Publication Bias: A Detailed Analysis of Experiments Published in ESEM

Rolando P. Reyes Ch.
Universidad de las Fuerzas Armadas ESPE
Sangolquí, Ecuador
rpreyes1@espe.edu.ec

Oscar Dieste
Universidad Politécnica de Madrid
Madrid, Boadilla del Monte, Spain
odieste@fi.upm.es

Efraín R. Fonseca C.
Universidad de las Fuerzas Armadas ESPE
Sangolquí, Ecuador
erfonseca@espe.edu.ec

Natalia Juristo
Universidad Politécnica de Madrid
Madrid, Boadilla del Monte, Spain
natalia@fi.upm.es

ABSTRACT

Background: Publication bias is the failure to publish the results of a study based on the direction or strength of the study findings. The existence of publication bias is firmly established in areas like medical research. Recent research suggests the existence of publication bias in Software Engineering. *Aims:* Finding out whether experiments published in the International Workshop on Empirical Software Engineering and Measurement (ESEM) are affected by publication bias. *Method:* We review experiments published in ESEM. We also survey with experimental researchers to triangulate our findings. *Results:* ESEM experiments do not define hypotheses and frequently perform multiple testing. One-tailed tests have a slightly higher rate of achieving statistically significant results. We could not find other practices associated with publication bias. *Conclusions:* Our results provide a more encouraging perspective of SE research than previous research: (1) ESEM publications do not seem to be strongly affected by biases and (2) we identify some practices that could be associated with p-hacking, but it is more likely that they are related to the conduction of exploratory research.

CCS CONCEPTS

• **General and reference → Surveys and overviews**; **Reference works**.

KEYWORDS

Research bias, publication bias, experimentation, statistical errors, literature review, survey, exploratory research

ACM Reference Format:
Rolando P. Reyes Ch., Oscar Dieste, Efraín R. Fonseca C., and Natalia Juristo. 2020. Publication Bias: A Detailed Analysis of Experiments Published in ESEM. In *Evaluation and Assessment in Software Engineering (EASE 2020), April 15–17, 2020, Trondheim, Norway.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3383219.3383233

EASE 2020, April 15–17, 2020, Trondheim, Norway

ACM ISBN 978-1-4503-7731-7/20/04…$15.00
https://doi.org/10.1145/3383219.3383233

1 INTRODUCTION

Publication bias is defined as *failure to publish the results of a study 'on the basis of the direction or strength of the study findings'* [5]. Typically, studies containing positive (i.e., direction), or statistically significant (i.e., strength) results, are more likely to be published than studies reporting negative or non-significant results [27].

Publication bias has been observed in disciplines such as medical research [6, 7], but not in others, e.g., personnel selection research [25, pp. 493-495]. In Software Engineering (SE), Jørgensen et al. [12] reviewed the ratio of statistically significant tests in 150 randomly selected experiments. Such a ratio (51%) supports the existence of substantial publication bias in SE.

Publication bias promotes questionable research practices [20, Figure 1]: Failure to control for cognitive biases, analytical flexibility or data dredging (*fishing* or *p-hacking*) [21]. Ioannidis [10], in a foundational paper, argued that most experimental results are false, pointing out almost the same reasons: little criticism of the posed research questions, excessive design and analysis flexibility, and research biases.

In a previous study [23], we searched for statistical errors in experiments published in the International Conference on Software Engineering (ICSE). We detected that a significant amount of experiments do not report statistical hypothesis, do not check the assumption of the inference tests, and perform multiple testing without any correction to avoid increasing type-I errors. These practices are typically associated with publication bias.

The Int'l Workshop on Empirical Software Engineering and Measurement (ESEM) is the flagship conference on experimental SE. In this research, we review experiments published in ESEM searching for practices associated with publication bias. We also survey with SE experimenters to triangulate the literature review results.

The **contribution of this paper is twofold**:

- We provide a more encouraging image of SE research than previous research [12, 23]. ICSE is the flagship SE conference, but it has a general, i.e., non-experimental, character. The same applies to the sources surveyed by Jørgensen and colleagues. Although the current situation can be improved, ESEM publications do not seem to be strongly affected by biases.
- We identify some practices that could be associated with *p-hacking*. However, a more likely explanation is our almost

complete ignorance about SE phenomena, that leads the conduction of a good deal of exploratory research.

This paper is organized as follows: Section 2 introduces the practices that may suggest the existence of publication bias. Section 3 state the research questions. Section 4 reports the literature review, and Section 5 the survey to experimenters. The paper finishes with the threats to validity in Section 6, and the conclusions in Section 7.

