Final version available at Elsevier : http://dx.doi.org/10.1016/j.jss.2015.03.065

Incorrect Results in Software Engineering Experiments: How to Improve Research Practices

Magne Jørgensen[1,2], Tore Dybå[2,3], Knut Liestøl[2], Dag I. K. Sjøberg[2]
[1]Simula Research Laboratory, Norway
[2]Department of Informatics, University of Oslo, Norway
[3]SINTEF, Norway

Abstract

Context: The trustworthiness of research results is a growing concern in many empirical disciplines.
Aim: The goals of this paper are to assess how much the trustworthiness of results reported in software engineering experiments is affected by researcher and publication bias and to suggest improved research practices.
Method: First, we conducted a small-scale survey to document the extent of researcher and publication biases in software engineering experiments. Then, we built a model that estimates the proportion of correct results for different levels of researcher and publication bias. A review of 150 randomly selected software engineering experiments published in the period 2002–2012 was conducted to provide input to the model.
Results: The survey indicates that researcher and publication bias is quite common. This finding is supported by the observation that the actual proportion of statistically significant results reported in the reviewed papers was about twice as high as the one expected assuming no researcher and publication bias. Our models suggest a high proportion of incorrect results even with quite conservative assumptions.
Conclusion: Research practices must improve to increase the trustworthiness of software engineering experiments. A key to this improvement is to avoid conducting studies with unsatisfactory low statistical power.

Appearances to the mind are of four kinds. Things either are what they appear to be; or they neither are, nor appear to be; or they are, and do not appear to be; or they are not, and yet appear to be. Rightly to aim in all these cases is the wise man's task.
Epictetus (AD 55-135), Discourses, Book 1, Chapter 27

1. Introduction

The cover article, *How Science goes wrong*, of the October 19th 2013 issue of *The Economist* describes the growing concern that the proportion of incorrect research results in many research domains is much higher than we would normally suppose, or like to think. If the proportion of incorrect results is high, the usefulness and trustworthiness of the research within the whole domain may be at stake. The much debated and cited paper from 2005, by J. P. A. Ioannidis, with the telling title: "*Why most published research findings are false*" [1], is the origin of much of the recent discussions and concerns. There is, however, nothing new with concerns related to data fabrication [2], publication bias (not publishing statistically non-significant results) [3, 4], researcher bias (flexible analyses that lead initially statistically non-significant results to become significant) [2, 5] and low statistical power (low likelihood of rejecting the hypothesis of no difference, the null hypothesis, even when there actually is a difference) [6]. Already in 1830, Babbage wrote about the decline in science, including what he called the "fraud of the observers" [7]. Babbage's list of questionable practices (frauds) is similar to those discussed in this paper. Researchers may feel a strong pressure to publish results, which sometimes leads to questionable or even unethical researcher practices [8].

Although the use of questionable research practices is not a new phenomenon, an increasingly competitive research environment, a "publish or perish" culture, may have increased the amount of such practices over the years [9], i.e., increasingly competitive academic environments seem to increase not only the scientists' productivity, but also their biases [10]. The use of questionable practices is hardly just a result of lack of knowledge about proper research practices. The survey reported in [11], for example, finds the amount of questionable research practices to be similar or, for some aspects, even increasing for researchers in the later stages of their research career.

The goal of this paper is to examine to what extent the trustworthiness problems observed in a wide range of research domains [8, 12-16] are present in the context of software engineering experiments. If such problems are present, there may be a need for changes in the current research practices.

The trustworthiness of a particular result of a study depends on the quality of the research method of that study and to what degree the result has been replicated by other, preferably independent, studies. In this paper, we assess the trustworthiness of the results within a domain as a whole. The approach we apply is limited to research results from statistical hypothesis testing and is based on a model that estimates the expected proportions of statistically significant results [1, 17, 18]. Input to this model includes the level of publication and researcher bias, and the statistical power of studies conducted in a research domain. A high level of publication and researcher bias increases the proportion of incorrect research results and inflates the effect sizes [19, 20]. Similarly, low statistical power is also likely to increase the proportion of incorrect results [21].

An illustration of the unfortunate consequence of strong publication bias, strong researcher bias and low statistical power on result trustworthiness is provided in Box 1.

Box 1: Do authors with longer names write more complex papers?

We wanted to test the following hypothesis: *Researchers with longer names write more complex texts than researchers with shorter names.* To test the hypothesis, we randomly selected twenty research papers using Google Scholar. For each of the papers, we collected information about the first author's family name and the complexity of the text in the paper. We found a strong and significant ($p<0.01$) correlation between the length of the name and the complexity of the text, where the complexity of the text was measured either using the Flesch-Kincaid [22] reading level or the number of words per paragraph. The correlation with name length was 0.6 for both complexity measures.

While our study contains no fabricated data, we do not believe that authors with longer names actually write more complex papers. It is more likely that our result is a consequence of three questionable, but perhaps not uncommon, research practices. The first questionable practice, which is an example of publication bias, was that we did not publish all the (fourteen!) complexity measures we tested, only the two ones that gave significant results. The second questionable practice, which is an example of researcher bias, was that we removed two outliers because we were unable to calculate the Flesch-Kinkaid measure on the text. While in principle defendable, we made the decision after looking at the effect it had on the results. Without the removal of these outliers, our results would not have been statistically significant. The third questionable practice, also an example of researcher bias, was that we changed the definition of the length of the name from the sum of the length of the first name and the family name, to the length of the family name only. This was defended by the observation that the first name was not available for all authors. We knew, however, that this decision would strengthen our results.

All the questionable research practices we used to create statistically significant results in this study would, we think, easily go unnoticed or feel well motivated by the reviewers and readers. In this case, where collecting data is inexpensive, a reviewer may question why the sample is not larger or why no replications have been conducted. While this may be a valid comment for this study, the sample size used ($n=20$) or less is common in software engineering experiments where collecting data is typically more costly. (X % of the 150 experiments in the review reported later in this article had a sample size ≤ 20.)

A similar experience of how easy it is to generate statistically significant, but incorrect, results when willing to use questionable practices and studies with low statistical power is reported in [23].

The remaining part of the paper is organized as follows: Section 2 reports on a small-scale survey on questionable statistical practices of software engineering researchers. Section 3 introduces models of the expected proportion of statistically significant results and the expected proportion of incorrect results. Section 4 reports on a review of the results of hypothesis tests of a random set of 150 software engineering experiments. Section 5 uses the models described in Section 3 to argue that there is a substantial amount of researcher and publication bias, and consequently a high rate of incorrect results in software engineering experiments. Section 6 uses the results to suggest improved research practices. Section 7 concludes.

2. A small-scale survey of questionable research practices

A web-based survey was conducted with questions about statistical research practices likely to contribute to publication and researcher biases. We sent a questionnaire to the 80 participants and program committee members of the joint conference of the 23rd International Workshop on Software Measurement (IWMS) and the 8th International Conference on Software Process and Product Measurement (Mensura). In addition, we sent the questionnaire to a few members of the Dutch Software Measurement Association. We clarified that the respondents would be anonymous and that no one, not even the researchers analysing the responses, would be able to identify their names.

We received 36 complete responses. For the purpose of the analysis in this section, we removed two responses where the researchers stated that they never used statistical hypothesis testing in their own research, leaving 34 responses. The four first questions (P1–P4) of the questionnaire were related to publication bias and the last three questions (R1–R3) to researcher biases. The questions and the responses are displayed in Table 1.

Table 1: Results from a survey on statistical practices

Research Practice	**Have experienced/done this in my own research**				
	Never	**Seldom**	**Occas.**	**Often**	**Don't know**
P1: Paper rejected due to non-significance[1]	14	6	8	4	4
P2: Paper not submitted due to non-significance[2]	16	6	8	4	1
P3: Not reported non-significant results[3]	17	8	4	4	2
P4: Not reported undesired results[4]	18	8	0	4	4
R1: Post hoc hypotheses[5]	11	4	12	6	1
R2: Post hoc outlier criteria[6]	14	5	9	3	3
R3: Flexible reporting of measures and analyses[7]	10	10	5	7	2

1: Reviewers stated that a reason for rejecting your paper was that the results of one or more hypothesis tests gave statistically non-significant results.

