. M. Risinger, N. Rudin, M. S. Taylor, et al., Sequential unmasking: a means of minimizing observer effects in forensic dna interpretation., Journal of Forensic Sciences 53 (2008) 1006–1007.

[35] R. J. Chenail, Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research., Qualitative Report 16 (2011) 255–262.

[36] D. L. Miller, One strategy for assessing the trustworthiness of qualitative research: Operationalizing the external audit. (1997).

[37] N. Kagan, Influencing human interaction. (1972).

[38] M. D. LeCompte, Bias in the biography: Bias and subjectivity in ethnographic research, Anthropology & Education Quarterly 18 (1987) 43–52.