2 BACKGROUND

Publication bias means papers containing statistically significant results have a higher likelihood of being published. Experimenters may be tempted to carry out some questionable practices that increase the chances to achieve statistical significance. According to Munafò et al. [20, Figure 1], these practices are[1]:

2.1 Failure to control for bias

Once experimental data is available, data visualization and exploratory analyses are usual. These practices cannot be questioned in principle, but they convey a risk: researchers may perceive patterns or regularities that suggest relationships between variables. Inadvertently (or not), these relationships can find their way into the study as genuine hypotheses.

HARKing (Hypothesizing After Results are Known) [15] means that *post hoc* hypotheses are presented and analyzed as they were *a priori* hypotheses. When the patterns/regularities are strong enough, *post hoc* hypotheses yield statistically significant results independently of any other consideration, e.g., statistical power.

HARKing can adopt different forms [15, pp. 197-198]. When *post hoc* hypotheses are presented as *a priori* hypotheses, HARKing cannot be detected. However, some other practices indicate (but do not confirm) that HARKING is present:

- The experiment does not contain hypotheses, but inference tests are applied to the data. Although these tests can be based on non-reported *a priori* hypotheses, such tests may be conducted on an opportunistic or exploratory manner.
- The experiment contains, besides the experimental hypothesis, additional hypotheses. They typically explore relationships between experimental and non-experimental (e.g., demographics) data. These hypotheses are usually termed *post hoc* hypotheses in the experimental literature.

2.2 Analytical flexibility

Experimenters have complete control of the data acquisition process and analysis procedures. Harmless decisions, e.g., outlier removal, dataset reduction, or the introduction of controlled variables (gender, experience), can influence the statistical significance of the tests[2]. Again, as in the case of the HARKing, it is generally impossible to assess the researchers' intention because it is not reflected in the written reports.

One exception is the choice of inferential tests. For most designs, simple tests suffice, such as ANOVA, or the corresponding non-parametric counterpart (Kruskal-Wallis). When unusual, sophisticated tests are used, that is an indication (again, not a confirmation) of *p-hacking* [29, pp. 147-150].

2.3 Data dredging

Data dredging is the practice of performing comparisons within a dataset to achieve statistically significant results [9, pp. 169]. This practice is favored when a large number of independent and, more frequently, dependent variables, are used in a study. Data dredging can take several forms:

- Performing multiple comparisons, beyond what is required to test the statistical hypotheses.
- Posing multiple hypotheses [3] is also an indication (again, not a proof) of data dredging. In most cases, these hypotheses test the relationship between a single independent variable and multiple dependent variables. In practice, they are equivalent to multiple testing.
- Making the analysis more powerful. There are several possible procedures (e.g., switching tests), but none of them guarantee that the power increases. One exception is switching the tails of the tests, from 2-tailed to 1-tailed. 2-tailed tests are inherently less powerful. 1-tailed tests are sometimes used not because existing knowledge suggests a directional effect, but to increase the chances of achieving statistically significant results.

3 RESEARCH QUESTIONS AND METHODOLOGY

Publication bias seems to affect SE [12] , as well as other more mature disciplines [6, 7]. However, experimental research in SE has not achieved standardization. At least in principle, experimental research may have different levels of rigor in each particular community. ESEM aims to be the "the premier conference for presenting research results related to empirical software engineering" [1]. Thus, we wonder:

RQ1: Is there evidence of publication bias at ESEM?

To answer this question, we have conducted a literature review of experiments published in ESEM between 2007 and 2016 (10 years in total), seeking signs of failure to control for cognitive biases, analytical flexibility, and data dredging.

The literature review is reported in Section 4. Unfortunately, most of the evidence is ambiguous. For instance, the lack of an explicit hypothesis definition does not automatically imply HARKing. Likewise, the switch of a 2-tailed test into a 1-tailed one suggests a *p-hacking*, but we cannot rule out alternative explanations, e.g., a mistake overlooked by authors and reviewers. Therefore, we propose another research question:

[1] Figure 1 in [20], and the caption of the same figure, are not totally consistent. We provide our personal interpretation herein.

[2] These practices can also be seen as instances of data dredging, e.g., see [9, pp. 169]. Different authors use the same terms with (slightly) different meanings. We provide here a coherent but necessarily partial picture.

[3] Multiple hypotheses are also an indication of HARKing, but it cannot be assessed in written reports.

RQ2: Why do researchers carry out questionable practices?

To answer this question, we have conducted a survey with SE experimenters (see Section 5). Their answers contextualize the literature review and triangulate our findings.

4 LITERATURE REVIEW

To answer RQ1, we have conducted a systematic literature review according to Kitchenham et al. [16]. The following sections describe the review objectives, design, execution, results, and the main findings obtained from the review.