2: You chose not to summarize and submit a paper, because the results of one or more of the hypothesis tests gave statistically non-significant results.

3: You chose not to report the outcome of one or more of the hypothesis tests (but submitted/published a paper with other statistically significant results from the same study or on the same topic), because the tests gave statistically non-significant results.

4: You chose not to report the outcome of one or more of the hypothesis tests (but submitted/published a paper with other statistically significant results from the same study or on the same topic), because the tests gave undesired results, e.g., results conflicting with the main message of the paper.

5: You reported the results of one or more hypothesis tests where at least one of the hypotheses was formulated after you had looked at the data.

6: You developed or changed the rules for whether to exclude data or not (e.g., outlier removal) after looking at the impact of doing so on the results.

7: You used several variants of a measure or several tests and reported only the measures and tests that gave the strongest results.

As can be seen in Table 1, practices likely to lead to publication bias were common among the respondents. A summary of the publication bias responses (excluding the category "Don't know") showed that 56% had experienced the rejection of a paper because it reported non-significant results, 53% had chosen not to submit a paper due to non-significant results, 48% had not reported non-significant results when reporting from a study and 40% had chosen not to report undesired results. Practices potentially leading to researcher bias were also common. We found that 67% had statistically tested and reported post hoc hypotheses, 55% had developed or modified outlier criteria after looking at the impact of doing so on the results, and 69% had only reported the best among several measures on the same test.

Self-report surveys on questionable research practices, even when reporting anonymously, are likely to underrepresent the true occurrences. Still, we found that between 40% and 69% of the respondents admitted to experiencing or using all these practices. The practices reported in our survey correspond well with those from a survey with similar questions in psychology [24]. In that survey, 63% of the respondents admitted that they had failed to report all of a study's dependent measures, 46% that they had selectively reported studies that "worked", and 38% that they had decided whether to exclude data after looking at the impact of doing so on the results[1]. The presence of researcher and publication biases also corresponds with responses from a health education research survey [25, 26], where 46% had witnessed first-hand that statistical techniques were selected "for [their] ability to provide [a] more favourable outcome" and 59% "reported only significant findings in published research".

The researchers in our study were allowed to comment on their answers. The comments included the following two very honest explanations of practices leading to publication and researcher bias:

[1] The results are those from the control group of the survey of the psychology researcher. The respondents of the "truth-telling incentive" group report slightly higher use of questionable practices.

- "*It's extremely hard to publish a journal paper without 'massaging' the data and the hypotheses first. If you do not do this, you will end up with no publications at all. I think journal editors and reviewers should do something, so that they encourage honest accounts of empirical work, and make researchers with non-significant results feel welcome.*"
- *"... unless authors do something really stupid, it's very easy to get away with post-hoc interventions. Sneaking up and making it to a journal publication is common and if many fellows practice it, why should we discriminate against ourselves by discarding the practice? The price appears to be too high for this."*

We should be careful about generalizing the results in Table 1 to most software engineering researchers. The respondents of our study do, however, resemble a typical population of empirically-oriented software engineering researchers, and we may at least use the results to argue that practices that lead to publication and researcher bias are present in the domain of software engineering. The results give us reason to suspect that the amount of questionable research practices can be a serious problem for the trustworthiness of the research results in software engineering experiments.

3. Modelling the impact of publication and researcher bias

We describe three models in this section. Section 3.1 describes a model of how publication and researcher bias affects the proportion of observed statistically significant findings. Section 3.2 describes a model of the proportion of correct findings. Section 3.3 describes a model of the strength of the evidence when reporting a statistically significant finding. The models are applied in Section 5, with input from the review in Section 4.

3.1 A model of the proportion of statistically significant tests

We consider situations where the correctness of simple hypotheses is evaluated, e.g., a hypothesis stating that there is no difference between to experimental procedures or no effect of a variable. If such a simple null hypothesis is false, we consider it a *true relationship*.

We use the concepts and variables described in Tables 2 and 3 to describe a model of the proportion of statistically significant hypothesis tests. The model is similar to the model in [1], with the exception that we add a variable for publication bias.

Table 2: Combinations of reported and actual relationships

	Actual relationship	
Reported relationship	**True relationship**	**False relationship**
Statistically significant (positive tests)	True positives (TP)	False positives (FP) (Type I error)
Statistically non-significant (negative tests)	False negatives (FN) (Type II error)	True negatives (TN)

A domain with reliable research results has few false negatives (Type II errors), i.e., true relationships not reported as statistically significant, and in particular few false positives (Type I error), i.e., non-existent relationships reported as statistically significant ("false alarms"). When the sample size of a study is given, there will be a trade-off between the elements in Table 2. Requiring a lower p-value of a study to claim statistical significance means that the proportion of false positives decreases (which is good) at the cost of an increase of false negatives (which is not so good) [27].

The observed proportion of significant tests (PST) equals the sum of true positive and false positive tests, divided by the total number of tests, i.e.,

(1) PST = (TP+FP)/(TP+FP+TN+FN)

Table 3: Variables used in the model

Variable	Description
ptr	Proportion of true relationships among those tested. If, for example, we have a domain where half of the hypotheses test true relationships, then *ptr* = 50%. The *ptr*-value is typically *not* known and will differ from domain to domain.
α	Probability of Type I error or significance level. α corresponds to the expected proportion of false positives among actually false relationships. Typically α is set to 5% in software engineering experiments, which means that we should expect, in the long run, that 5% of the tests give a statistically significant result when there is no true relationship.
β	Probability of Type II error. β corresponds to the expected proportion of false negatives. In software engineering experiments the median β has been estimated to be about 70% [28], which means that we should expect, in the long run, that as much as 70% of the true relationship gives a statistically non-significant test result. The statistical power equals $(1-\beta)$ [29].
rb	Researcher bias. *rb* is the proportion of statistically non-significant results that becomes significant through questionable research or analysis practices, such as those described as R1-R3 in Table 1. For a more complete set of research practices that lead to researcher bias, see [24].
pb	Publication bias. *pb* is the proportion of statistically non-significant results that are not reported. Not reporting non-significant results includes both the situation where papers are not accepted or submitted due to non-significant findings and the situation where one or more non-significant results within a study are not reported, i.e., all the situations described as P1-P4 in Table 1.

