The Journal of Systems and Software 116 (2016) 133–145

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Incorrect results in software engineering experiments: How to improve research practices

Magne Jørgensen [a,b,*], Tore Dybå [b,c], Knut Liestøl [b], Dag I.K. Sjøberg [b]

[a] Simula Research Laboratory, P. O. Box 134, NO-1325 Lysaker, Norway
[b] University of Oslo, Norway
[c] SINTEF, Norway

ARTICLE INFO

Article history:
Received 6 October 2014
Revised 11 March 2015
Accepted 19 March 2015
Available online 28 March 2015

Keywords:
Controlled experiments
Empirical software engineering
Statistical hypothesis testing

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, given typical statistical power and significance levels, and to suggest improved research practices.
Method: First, we conducted a small-scale survey to document the presence 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 2002–2013 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.

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 in a domain 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*" (Ioannidis, 2005), is the origin of much of the recent discussions and concerns. There is, however, nothing new with concerns related to publication bias (not publishing statistically non-significant results) (Lane and Dunlap, 1978; Tannock, 1996), researcher bias (flexible analyses that lead initially statistically non-significant results to become significant) (Dingell, 1993; Masicampo and Lalande, 2012) and low statistical power (low likelihood of rejecting the hypothesis of no difference, the null hypothesis, even when there is a difference) (Tversky and Kahneman, 1971). Already in 1830, Babbage wrote about the decline in science, including what he called the "fraud of the observers" (Babbage, 1830). 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 (Bakker et al., 2012).

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 (Fanelli, 2012), i.e., increasingly competitive academic environments seem to increase not only the scientists' productivity, but also their biases (Fanelli, 2010). The use of questionable practices is hardly just a result of lack of knowledge about proper research practices. The survey reported in Martinson et al., 2005, 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 (Bakker et al., 2012; Bofetta et al., 2008; Farthing, 2014; Francis, 2012; Kepes and McDaniel, 2013; Prinz et al., 2011) are present in the context of software engineering experiments. If such problems are

* Corresponding author. Tel.: +47 924 333 55.
E-mail address: magnej@simula.no (M. Jørgensen).

http://dx.doi.org/10.1016/j.jss.2015.03.065
0164-1212/

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 (Ioannidis, 2005; Ioannidis and Trikalinos, 2007; Schimmack, 2012). 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 (Hedges, 1984; Ioannidis, 2008). The results presented in Shepperd et al., 2014 give strong indications that there are severe problems in the validity of some empirical software engineering results. Similarly, low statistical power is also likely to increase the proportion of incorrect results (Button et al., 2013).

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. The result of publication and researcher bias in a study with low statistical power

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 20 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 (Kincaid et al., 1975) 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 (14!) 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–Kincaid measure on the text. While in principle defendable, we made the outlier 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, sample sizes around 20 and less is common in software engineering experiments, where data collection typically is more expensive.

As much as 36% of the 196 software engineering experiments in the review reported later in this article had a sample size of 20 or less. Almost half of the experiments (47%) had a sample size of 25 or less.

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 Simmons et al., 2011. A study demonstrating how easy it may be to produce meaningless results in software engineering, amongst other based on researcher and publication bias, is presented in Zeller et al., 2011.

The remaining part of the paper is organised 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 papers describing in total 196 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 calculate the expected 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.

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 at least once. 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 at least once. Much fewer of the participants (10–22%), but still a noticeable proportion, admitted experiencing/conducting each of the questionable practises often.

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 these practices at least once. The practices and responses reported in our survey correspond well with those from a survey with similar questions in psychology