4.1 Review objectives

The literature review aims to identify whether experiments published at ESEM carry out practices (described in Section 2) that suggest publication bias. More concretely, we propose the following review objectives:

- **Failure to control for bias**:
  (1) Are hypotheses explicitly defined?
  (2) Are there *post hoc* tests?
- **Analysis flexibility**:
  (3) Which analysis procedures do ESEM experiments apply?
- **Data dredging**:
  (4) How many explicit hypotheses are defined per experiment?
  (5) How many tests are conducted per experiment?
  (6) What is the ratio of statistically significant tests?
  (7) Does the hypothesis tail match the analysis tail?
  (8) Are 1-tailed tests used to increase power?

4.2 Inclusion and exclusion criteria

Papers qualify as experimental when they have:

- *Group assignment*: experimental units are assigned (randomly or not) to groups.
- *Comparative goal*: they aim to compare some response variables across groups.
- *Inference*: (frequentist) statistical tests are used to reveal differences among groups.
- *Experimental data*: The data is generated as a result of the experimental manipulations; previously existing data is not used.

The 1st inclusion criterion asserts the paper's experimental (or quasi-experimental) character [26]. The 2nd and 3rd guarantee the use of hypothesis testing. The 3rd criterion would have excluded experiments analyzed under the Bayesian framework if there were any; none of the reviewed papers applied Bayesian statistics.

A large number of papers use existing datasets, e.g., PROMISE repository [24], data extracted from open source repositories, etc. The studies that rely on these data do not have an experimental character, i.e., the data is not obtained as the result of some experimental manipulation. As the data is pre-existing (either readily available, as the case of the PROMISE repository, or available to be processed, as the case of open-source repositories), such studies can be properly termed as observational studies. The 4th criterion makes a distinction between experimental (or quasi-experimental) and observational studies. The latter are excluded from our literature review. Such exclusion does not lead to negative results. In fact, observational studies depart widely from usual experimental standards; their inclusion would have made the review results considerably worse.

4.3 Execution

We reviewed experimental papers published in ESEM between 2007-2016. Two researchers (R. Reyes & O. Dieste) screened[4] the titles, abstracts, and keywords of all papers independently, using the ACM and IEEE digital libraries. Discrepancies were solved by consensus.

Fleiss' $\kappa = 0.62$, representing substantial agreement [8]. In total, 387 papers were screened and 55 papers initially selected. After a detailed review, 6 papers were removed because they did not meet the inclusion criteria. Finally, the primary study set was composed of 49 experiments.

The first author (R. Reyes) created an extraction form[5], which stems from the review objectives described in section 4.1. Two researchers (O. Dieste & R. Fonseca) piloted the form, suggesting several changes and general improvements to the wording. Two researchers (R. Reyes & O. Dieste) extracted the information from the primary studies independently. Both datasets were later compared and corrective measures were taken in case of disagreement.

4.4 Review results

The collected data were summarized using 2-way tables and tree-like representations. The results are described below.

4.4.1 Are hypotheses explicitly defined? Shortly after commencing the review, we realized that the primary studies define hypotheses at different levels of abstraction:

- Research goal or aim: A high-level statement describing the purpose of the experiment, e.g.:
  We wanted to find out whether CFT are less error-prone and whether the participants would favor CFT over FT [13].
- Research hypothesis: A declaration of the relationship between independent and dependent variables that drives the research design [19]. For instance:
  H_{21}*: When using CFT, consistency between the system description and the safety analysis model is perceived differently than when using FT.* [13]
- Research questions: A research goal breaks down into several research hypothesis rather frequently. Although unnecessary from a technical viewpoint, researchers tend to create *research questions* to report this refinement process explicitly. For instance, the research questions below appear also in [13]:
  RQ_1*: Will the application of the CFT yield the same quality of the resulting safety model as a model built with FT?*

[4]Details are available at https://goo.gl/PSjjQu.
[5]The form and the raw data are available at https://goo.gl/PSjjQu.

RQ_2: Is CFT perceived differently than FT with regard to consistency, clarity, and maintainability?

- Statistical hypothesis: Finally, research hypotheses can adopt an analytic formulation, using the usual null/alternate format. The statistical hypothesis contains the keys elements (estimators, tails, etc.) that drive statistical analysis [19]. Again using [13] as example:
  $H0_{21}: \mu CFT = \mu FT$
  $H_{21}: \mu CFT \neq \mu FT$

The previous elements (research goal/question/hypothesis, and statistical hypothesis) do not usually appear together in the same experiment. Jung et al. [13] is one of the few cases in which the four are explicitly reported. In turn, it is rather common that one or several are missing (in particular, the research hypotheses and the statistical hypotheses). The same problem has been observed in other areas [2].

Fig. 1 summarizes how ESEM experiments define hypotheses. Almost all experiments (96%) specify the research goal. However, only 55% of them contain research hypotheses and, looking further down, the declaration of statistical hypotheses (29%) decreases considerably.