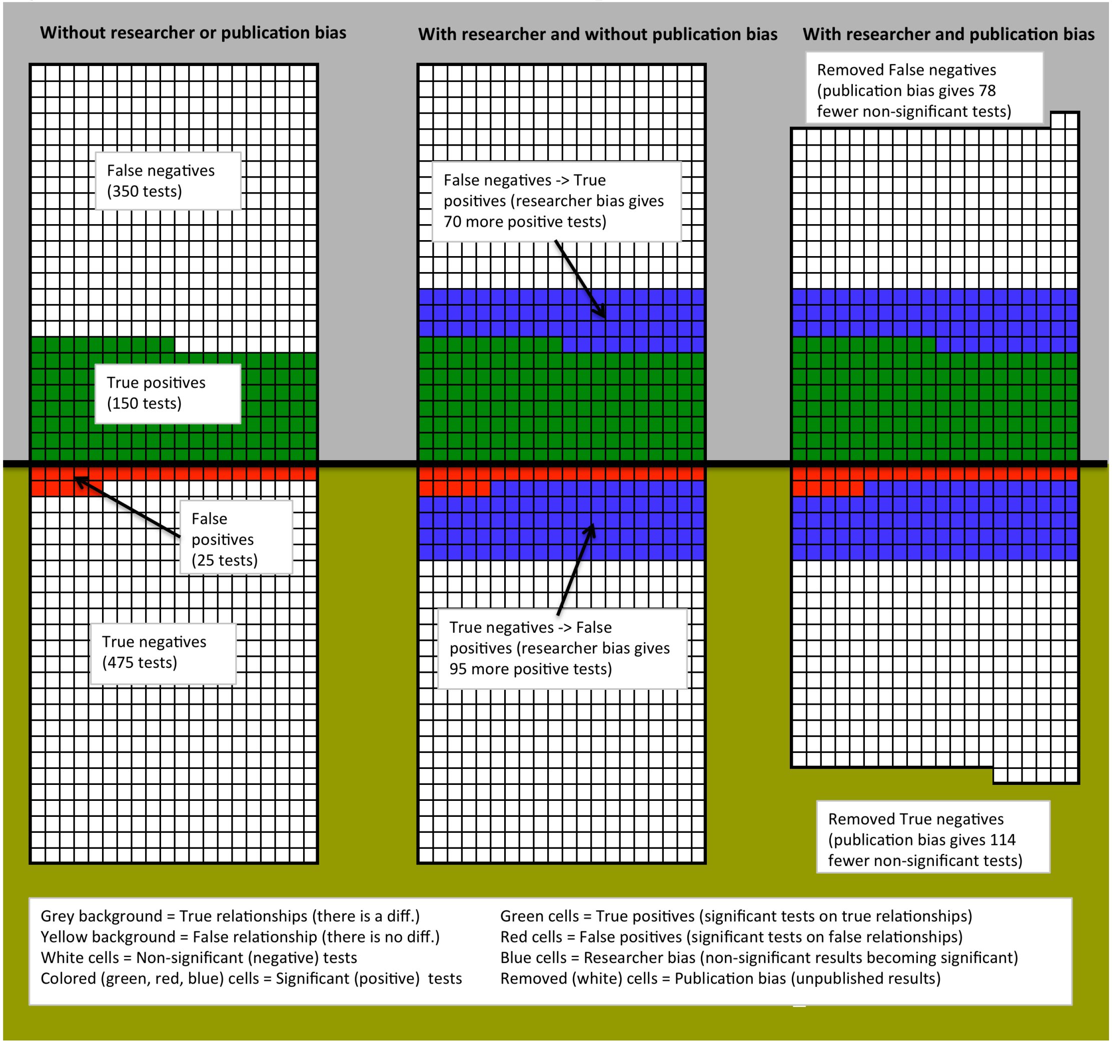
Figure 1 illustrates the relationships between the variables. Each of the 1000 cells represents one hypothesis test. As can be seen, we have assumed that 50% of the tested hypotheses are actually true (*ptr* = 0.5), a significance level (α) of 5% and a statistical power ($1-\beta$) of 30%. The left-hand part of Figure 1 illustrates the situation with no researcher and publication bias, the middle part a situation with 20% researcher bias and no publication bias, and the right-hand part a situation with 20% researcher and 30% publication bias. The cells above the thick horizontal black line (grey background) are the relationships that are actually true, while those below the line (yellow background) are those that are actually false. White cells denote tests that are non-significant, while the coloured cells (green = true positives, red = false positives, blue = initially non-significant test appearing significant due to researcher bias) denote statistically significant tests.

We assume that a statistical power of 30% means that 30% of the true relationships, i.e., 500 x 0.3 = 150 (15%) of the cells in Figure 1, are true positives (green cells) for a situation with no researcher or publication bias (left-hand situation). The remaining true relationships should correspond to negative tests; i.e., there are 500–150 = 350 false negatives (white cells, upper half). A significance level of 5% means that we will expect that 5% of the tests will (falsely) show a positive test when there is no relationship; i.e., 500 x 0.05 = 25 (2.5%) of the cells are false positives (red cells). The remaining false relationships correspond to negative tests; i.e., there are 500–25 = 475 true negatives (white cells, lower half). The expected number of statistically significant tests, assuming no researcher or publication bias, is consequently 150 (true positives) + 25 (false positives) = 175; i.e., 17.5% of the total tests are expected to be statistically significant in the situation with no researcher and publication bias.

More generally, the expected proportion of statistically significant relationships given no researcher and publication bias (PST_0) can be expressed as:

$$(2)\ PST_0 = (1-\beta)ptr + \alpha(1-ptr)$$

Figure 1: Illustration of the effect of researcher and publication bias


Without researcher or publication bias
With researcher and without publication bias
With researcher and publication bias
Removed False negatives (publication bias gives 78 fewer non-significant tests)
False negatives (350 tests)
False negatives -> True positives (researcher bias gives 70 more positive tests)
True positives (150 tests)
False positives (25 tests)
True negatives (475 tests)
True negatives -> False positives (researcher bias gives 95 more positive tests)
Removed True negatives (publication bias gives 114 fewer non-significant tests)
Grey background = True relationships (there is a diff.)
Yellow background = False relationship (there is no diff.)
White cells = Non-significant (negative) tests
Colored (green, red, blue) cells = Significant (positive) tests
Green cells = True positives (significant tests on true relationships)
Red cells = False positives (significant tests on false relationships)
Blue cells = Researcher bias (non-significant results becoming significant)
Removed (white) cells = Publication bias (unpublished results)

Adding 20% researcher bias, as we do in the middle part of Figure 1, means that 20% of the initially statistically non-significant relationships become statistically significant. We assume in our calculations that the researcher bias affects the true negatives and the false negatives equally.[2] The researcher bias implies that 20% (=70) of the 350 false negatives and 20% (=95) of the 475 true negatives become statistically significant. There will now *not* be 175 (as in the left-hand part of Figure 1) statistically significant results, but instead 175 + 70 + 95 = 340; i.e., 34% of the total tests will be expected to be statistically significant. The expected observed proportion of statistically significant findings (PST_1) when adding researcher bias can be expressed as:

$$(3)\ PST_1 = PST_0 + \beta \cdot rb \cdot ptr + (1-\alpha) \cdot rb \cdot (1-ptr)$$

Adding 30% publication bias, as we do in the right-hand part of Figure 1, means that 30% of the statistically non-significant relationships are removed from the total set of hypothesis testing. As for the researcher bias, we assume that this will affect the true negatives and the false negatives equally. Adding publication bias implies that 30% of the remaining 660 (=1000–340) statistically non-significant tests, i.e., 30% x 660 = 198 non-significant tests, are removed. There are now just 1000–198 = 802 tests left. There are still 340 statistically significant tests, but they now constitute *not* 34% of the total number of reported tests, but instead 340/802 = 42%. The expected observed proportion of

[2] This assumption may not be valid if the studies with initially false negatives tend to have much lower p-values than the studies with initially true negatives. In that case, it may be easier to achieve statistically significant tests for false negatives (tests on true relationships) than for true negatives (tests on false relationships). An improvement of the model would, consequently, be to allocate a higher proportion of researcher bias to the false than to the true negatives. For a situation with studies with low statistical power, as in our context, we expect that the difference in the ease of achieving statistically significant results through researcher bias is approximately the same for these two situations, which defends our assumptions.

statistically significant findings (PST_2) when adding researcher and publication bias can be expressed as:

$$(4)\ PST_2 = PST_1/(PST_1 + (1 - PST_1) \cdot (1 - pb))$$

We will use the expression in (4) to examine the levels of researcher and publication bias needed to produce the observed proportion of statistically significant results.

3.2 A model of the proportion of correct tests

In Figure 1 we have that the situation without researcher and publication bias (left-hand part), gives 150 true positives and 475 true negatives. This means that 150 + 475 = 625 of the 1000 tests (62.5%) are correct. Perhaps even more important, as many as 150 out of the 175 (=86%) statistically significant tests are correct.[3] When adding 20% researcher bias and 30% publication bias (right-hand part), the expected number of true positives (blue and green cells in the upper half) becomes 220, and the number of true negatives (white lower half) becomes 266. The number of tests is now reduced to 802, which means that the tests give the correct results (266 + 220)/802 = 61% of the times. This is about the same proportion as in the situation without researcher and publication bias, and is mainly a result of the low statistical power assumed in the illustration. The inclusion of researcher and publication bias gives, on the other hand, that only 220 (65%) of the 340 statistically significant tests now give the correct result; i.e., 35% of the statistically significant results are incorrect.

We can express the proportion of observed correct results of a set of tests, i.e., the Proportion of Correct Results (PCR)[4] and Proportion of Correct Results among those statistically Significant (PCRS)[5], as:

$$(5)\ PCR = \frac{PTP+PFN}{PTP+PFN+PFP+PTN} \text{ and}$$

$$(6)\ PCRS = \frac{PTP}{PTP+PFP},$$

where PTP is the proportion of true positives, PFN the proportion of false negatives, PFP the proportion of false positives and PTN the proportion of true negatives. These proportions can be expressed as:

$$(7)\ PTP = ptr \cdot (1 - \beta) + \ ptr \cdot \beta \cdot rb$$

$$(8)\ PFN = ptr \cdot \beta \cdot (1 - rb) \cdot (1 - pb)$$

$$(9)\ PFP = (1 - ptr) \cdot \alpha + (1 - ptr) \cdot (1 - \alpha) \cdot rb$$

$$(10)\ PTN = (1 - ptr) \cdot (1 - \alpha) \cdot (1 - rb) \cdot (1 - pb)$$

The above expressions are based on the same ideas as those reported in [30], but add the effect of publication and researcher bias. Note that, with publication bias, the sum of the proportions will not be 100%; i.e., we calculate the proportion of correct results among the reported tests, not among all tests actually conducted.

3.3 A model of the strength of the evidence from a statistically significant test

Expression (6), which estimates the proportion of correct results among the statistically significant tests, includes no reference to the level of publication bias. This is as expected, since publication bias affects only the non-significant results, but should not be taken to mean that publication bias is harmless for the reliability of reported statistically significant results.

A potential side-effect of publication bias is, for example, that it makes it easier to publish studies that test many hypotheses on topics where the proportion of true relationships (*ptr*) is low. As can be derived from (6), a decrease of *ptr* will decrease the proportion of correct, statistically significant results.

[3] This illustrates how misleading it is to think that the significance level of a hypothesis test tells us how probable it is that a null hypothesis is true. In this scenario, the proportion of true null hypotheses when observing $p<0.05$ is 14%. Adding researcher and publication bias, as we do in Figure 1, increases this proportion to 35%.

[4] Frequently denoted Accuracy (ACC)

[5] Frequently denoted Positive Predictive Value (PPV)

We will now demonstrate the effect of publication bias on result correctness through the use of a Bayes Factor [31] (see expression (11)). The considerations leading to Expression (11) are described in Appendix 1.

$$(11)\ \mathrm{BF} = \frac{(1-\beta)+\beta\cdot rb}{(1-\beta\cdot pb)+\ \beta\cdot rb\cdot pb} \Big/ \frac{(\alpha+(1-\alpha)\cdot rb)}{(1-pb)+\alpha\cdot pb+(1-\alpha)\cdot rb\cdot pb}$$

The Bayes Factor tells us how much the odds[6] of the alternative hypotheses (true relationship) increase after observing that a test gave statistically significant results. While classical hypothesis testing only considers the evidence against the null hypothesis, i.e., how unlikely it is to observe the data actually observed or more extreme data given that there is no difference, the Bayes Factor compares the strength of the evidence in favor of the null hypothesis (false relationship) with the strength of the evidence of the alternative hypothesis. The Bayes Factor may consequently be interpreted as a measure of how much we should update our belief, or how much our a priori odds should change, based on the collected evidence. A Bayes Factor of 1 means that the evidence equally favours the null and the alternative hypothesis. Values between 1 and 3 are typically interpreted as “not worth more than a bare mention”, between 3 and 20 as “positive”, between 20 and 150 as “strong”, and higher than 150 as “very strong” [31]. As pointed out in [32], while the Bayes Factor and traditional hypothesis testing almost always agree on which hypothesis is better supported by the data, they may disagree about the strength of this support. In [32], examining 855 t-tests from experiments in psychology, p-values between 0.01 and 0.05 corresponded with Bayes Factors of less than 3, i.e., “not worth more than a bare mention”, in as much as 70% of the tests.

4. The proportion of statistically significant results in software engineering experiments

To find the proportion of reported statistically significant tests among all reported tests in software engineering experiments, we conducted a systematic literature review. The design of the review process is displayed in Table 4.

[6] Odds = Probability / (1-Probability)

Table 4: The review process

Characteristic	Description
Population	All reported software engineering experiments applying statistical hypothesis testing, including quasi-experiments (experiments without random allocation of treatment) in the period 2002-2013.
Sample	25 randomly sampled papers from each of the periods: 2002–2003, 2004–2005, 2006–2007, 2008–2009, 2010–2011, and 2012–2013. In total, 150 papers.
Search process	Full text search with *Google Scholar* using the term: "software engineering" AND "experiment" AND "hypothesis" for each of the sample periods.
Inclusion criteria	At least one statistical hypothesis test.
Exclusion criteria	Studies with all hypotheses stated "post hoc" (derived from analyses of the collected data) and studies with "unfocused" hypotheses, i.e., where a large number of hypotheses derived from a very general hypothesis were tested.
Data collected	For each experiment, we collected the following information: 1) Paper reference (year, authors, title and source) 2) Study unit (students, professionals, projects, etc.) 3) Sample size (total number of subjects in study) 4) Number of treatments, including control group 5) Significance level chosen for study (1%, 5%, etc.) 6) Number of hypotheses tested 7) Number of non-significant tests 8) Number of tests with p-value less than 0.01 9) Number of tests with p-value between 0.01 and 0.05 10) Number of tests reported as significant, but without exact p-value A paper could include more than one experiment.
Review process	Three of the authors participated in the review process. The first author reviewed all the studies, while the second and fourth authors reviewed 50% of the studies each; i.e., all studies were reviewed by two of the authors. When there were disagreements on data collection or interpretations of study results, the paper was re-reviewed and discussed until agreement on the data collection was reached.
Synthesis	The data was summarized and the number of experiments, the median sample size, the number of hypothesis tests and the proportions of test results with p<0.01 and p<0.05 were calculated.

The results from the review are summarized in Table 5.

Table 5: Results from the review

	Total	**2002–2003**	**2004–2005**	**2006–2007**	**2008–2009**	**2010–2011**	**2012–2013**
No. papers	150	25	25	25	25	25	25
No. experiments	196	30	31	32	37	35	31
Median sample size	29	47	33	32	23	26	27
No. hypothesis tests	1279	212	210	251	220	215	171
p<0.05[1]	52%	53%	59%	52%	46%	52%	54%
p<0.01[2]	29%	27%	32%	31%	25%	33%	26%

1: Proportion of statistical hypotheses tests with reported p-value lower than 0.05

2: Proportion of statistical hypotheses tests with reported p-value lower than 0.01. A few tests only reported that p<0.05 without reporting the exact p-value. We assumed that half of these tests had p-values less than 0.01 and half had p-values between 0.01 and 0.05.

Table 5 shows that 52% of the hypothesis tests in the reviewed software engineering experiments resulted in statistically significant results when assuming a significance level of $\alpha = 0.05$. Furthermore, 29% of the hypothesis tests had p-values less than 0.01. The median sample size is 29 subjects only.[7]

[7] The low sample size may reflect the typical number of students in a software engineering class. Availability of subjects may consequently be the practical reason for the very low statistical power of software engineering experiments. This is further supported by the observation that around 90% of the experiments in our review used students as subjects; that is, the proportion of professionals as subjects in software engineering experiments has not increased since the previous decade (1993–2002) [Sjøberg et al. 2005].

Among the 150 studies, 140 (93%) reported at least one statistically significant hypothesis test. This is consistent with a situation where software engineering experiments are more likely to be published when they produce a statistically significant result. The alternative explanation is that nearly all experiments include the test of at least one true relationship and have sufficient statistical power to produce a positive test. The low sample size, and consequently low statistical power, of many of the experiments suggest that the first explanation is more likely than the second one.