The tails of the tests are defined in a larger proportion (49%) than the statistical hypotheses. The reason is that researchers specify the type of tail in the research hypothesis rather frequently. The leftmost branch corresponds with the orthodox use of hypothesis testing; only 12% of the experiments conform to it.

27% of the experiments do not include research or statistical hypotheses. This is rather unusual since a declaration of purpose (which variables are being examined, and why) predates experimental operation, not to mention it goes against the recommendations of standard experimentation textbooks and guidelines, e.g.: Juristo & Moreno explicitly mention research hypotheses [14, pp. 49]; both Wohlin et al. and Jedlitschka et al. emphasize the usage of statistical hypotheses [11, 30]. The underlying reason seems to be the widespread usage of research questions instead of hypotheses. Table 1 shows that 70% of the papers contain RQs but not research hypotheses. Table 2 reports the same figure[6] (70%) for the statistical hypotheses.

Table 1: RQs vs. Research Hypotheses Crosstab

	Have research hypothesis	Do not have research hypothesis	Total
Have RQs	14 (29%)	20 (41%)	**34 (70%)**
Do not have RQs	13 (26%)	2 (4%)	**15 (30%)**
Total	**27 (55%)**	**22 (45%)**	

Table 2: RQs vs. statistical hypotheses crosstab

	Have statistical hypothesis	Do not have statistical hypothesis	Total
Have RQs	10 (21%)	24 (49%)	**34 (70%)**
Do not have RQs	4 (8%)	11 (22%)	**15 (30%)**
Total	**14 (29%)**	**35 (71%)**	

[6]The values are alike by chance; it is not a mistake.

4.4.2 Are there post hoc tests? Only $\frac{15}{49} = 31\%$ of the experiments contain *post hoc* tests. *Post hoc* tests were conducted after the inference tests. In 12 out of 15 cases (80% of the total), inference tests already yielded statistically significant results, so *post hoc* tests were unnecessary (from the publication bias perspective).

The *post hoc* tests performed in each experiment have their own peculiar characteristics, but they can be roughly classified into three groups:

- Subjects are decomposed into subgroups on the basis of some characteristic, e.g., experience, and the hypotheses are re-tested.
- During hypothesis testing, some independent variables that have a role in the experimental design were not examined. They are examined during *post hoc* testing.
- Correlations (usually bivariate) are run between independent and dependent variables.

4.4.3 Which analysis procedures do ESEM experiments apply? Fig. 2 shows (besides other aspects that we will not discuss in this moment) the types of tests used in ESEM experiments.

None of the tests is uncommon; actually, they constitute a rather basic statistical toolset, e.g., t-test, Mann-Whitney, Wilcoxon, ANOVA, Kruskal-Wallis, etc. The *Friedman* and *McNemar* tests (a non-parametric repeated-measures and an alternative to the Fisher exact test, respectively) are somehow unusual, but by no means "sophisticated".

4.4.4 How many explicit hypotheses are defined per experiment? Fig. 3 shows a histogram. The x-axis represents the number of hypotheses defined per experiment, and the y-axis the number of experiments of each kind. Most of the experiments declare $\leq$ 2 hypotheses (2.41 on average). Experiments with more than 4 hypotheses are rare.

4.4.5 How many tests are conducted per experiment? There is a large difference between the number of hypotheses per experiment and the number of tests actually conducted. As shown in Fig. 4, most of the experiments conduct $\leq$ 12 tests. The average is 7.95, roughly 3 times the number of hypotheses.

The reason for the difference is that research hypotheses typically make reference to the main factor of interest, but not to other independent variables of the design, e.g., task, type of object, etc. In some cases, e.g., non-normality, the independent variables are not analyzed jointly, e.g., using ANOVA, but they are tested in sequence using non-parametric tests. This increases the number of tests as compared to the number of hypotheses.

4.4.6 What is the ratio of statistically significant tests? On average, an experiment published in ESEM contains 3.87 statistically significant tests on average. However, there is a large variation at the level of individual experiments. As shown in Fig. 5, roughly half of the papers have ≤ 1 significant tests only. Overall, there are 179 statistically significant tests out of 480 conducted tests. This gives a significant test ratio of $\frac{179}{480} = 0.37$.

4.4.7 Does the hypothesis tail match the analysis tail? In addition to the statistical tests used in ESEM experiments, Fig. 2 shows the tails associated with those tests. Each test is related to two tails: the one defined in the corresponding research/statistical hypothesis and the actual tail employed in the analysis.

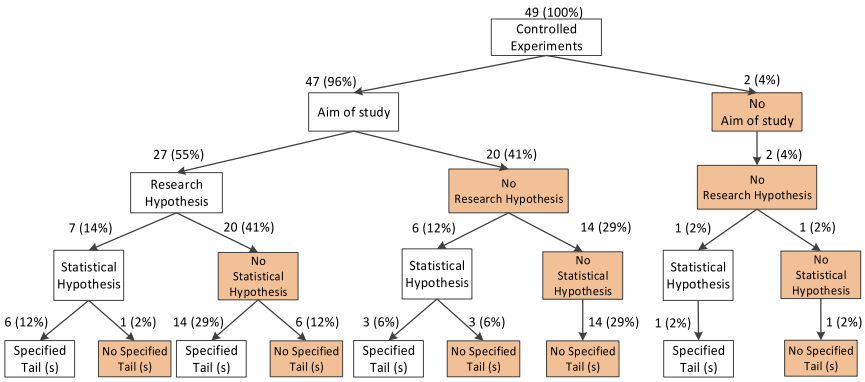

Figure 1: Relationships among aims, research hypotheses, statistical hypotheses and tails

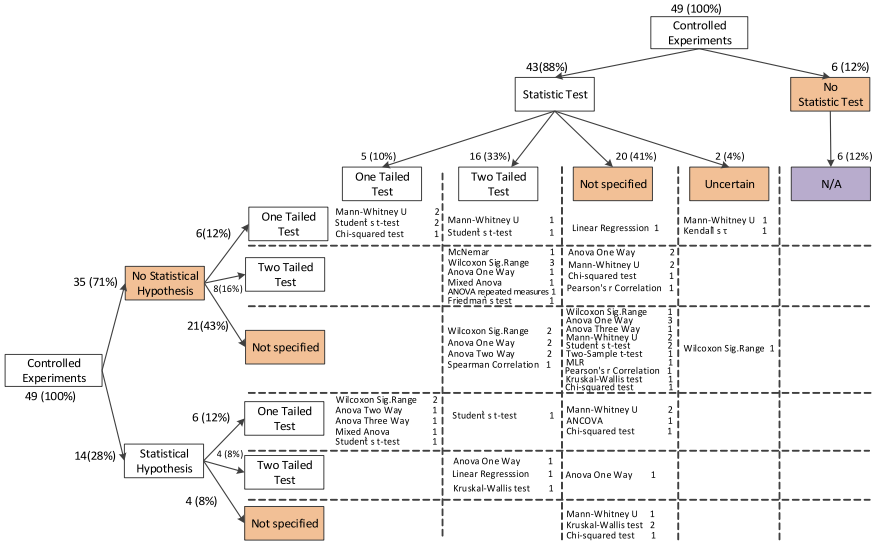

Figure 2: Relationship between statistical hypotheses and tests

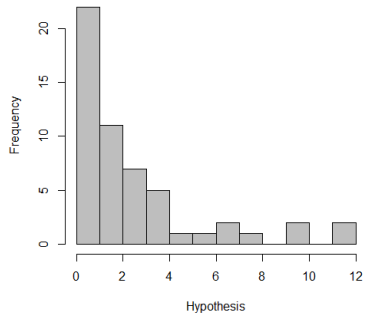

Figure 3: Number of hypotheses are defined per experiment

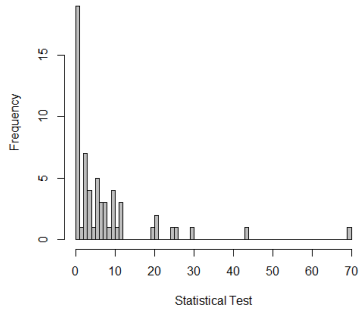

Figure 4: How many tests are conducted per experiment?

In almost all cases, the tail of the hypothesis and the tail of the test match. Only in three cases, 1-tailed hypotheses were tested using

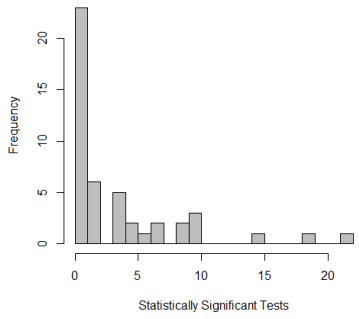

Figure 5: How many statistically significant tests are conducted per experiment?

2-tailed tests. The most likely cause is a mistake in the definition of the hypotheses' tails. Switching from a 1-tailed to a 2-tailed test does not increase power, but actually decreases it. Those mistakes cannot be thus associated with research bias.

A challenge to this interpretation comes from the fact that, in a large number of cases, the definition of the tail is missing either in the research/statistical hypothesis and/or the test. In our opinion, the lack of definition of the tests cannot be understood as evidence of publication bias. The vast majority of tests specifying the type of tail are 2-tailed; there is not any reason to think that *all* the other tests are 1-tailed. In this situation, concealing the information about the tails would have had no effect in terms of power.