The study reported in [33] found a proportion of statistically significant results in the field of computer science of about 80%, i.e., a much higher proportion than we found. This difference in results may be caused by a difference between the broader field of computer science and its subset of experimental software engineering. More likely, it is caused by a difference in review methods. It seems as if the review in [33] only examines whether a paper's "main" hypothesis is fully or partly supported, i.e., one hypothesis per paper. We, on the other hand, study all reported hypothesis tests of an included experiment. Only examining the hypothesis reported as the "main" one may easily lead to a new type of "publication bias", because an author may tend to emphasize the result that is statistically significant as the main result.

5. Researcher and publication bias and result correctness

To estimate the presence of researcher and publication bias in software engineering experiments and its effect on result correctness, we apply the models introduced in Sections 3.1–3.3. For this purpose, we need to make assumptions about the significance level (α), the statistical power (1-β) and the proportion of true relationships test (*ptr*). Table 6 displays and motivates the assumptions.

Table 6: Variable value assumptions

Variable	Value	Motivation
α	0.05	In our review of software engineering experiments, we found that almost 90% of the tests used 0.05 as the threshold for statistical significance.
(1-β)	0.3	A median statistical power of about 0.3, for medium–large effect sizes, was reported from a review of software engineering experiments in [28]. A medium–large effect size for software engineering experiments was documented in an analysis of the same experiments in [34]. The similarity between our set of studies and those in [28] with respect to sample size and the proportion of statistically significant results motivates the use of 0.3 as the median statistical power; i.e., there has not been much change since the previous review was conducted. (The median sample size was 34 in the previous review and 29 in our review, and the proportion of statistically significant results was 49% in the previous review [34] and 52% in our review.)
ptr	0.7	The proportion of true relationships tested by statistical hypothesis testing in software engineering experiments is unknown. To be on the safe side, i.e., to avoid the critique that we make assumptions that make the reliability of results from software engineering experiments look worse than it is, we assume a *ptr*-value that corresponds to a situation where as much as 70% of the tested hypotheses are true.[8] A *ptr*-value of about 0.7 is what is assumed to be the proportion of true relationships in a confirmatory meta-analysis situation in [1], i.e., in situations where one has good reasons to believe that there is a true relationship.

The expected proportions of statistically significant findings for different levels of researcher and publication bias by using the expression in (4) and the values in Table 6, are displayed in Table 7. The values in bold type are examples of combinations of researcher and publication bias with values close to the observed proportion of statistically significant results (52%).

[8] It may be argued that there will always be a true relationship, i.e., that it is unlikely that something is exactly the same, and that it is only a matter of sampling size whether we find a statistically significant difference or not. While this points out the problem of using statistical significance as a measure of practical significance, the validity of our model estimates is not affected. In our context, we may assume that a true relationship means a relationship that will be found to be significant for reasonable sample sizes.

Table 7: Expected median proportions of significant findings

		Researcher bias					
		0	0.1	0.2	0.3	0.4	0.5
Publication bias	0	23%	30%	38%	46%	**54%**	61%
	0.1	24%	33%	41%	48%	56%	63%
	0.2	27%	35%	43%	**51%**	59%	66%
	0.3	29%	38%	47%	55%	62%	69%
	0.4	33%	42%	**50%**	58%	66%	72%
	0.5	37%	46%	55%	63%	70%	76%
	0.6	42%	**52%**	60%	68%	74%	80%
	0.7	49%	59%	67%	74%	79%	84%
	0.8	59%	68%	75%	81%	85%	89%

Table 7 shows that 52% of the statistically significant tests that we observed in our review do not match a situation with no or low researcher and publication bias. In the case of no researcher and publication bias (*rb*=*pb*=0), we should observe only 23% of statistically significant findings. Even if all the hypotheses in software engineering experiments test true relationships (*ptr* = 1.0), we should not observe more statistically significant tests than are predicted by the typical statistical power, i.e., around 30%.

To estimate the reliability of reported findings in software engineering experiments, we use the four scenarios (combinations) of researcher and publication bias with the best match between the observed and the expected proportion of statistically significant findings, i.e., those corresponding to the values in bold letters in Table 7. These are the scenarios with researcher bias 0.4 and publication bias 0.0, researcher bias 0.3 and publication bias 0.2, researcher bias 0.2 and publication bias 0.4, and researcher bias 0.1 and publication bias 0.6. The scenario with no publication bias is unlikely. We therefore include only the other three scenarios in the following discussion of the result correctness.

Table 8 displays the values for the expressions in (5)–(10), i.e., proportion of true positives (PTP), proportion of false negatives (PFN), proportion of false positives (PFP), proportion of true negatives (PTN), proportion of correct results (PCR) and proportion of correct results among those that are statistically significant (PCRS) for the three chosen scenarios. We denote the proportion of not reported test results "PNP".

Table 8: Result reliability for selected scenarios

Scenario	**rb**	**pb**	**PTP**	**PFN**	**PFP**	**PTN**	**PNP**	**PCR**	**PCRS**
1	0.3	0.2	36%	27%	10%	16%	11%	58%	78%
2	0.2	0.4	31%	24%	7%	14%	25%	59%	81%
3	0.1	0.6	26%	18%	4%	10%	42%	62%	86%

In [1] it is claimed that "most published research findings are false". While this is not the case for software engineering experiments, the situation is, we believe, unsatisfactory. As can be seen in Table 8, the proportion of correct outcomes (PCR) is only 58–62%, i.e., slightly more than 50-50 of correct and incorrect results. The low proportion is, as indicated by the high proportion of false negatives (PFN), to a large degree caused by the low statistical power of most software engineering experiments. The correctness of findings reported to be statistically significant (PCRS) is 76–86%; i.e., 14–24% of the statistically significant results will be incorrect. As pointed out earlier, the PCRS measure fails to take into account the decrease in evidence strength in situations with high publication bias, such as in Scenario 1. We discuss this effect later in this section.

Our assumption on 70% true relationships (*ptr* = 0.7) among those tested is very conservative. If we, perhaps more realistically, assume that there are 50% true relationships among those examined (*ptr* = 0.5), the PCR value does not change much (the new interval is 59–64%), but the PCRS value does. It is now 60–72%; i.e., as much as 28–40% of the claimed statistically significant results will in this case be incorrect! We will hardly ever know the actual *ptr* value, but there seem to be a few research domains where they think it is reasonable to assume a *ptr* value higher than 0.5. If there are topics or sub-domains where the proportion of true relationships is as low as 30% (*ptr* = 0.3), then the claim "most published research findings are false" may be the case for software engineering experiments as well (PCR between 34 and 40%, and PCRS between 48 and 61%).

To test the robustness of the results in Tables 7 and 8, we re-calculated PCR and PCRS with the following two changes in assumption:

- It may be argued that the chosen level of significance places too little emphasis on tests that are highly significant; i.e., the calculations do not sufficiently acknowledge that many tests will be significant at, for example, the 1% level. The consequence is that our model may give a too negative estimate of the research correctness (see [35] for a discussion on this critique). Our review of the software engineering experiments gave that about 50% of the statistically significant ($p < \alpha = 0.05$) results had p-values less than 0.01. We therefore modelled a situation where 50% of the studies had an α of 0.01 and 50% of the studies an α of 0.05, applying the same assumptions as before. This gave results very similar to those reported in Tables 7 and 8.
- The statistical power of software engineering experiments varies and it may give a biased outcome to use the median statistical power to represent the set of all experiments, especially if the distribution of statistical power values is strongly skewed. We therefore re-calculated the proportion of correct results using the empirical distribution of statistical power of the software engineering experiments reported in [28], i.e., for a similar set of software engineering experiments. This simulation, where we randomly selected 1000 statistical power values from the empirical distribution, again gave results that were very similar to those reported in Tables 7 and 8.

Applying the Bayesian Factor expression in (11) for selected levels of researcher and publication bias, and assuming as before a statistical power of 0.3 and a significance level of 0.05, further supports the unsatisfactory situation resulting from researcher and publication bias (see Figure 2).