Although not related to publication bias, we wish to point out in passing some inconsistencies in Fig. 2. ANOVA's are inherently 2-tailed. Even so, they have been used as 1-tailed tests in four cases. Likewise, Chi-squared tests have been used in the primary studies to test contingency tables exclusively. Contingency tables are inherently 2-tailed, but in one case the authors claim that the test is 1-tailed. Similar problems could have happened in other tests whose tails are not explicitly declared. In any case, these problems do not seem to be associated with publication bias, but to a shallow spread of statistical knowledge in the SE community.

4.4.8 Are 1-tailed tests used to increase power? Table 3 shows the relationship between the types of tails of the statistical tests (notice that not all papers define the tails) and the results of the corresponding inference tests. When 1-tailed tests are conducted, the ratio of statistically significant tests is $\frac{51}{51+26} = 66\%$. In the case of 2-tailed tests, the ratio is $\frac{61}{61+74} = 45\%$. The obvious conclusion is that 1-tailed tests are associated with higher rates of statistically significant tests.

4.5 Review findings

For our perspective, the main findings of the review are:

- **Failure to control for bias**:
  (1) Are hypotheses explicitly defined? **No**. Only 55% of the papers contain a research hypothesis. However, it is doubtful that the lack of definition is associated with HARKing. When the research hypotheses are not present, the authors provide research questions. Hypotheses may or may not define the type of tails. The current situation can be explained more easily by inconsistent reporting and/or limited statistical expertise than HARKing.
  (2) Are there *post hoc* tests? **Yes, but in lower rates than expected**. Only 31% of the papers contain *post hoc* tests. Most of these *post hoc* tests (80%) are unnecessary because the experiment already has produced statistically significant results.
- **Analysis flexibility**:
  (3) Which analysis procedures do ESEM experiments apply? **Regular procedures**, such as t-test, ANOVA or their non-parametric counterparts. Sophisticated tests are not used.
- **Data dredging**:
  (4) How many explicit hypotheses are defined per experiment? **2.41 hypotheses on average**. Most of the experiments declare ≤ 2 hypotheses.
  (5) How many tests are conducted per experiment? **7.95 tests in average**. There are more tests than hypotheses due to the existence of independent variables not mentioned in the hypotheses.
  (6) What is the ratio of statistically significant tests? **0.37**. This value is higher than the ratio reported by Jørgensen et al. [12] (51%).
  (7) Does the hypothesis tail match the analysis tail? **Yes, in general**. There are some inconsistencies and errors, but no evidence of bias. Most of the tests lack an explicit definition of the type of tail. This problem seems to be connected to the confusion around research/statistical hypotheses and tails.
  (8) Are 1-tailed tests used to increase power? The researchers' intentions cannot be inferred from the data. However, the high number of statistically significant results associated to 1-tailed tests **suggests a positive answer**.

Table 3: Relationship between the tail type and the significance of statistical tests

		Results by Total Hypothesis		
		Sig.	No. Sig.	Uncertain
Type of tail selected	1-Tailed	51	26	0
	2-Tailed	61	74	5
	Not specified	53	142	2
	N/A	1	23	5
	Uncertain	15	34	3
Total		**181**	**299**	**15**

5 SURVEY TO SE EXPERIMENTERS

The literature review reported in the previous section shows some issues in the conduction of experimental research in ESEM, e.g., the inconsistent use of research/statistical hypotheses. RQ2 aims to clarify these issues. We performed a survey with SE experimenters, inquiring how they plan an experimental study and their beliefs about the associated concepts, e.g., the type of tails. The sections below report the survey design, execution, and results.

5.1 Survey design

The survey design is based on Kitchenham et al. [18] and Punter et al.'s [22] guidelines. The questions try to clarify why researchers perform some practices that surfaced during the literature review, in particular:

(1) The lack of clear relationships between research and statistical hypotheses, and the associated tails.
(2) The seemingly arbitrary choice of 1-tailed and 2-tailed tests.
(3) The errors and inconsistencies in statistical tests.

The first version of the survey consisted of 21 mandatory questions and six optional questions. Several questions address the same topic to avoid misinterpretations; this makes the survey more time consuming than we initially wanted. The respondent is allowed to add comments or opinions using free text. Five mandatory questions were removed, and eight new ones (mandatory and optional) were created after a thorough review by O. Dieste and N. Juristo. In a second stage, two independent researchers (M. Solari y O. Gómez) piloted the survey; their feedback improved the text of 4 optional questions. The final version of the survey is available at https://goo.gl/QS1ati.

The population is defined as any SE experimenter. The sample was collected as follows: We collected the emails of the experimenters who published experiments in the most representative conferences and journals of the experimental Software Engineering community[7] between 2012 and 2015, i.e., ESEM, the International Journal on Empirical Software Engineering (EMSE), the IEEE Transactions on Software Engineering (TSE), and the International Conference on Software Engineering (ICSE). We obtained a total of 403 authors' emails (95 of ESEM, 93 of EMSE, 124 of TSE, and 91 of ICSE). When the same author appears in several outlets, she/he is only considered once (in the following order: ESEM, EMSE, TSE, ICSE).

5.2 Execution

Respondents were contacted by email and periodically reminded. The data were collected in approximately 3 months. Each group of authors (ESEM, EMSE, TSE, ICSE) responded to the survey at a different time, so the provenance of each respondent could be identified (in all other respects, the identity of the respondent was not known to the researchers).

The experimenters completed the survey in 8-12 minutes (depending on the feedback provided). A survey of this duration is unlikely to produce fatigue, to the point of representing a threat to validity [18]. 45 researchers answered the questionnaire (29 in full, 16 partially). The response ratio was 11.20% (45/403), which is comparable with the response ratios achieved by other SE surveys, e.g., [4]. The data obtained are available at https://goo.gl/1X3Px9.

5.3 Results

The survey results are summarized in trees and double-entry tables, as in the previous section. The respondents' comments (submitted as free text) have been coded to highlight the underlying themes.

[7]Retrospectively, we believe that the Information and Software Technology journal (IST) had to be included as well. Nevertheless, notice that the authors that publish at IST and the other four outlets overlap considerably.

5.3.1 Relationship between the research hypothesis, statistical hypothesis, and tails. The vast majority of researchers (86%) claim that they include research hypotheses in their reports, as shown in Figure 6. This number is 30 points higher than Fig. 1, where only 55% of the experiments contain research hypotheses. The same pattern (30% difference) appears in the statistical hypotheses. 59% of researchers say that they always include statistical hypotheses in their investigations, while Fig. 1 indicates that this only occurs in 29% of the cases.

It could be argued that the articles published in ESEM are particularly defective. Fig. 7 shows the equivalent to Fig. 1, but restricting the responses to the experimenters who have published in ESEM. The experimenters argue that they include research/statistical hypotheses in 80% and 60% of the cases, respectively.

Figure 6 shows that researchers know the hypothetical-deductive method. Only a minority of researchers (3%) state that they do not use statistical hypotheses, or specify the tails of alternative hypotheses (17%). It stands in stark contrast with the literature review results, where the figures are the opposite (71% and 51% respectively).

When a participant provided a response apparently in opposition to the usual practice (e.g., she/he does not define statistical hypotheses), the survey asked for the reasons why. After a process of refinement and classification (available at https://goo.gl/FTxnV5), we obtained the following advice from experimenters:

- **Research hypothesis:** There are two noteworthy aspects: a) A significant number of researchers believe that the research hypothesis can be easily inferred, and b) *exploratory studies*[8] do not need research hypotheses.
- **Statistical hypothesis**: The reasons given are essentially the same as above: a) they are obvious, or b) they are not useful in exploratory studies. A third reason is that a decision about which hypotheses to test cannot be taken during the experimental design and it should be deferred until the experiment has been executed. This goes against the recommended practice, although a more likely interpretation is that the respondent is referring to conducting exploratory studies.
- **Tails**: The situation is, again, quite similar: a) the tails are evident or can be inferred, or b) they are useless since they can not be defined during the design (that is, in exploratory studies). The researchers show a preference for the use of two tails, which could be related, again, with the exploratory studies, since the lack of the direction of the effect is, precisely, one of their main characteristics.

5.3.2 Choice of 1-tailed or 2-tailed tests. In the previous section, we have reported that experimenters exhibit a preference for 2-tailed tests. The Tables 4 and 5 explain the reason why. In 23% of cases, researchers indicate that 2-tailed tests should always, or in most situations, be used. 1-tail tests are generally discouraged. In 14% of the cases, experimenters believe that they should never be used.

The reasons given for the choice of 1-tailed or 2-tailed tests are completely reasonable. In 72% of cases, the researchers indicate

[8]Experimenters differentiate between exploratory and confirmatory studies. The former seek relationships between variables, whereas the latter try to confirm/reject given research/statistical hypotheses.