Figure 2: Decrease in strength of evidence with publication and researcher bias

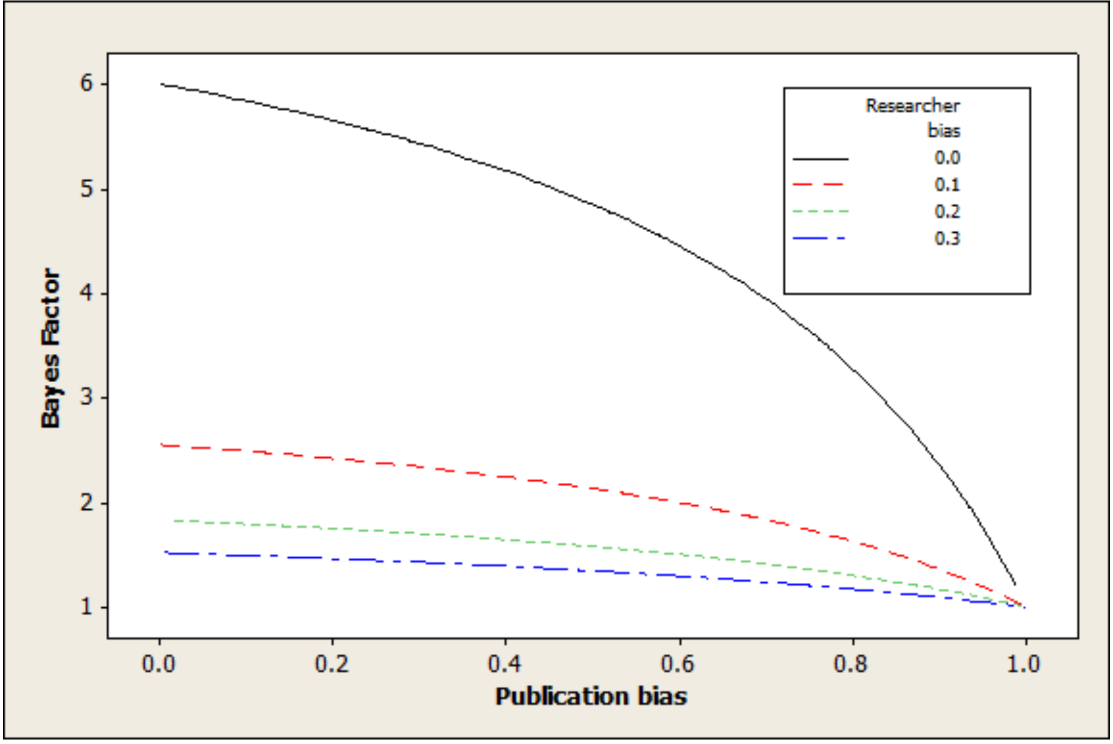


Figure 2 indicates that particularly the researcher bias affects the strength of the evidence from statistically significant findings. The three scenarios in Table 8, for example, result in Bayes Factors between 1.5 and 2.0, i.e., results categorized as “not worth more than a bare mention”. A Bayes Factor of 1.0 means that the evidence supports the alternative hypothesis (true relationship) and the null hypothesis (false relationship) equally, while a Bayes Factor of 2.0 means that the data is just twice as likely to observe when the relationship is true compared with when the relationship is false. Clearly, Bayes Factors in the interval 1.5 to 2.0 cannot be considered to provide strong evidence in support of claiming a relationship based on a statistically significant finding.

Figure 2 also shows that the maximum strength of evidence is a Bayes Factor of 6, which is categorized as “positive” but not as “strong evidence”. The low strength of evidence, even in an unbiased situation, is due to the very low average statistical power of software engineering experiments. As a comparison, a situation where studies have a statistical power of 0.8, where we observe statistical significance at $p < \alpha = 0.01$, and where there is no researcher bias and publication bias, would give a Bayes Factor of 80, i.e., “strong evidence”.

6. What should be done to improve research practices?

A high level of reliability of research results is a prerequisite for the use of scientific studies as input to evidence-based practices in software engineering [36]. Undoubtedly, there are many software engineering experiments of high quality. The results from the analyses of this paper suggest, nevertheless, that there is a need to improve research practices. Meta-studies on a particular research question, replications and/or careful reviews of the research quality of individual studies may improve the reliability of applied research results, but cannot fully remove the unfortunate effects of low statistical power and strong publication and researcher bias within a research domain.

To address the identified challenges related to publication and researcher bias, we recommend the following improvement in researcher and reviewer practices (described in more detail in Table 9):

- Avoid studies with low statistical power
- Avoid studies with many statistical tests
- Emphasize effect sizes and their confidence intervals
- Improve the reporting of the study design, analysis and results
- Increase the acceptance of non-significant research results
- Increase the number of replications and meta-analyses of studies
- Make the raw data and details about the research process available

There is no shortage of recommendations on how to conduct empirical software engineering studies (see, for example, [37-40]), and all our advice is included in previously reported, more general and more comprehensive guidelines and textbooks. What we add is a focus on the need to reduce researcher and publication bias by quantifying its negative effects. We also suggest changes in reviewer and editorial practices to help overcome the problem. We are aware that our suggestions, to some extent, are in conflict with the need for more empirical studies to answer the set of questions of high industrial importance [41]. Higher statistical power of studies, for example, means that a study may be more costly and require greater research competence, and, therefore, fewer empirical studies may be conducted. In spite of this consequence, we believe that it is not acceptable to ignore the identified challenges of research reliability.

Previously published advice has typically not had much impact on research practice. The similarity, or even decrease, in sample size from the survey on software engineering experiments from 1993–2002 [28] and 2002–2013 (this paper), for example, suggests that the clear advice about the need to increase statistical power has had little impact on actual statistical power. Reasons for the deviations between best and actual practice may include:

- Lack of capability, e.g., due to a lack of financial resources or access to a large number of developers, to conduct large-scale experiments with sufficient statistical power. Classes of student programmers are consequently the easiest, and perhaps the only feasible, option.
- Publication mechanisms implicitly reward the use of questionable practices [8]. As an illustration, there is an increased probability of finding at least one statistically significant (publishable) finding when conducting many small studies with many hypotheses instead of one larger study with a fewer, well-defined hypotheses.
- Questionable practices, including very low statistical power, are common even among senior researchers. If they can, why shouldn't I, i.e., "do as the others".
- It is harder to publish a replication of the findings of other researchers than to be the first one with an interesting, statistically significant, finding.
- There is little to gain and much to lose from making data available. Hiding their own data, e.g., claiming that they are confidential, makes it much harder for other researchers to find embarrassing errors or weaknesses in the analyses.

Changing software engineering researchers' behaviour from questionable to more proper research practices is, we believe, not so much about better training in empirical studies, although that is important as well. Even more important is the creation of mechanisms that reward good practices. Such mechanisms must include changes in the reviewing process and paper acceptance criteria. For each recommended practice in Table 9, we therefore include what we believe are useful changes for reviewing policies.

Table 9: Improved research practices, and suggested consequences, for software engineering experiments with hypothesis testing

Advice	Practical consequences

Avoid low statistical power [21, 40, 42]	Researchers: • Carry out analyses of statistical power before running a study. • Cancel or re-design studies with unsatisfactory low statistical power. • Do not use the observed (post hoc) statistical power as an indicator for the statistical power of the study. Reviewers: • Require that papers include a discussion on the desired level of statistical power and the implied sample size as part of the design of the study. • Reject studies with unsatisfactory low statistical power for reasonable effect sizes, e.g., studies with statistical power of less than 0.6 for meaningful effect sizes, regardless of statistical significance.
Avoid complex studies with many statistical tests [39, 43]	Researchers: • Keep the design of the experiment simple and transparent. • Include few hypotheses and variables in your study. Know that an increase in the number of hypothesis tests reduces the strength of evidence from each test. Reviewers: • Reward studies with a simple design and few hypotheses and variables. • Reject studies with a high number of statistical hypothesis tests, especially when it is likely that a substantial proportion of the statistical significant tests are due to chance or formulated post hoc, or when there are indications of publication bias.
Improve the reporting of study design, analyses and results [39, 44, 45]	Researchers: • 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). Avoid statistical tests on exploratory hypotheses. • Report on all evaluated tests and measures, especially when measures are on variants of the same construct. • Decide on inclusion/exclusion (outlier) criteria and statistical instruments in advance. • Avoid confusing statistical and practical significance. Reviewers: • Request that the authors make a statement where they declare that, for all experiments, they have accurately reported all hypothesis tests, measures, conditions, data exclusions, and how they determined their sample sizes. Making such a statement may be a mandatory step in the submission process for papers that report experimental studies, similarly to the declaration of vested interests in many research domains.
Emphasize effect sizes and their confidence intervals [46]	Researchers: • Many, probably most, research questions are better formulated as "How large is the effect?" rather than "Is there an effect?" Such questions should be answered by reporting confidence intervals of effect sizes, rather than by the use of p-values. • Use the confidence intervals of the effect sizes as the main means to interpret the importance and precision of the results, not the p-values. Reviewers: • Request that confidence intervals and their effect sizes are reported. Acknowledge that a confidence interval of effect sizes that includes "no effect", e.g., zero difference in mean values, can be very informative. This is especially the case when the confidence interval is narrow.
Accept non-significant results [47, 48]	Researchers: • Report well-powered studies and tests with statistically non-significant results the same way you would do when finding statistically significant results. Reviewers: • Accept non-significant results from studies with good research quality

	and reasonable statistical power.
More replications and meta-studies [39, 49]	Researchers: • Conduct replications of your own and others' studies to increase result robustness. Replications do not have to be, and frequently should not be, *identical* replications. Publish results even if you do not manage to find the same result as the original study. Emphasize the replication of effect size, less than the statistical significance.[9] • Conduct meta-studies based on the original study and its replications. Adjust for strength of study, publication bias and look for indicators of researcher bias in individual studies. Reviewers: • Replications, especially of other researchers' studies, and meta-studies should be welcomed, even when presenting non-significant results. As with other experiments, the statistical power of replications should be explicitly considered as part of the design of the replication.
Make data available [39]	Researchers: • Unless there are very good reasons for not disclosing the raw data, make them openly available, at least on request to the author. Include information about the data collection and analysis that might be needed to properly use the results of your study. Reviewers: • To publish a paper, it should be required that the data is made openly available unless strong confidentiality reasons prohibit it.

5. Conclusion

Experiments in software engineering are subject to publication and researcher biases. We document these biases found in our small-scale survey in a follow-up literature review that demonstrates that an unbiased situation does not match the observed proportion of statistically significant tests. The biases are also found in meta-studies of software engineering topics [50, 51]. As a consequence, the reliability of results reported in software engineering experiments, even when assuming a best-case (conservative) scenario, is unsatisfactory. The unsatisfactory reliability of the research results implies that there is a need for improvement in research and review practices.

If followed, our advice, we believe, has the potential to lead to a substantial improvement in the reliability of research results. Unfortunately, the current publishing mechanisms do not always reward good research practices. Consequently, we urge the research community to change the review practices, i.e., reconsider what is accepted and not accepted by software engineering journals and conferences. In particular, researchers in software engineering need to strengthen the statistical power of their studies, and reviewers of software engineering experiments should require that the researchers derive the population size of their studies based on consideration of reasonable levels of statistical power.

References

[1] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS medicine,* vol. 2, p. e124, 2005.

[2] J. D. Dingell, "Misconduct in medical research," *New England Journal of Medicine,* vol. 328, pp. 1610-1615, 1993.

[3] D. M. Lane and W. P. Dunlap, "Estimating effect size: Bias resulting from the significance criterion in editorial decisions," *British Journal of Mathematical and Statistical Psychology,* vol. 31, pp. 107-112, 1978.

[4] I. F. Tannock, "False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons," *Journal of the National Cancer Institute,* vol. 88, pp. 206-207, 1996.

[5] E. Masicampo and D. R. Lalande, "A peculiar prevalence of p values just below. 05," *The Quarterly Journal of Experimental Psychology,* vol. 65, pp. 2271-2279, 2012.

[9] It may feel natural that a replication of a study that finds $p<0.05$ should have a good chance of replicating the significance. This is frequently *not* the case and part of the fallacy "belief in the law of small numbers" [6]: While one should expect to find about the same effect size in a replication, one would need to strongly *increase* the number of subjects to have, for example, an 80% chance of replicating a slightly significant finding.

[6] A. Tversky and D. Kahneman, "Belief in the law of small numbers," *Psychological Bulletin,* vol. 76, p. 105, 1971.

[7] C. Babbage, *Reflections on the Decline of Science in England, and on Some of Its Causes*: B. Fellowes, 1830.

[8] M. Bakker, A. van Dijk, and J. M. Wicherts, "The rules of the game called psychological science," *Perspectives on Psychological Science,* vol. 7, pp. 543-554, 2012.

[9] D. Fanelli, "Negative results are disappearing from most disciplines and countries," *Scientometrics,* vol. 90, pp. 891-904, 2012.

[10] D. Fanelli, "Do pressures to publish increase scientists' bias? An empirical support from US States Data," *PloS one,* vol. 5, p. e10271, 2010.

[11] B. C. Martinson, M. S. Anderson, and R. De Vries, "Scientists behaving badly," *Nature,* vol. 435, pp. 737-738, 2005.

[12] P. Bofetta, J. K. McLaughlin, C. La Vecchia, R. E. Tarone, L. Lipworth, and W. J. Blot, "False-positive results in cancer epidemiology: a plea for epistemological modesty," *Journal of the National Cancer Institute,* vol. 100, pp. 988-995, 2008.

[13] F. Prinz, T. Schlange, and K. Asadullah, "Believe it or not: how much can we rely on published data on potential drug targets?," *Nature reviews Drug discovery,* vol. 10, pp. 712-712, 2011.

[14] M. J. Farthing, "Research misconduct: A grand global challenge for the 21st Century," *Journal of gastroenterology and hepatology,* 2013.

[15] G. Francis, "Too good to be true: Publication bias in two prominent studies from experimental psychology," *Psychonomic Bulletin & Review,* vol. 19, pp. 151-156, 2012.

[16] S. Kepes and M. A. McDaniel, "How Trustworthy Is the Scientific Literature in Industrial and Organizational Psychology?," *Industrial and Organizational Psychology,* vol. 6, pp. 252-268, 2013.

[17] J. P. A. Ioannidis and T. A. Trikalinos, "An exploratory test for an excess of significant findings," *Clinical Trials,* vol. 4, pp. 245-253, 2007.

[18] U. Schimmack, "The ironic effect of significant results on the credibility of multiple-study articles," *Psychological Methods 17.4 (2012): 551.,* vol. 17, pp. 551-566, 2012.

[19] J. P. A. Ioannidis, "Why most discovered true associations are inflated," *Epidemiology,* vol. 19, pp. 640-648, 2008.

[20] L. V. Hedges, "Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences," *Journal of Educational and Behavioral Statistics,* vol. 9, pp. 61-85, 1984.

[21] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafo, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience,* vol. 14, pp. 365-376, 2013.

[22] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," DTIC Document 1975.

[23] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological science,* vol. 22, pp. 1359-1366, 2011.

[24] L. K. John, G. Loewenstein, and D. Prelec, "Measuring the prevalence of questionable research practices with incentives for truth telling," *Psychological science,* vol. 23, pp. 524-532, 2012.

[25] M. Kattenbraker, "Health education research and publication: ethical considerations and the response of health educators [Thesis]," *Carbondale, IL: School of Graduate, Southern Illinois University Carbondale,* p. 4, 2007.

[26] D. Fanelli, "How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data," *PloS one,* vol. 4, p. e5738, 2009.

[27] J. Ioannidis, I. Hozo, and B. Djulbegovic, "Optimal type I and type II error pairs when the available sample size is fixed," *Journal of clinical epidemiology,* vol. 66, pp. 903-910. e2, 2013.

[28] T. Dybå, V. B. Kampenes, and D. I. Sjøberg, "A systematic review of statistical power in software engineering experiments," *Information and Software Technology,* vol. 48, pp. 745-755, 2006.

[29] J. Cohen, "A power primer," *Psychological Bulletin,* vol. 112, p. 155, 1992.

[30] S. Wacholder, S. Chanock, M. Garcia-Closas, and N. Rothman, "Assessing the probability that a positive report is false: an approach for molecular epidemiology studies," *Journal of the National Cancer Institute,* vol. 96, pp. 434-442, 2004.

[31] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association,* vol. 90, pp. 773-795, 1995.

[32] R. Wetzels, D. Matzke, M. D. Lee, J. N. Rouder, G. J. Iverson, and E.-J. Wagenmakers, "Statistical evidence in experimental psychology an empirical comparison using 855 t tests," *Perspectives on Psychological Science,* vol. 6, pp. 291-298, 2011.

[33] D. Fanelli, ""Positive" results increase down the hierarchy of the sciences," *PloS one,* vol. 5, p. e10068, 2010.

[34] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. Sjøberg, "A systematic review of effect size in software engineering experiments," *Information and Software Technology,* vol. 49, pp. 1073-1086, 2007.

[35] S. Goodman and S. Greenland, "Why most published research findings are false: problems in the analysis," *PLoS medicine,* vol. 4, 2007.

[36] T. Dybå, B. Kitchenham, and M. Jørgensen, "Evidence-based software engineering for practitioners," *IEEE Software,* pp. 58-65, 2005.

[37] C. Wohlin, P. Runeson, M. Høst, M. C. Ohlsson, B. Regnell, and A. Wesslen, *Experimentation in software engineering*: Springer, 2012.

[38] D. E. Perry, A. A. Porter, and L. G. Votta, "Empirical studies of software engineering: a roadmap," in *Proceedings of the conference on The future of Software engineering*, 2000, pp. 345-355.

[39] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *Software Engineering, IEEE Transactions on,* vol. 28, pp. 721-734, 2002.

[40] J. Miller, J. Daly, M. Wood, M. Roper, and A. Brooks, "Statistical power and its subcomponents‚Äîmissing and misunderstood concepts in empirical software engineering research," *Information and Software Technology,* vol. 39, pp. 285-295, 1997.

[41] D. I. Sjoberg, T. Dyba, and M. Jorgensen, "The future of empirical methods in software engineering research," in *Future of Software Engineering, 2007. FOSE'07*, 2007, pp. 358-378.

[42] M. Ingre, "Why small low-powered studies are worse than large high-powered studies and how to protect against ‚Äútrivial‚Äù findings in research: Comment on Friston (2012)," *Neuroimage,* vol. 81, pp. 496-498, 2013.

[43] J. Cohen, "Things I have learned (so far)," *American psychologist,* vol. 45, p. 1304, 1990.

[44] R. E. Kirk, "Practical significance: A concept whose time has come," *Educational and psychological measurement,* vol. 56, pp. 746-759, 1996.

[45] F. A. Dahl, M. Grotle, J. Benth, and B. Natvig, "Data splitting as a countermeasure against hypothesis fishing: with a case study of predictors for low back pain," *European journal of epidemiology,* vol. 23, pp. 237-242, 2008.

[46] G. Cumming, "The New Statistics Why and How," *Psychological science,* vol. 25, pp. 7-29, 2014.

[47] M. A. van Assen, R. C. van Aert, M. l. B. Nuijten, and J. M. Wicherts, "Why Publishing Everything Is More Effective than Selective Publishing of Statistically Significant Results," *PloS one,* vol. 9, p. e84896, 2014.

[48] M. Levine and M. H. Ensom, "Post hoc power analysis: an idea whose time has passed?," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy,* vol. 21, pp. 405-409, 2001.

[49] K. Murayama, R. Pekrun, and K. Fiedler, "Research practices that can prevent an inflation of false-positive rates," *Personality and Social Psychology Review,* p. 1088868313496330, 2013.

[50] J. E. Hannay, E. Arisholm, H. Engvik, and D. I. Sjoberg, "Effects of personality on pair programming," *Software Engineering, IEEE Transactions on,* vol. 36, pp. 61-80, 2010.

[51] M. Ciolkowski, "What do we know about perspective-based reading? An approach for quantitative aggregation in software engineering," in *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*, 2009, pp. 133-144.

Appendix: The effect of significance level, statistical power, researcher bias and publication bias on the reliability of statistically significant results

In the following, we assume that our null hypothesis (H_0) is that there is no relationship, and the alternative hypothesis (H_a) is that there is a relationship between two variables. The observation that a test is statistically significant at level α is denoted S_α. The other variables are as described in Table X, Section Y. Notice that the model is based on the same structure and assumptions as the one in Section X.

Scenario 1: No researcher or publication bias (BF_0)

The Bayes Factor is in this case the probability of observing a statistically significant relationship when there is one, i.e., the statistical power of a study, divided by the probability of observing a statistically significant relationship if there is none, i.e., the significance level of a study.

$$BF_0 = p(S_\propto \mid H_a)/p(S_\propto \mid H_0) = (1-\beta)/\alpha$$

Scenario 2: Researcher bias, no publication bias (BF_1)

The probability of observing significant results given that the relationship is true, i.e., $p(S_\propto \mid H_a)$, increases with the probability of observing true negatives multiplied by the researcher bias, i.e., with β · rb. The probability of observing significant results given that the relationship is not true, $p(S_\propto \mid H_0)$, increases with the probability of false negatives times the researcher bias, i.e., (1- α) · rb.

$$BF_1 = p(S_\propto \mid H_a \wedge rb)/p(S_\propto \mid H_0 \wedge rb) = ((1-\beta)+\beta\cdot rb)/(\alpha+(1-\alpha)\cdot rb)$$

Scenario 3: Researcher and publication bias (BF_2)

The probability of observing non-significant tests decreases and, accordingly, the probability of observing significant tests increases. The Bayes Factor expression is based on the following elements, which are assumed to be true for both H = H_0 and H = H_a:

- The probability of observing a non-significant finding = 1 – the probability of observing a significant finding = 1 - $p(S_\propto \mid H)$.
- A publication bias means that the probability of not reporting a non-significant finding, given that a non-significant finding has been found, is (1- $p(S_\propto \mid H)$) · pb. The new (reduced) bias is the initial probability of reporting a finding minus the probability of not reporting a finding, i.e., 1 - (1 - $p(S_\propto \mid H)$) · pb.

$$BF_2 = \frac{p(S_\propto \mid H_a \wedge rb)}{1-(1-p(S_\propto \mid H_a \wedge rb))\cdot pb} \Bigg/ \frac{p(S_\propto \mid H_0 \wedge rb)}{1-(1-p(S_\propto \mid H_0 \wedge rb))\cdot pb}$$

$$= \frac{(1-\beta)+\beta\cdot rb}{1-(1-((1-\beta)+\beta\cdot rb))\cdot pb} \Bigg/ \frac{(\alpha+(1-\alpha)\cdot rb)}{1-(1-(\alpha+(1-\alpha)\cdot rb))\cdot pb}$$

To better illustrate the effect of publication bias on evidence strength, we may express this as:

$$= \frac{(1-\beta)+\beta\cdot rb}{(1-\beta\cdot pb)+\ \beta\cdot rb\cdot pb} \Bigg/ \frac{(\alpha+(1-\alpha)\cdot rb)}{(1-pb)+\alpha\cdot pb+(1-\alpha)\cdot rb\cdot pb}$$

As can be derived, the higher the publication bias, i.e., the closer *pb* is to 1, the lower the Bayes Factor. In the extreme case, applying the BF-expression with a total publication bias (*pb* = 1) gives a Bayes Factor of 1, i.e., there is no added value from observing a significant finding in a domain where none of the statistically non-significant findings are reported. This corresponds well, we believe, with common sense.