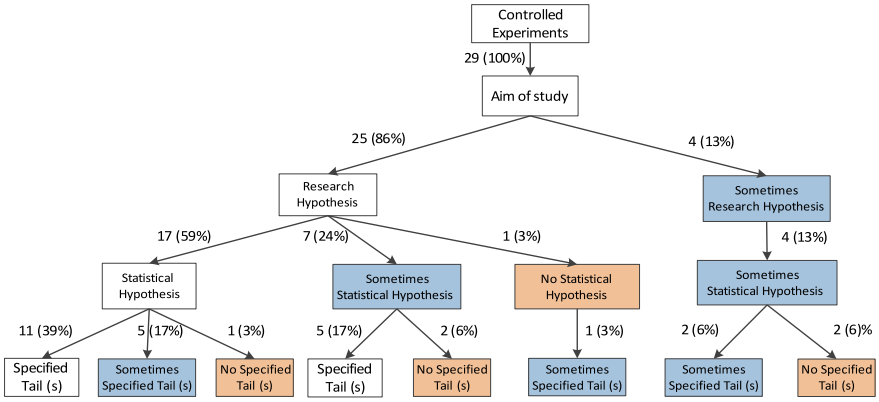

Figure 6: Relationship between research hypotheses and statistical hypotheses

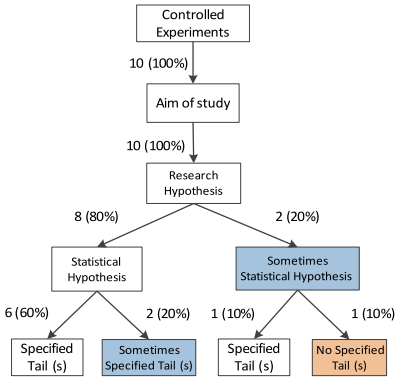

Figure 7: Relationships between aims, research hypotheses, statistical hypotheses, and tails

Table 4: Situations in which 1-tailed test should be used

When 1-tailed tests should be used?	Frequency	Questionable?
-	1	
When the effect has a clear direction	3	
When the researcher is only interested in certain effect direction	3	
The effect can only occur in a certain direction, according to prior knowledge	4	
It depends on the existing knowledge about a phenomenon	1	
The effect can only occur in one direction	1	
The hypothesis is directional	9	
The distributions are asymmetrical	1	Probably
Never in SE	3	
Never, unless there is a good reason to assume an effect in a certain direction	1	
Post-hoc tests	1	Yes
Test the alternative hypothesis	1	Maybe
Total	**29**	

Table 5: Situations in which 2-tailed test should be used

When 2-tailed tests should be used?	Frequency	Questionable?
-	3	
Any difference is of interest	3	
The statistical power is enough	1	Yes
In most instances	5	
There is more than one independent variable	1	Yes
The statistical distribution has two tails	1	Yes
The null hypothesis suggests that there is no difference between two distributions	2	Yes
No, there should be prior knowledge that indicates the direction of the effect	1	
The effect direction is unknown	11	
Always, since in SE it is unlikely that the effects have a well-defined direction	2	
Total	**29**	

that the 1-tailed tests should be used when: a) the direction of the effect is known, or b) the experimenter is only interested in one of the directions of the effect. For the 2-tailed tests, the situation is similar; in 47% of cases, the reason given is the uncertainty of the effect's direction.

A remarkable aspect in Tables 4 and 5 is that some responses are questionable, perhaps pointing at issues with experimental statistics.

5.3.3 Usage of statistical tests. We have tried to find out to what extent the researchers correctly handle the concepts of 1/2 tails at the level of statistical tests. Hence, we have made 3 specific questions:

- **Which tails can be used with which statistical tests?** The responses (experimenters could select both the 1-tail and 2-tail options, so the percentages do not add up to 100%) are shown in Table 6. Some answers were incorrect (shaded cells in the table). 31% of experimenters say they use one tail with the ANOVA, and 41% with the Kruskal-Wallis tests when both are inherently 2-tailed. A similar figure (31%) is

Table 6: Which statistical tests do you use 1-tailed tests or 2-tailed test with?

	1-tailed	2-tailed
Student's t-test (paired/unpaired)	62%	62%
Mann-Whitney / Wilcoxon (paired/unpaired)	59%	79%
Chi-squared test	**31%**	59%
ANOVA	**28%**	66%
ANOVA With Repeated Measures	**10%**	38%
Kruskal-Wallis test	**41%**	55%
Linear regression	**3%**	31%

obtained for the Chi-squared test, which is almost universally used for contingency table analysis, also inherently 2-tailed. It is highly unlikely that the experimenters are thinking about the underlying statistic. For instance, F-test is used to assess the statistical significance in the case of the ANOVA. The F-test is 1-tailed, but the ANOVA is 2-tailed. However, the percentage of responses in the 2-tailed column for ANOVA is higher (and also correct) than the percentage of the 1-tailed column; in our opinion, the answers in the 1-tailed column are mistakes.

- **Have you tried several statistical techniques during the analysis?:** 7% of researchers say they do it. 52% of researchers say they have done so, at least on some occasions.
- **Have you ever made a change in the type of tail during the analysis?** The vast majority of the experimenters never switched 1-tailed tests (from $\leq$ to $\geq$, and *vice versa*) or 2-tailed tests into 1-tailed ones. However, a sizable number of experimenters (21%) switched from 2-tailed to 1-tailed tests, as shown in Table 7. When this occurs, significant results were obtained in 10.5% of cases.

Table 7: Did you ever change a 2-tailed hypothesis into a 1-tailed one?

			When you switched the 2-tailed hypothesis into a 1-tailed one, did it turn out that a statistically significant result came to light?	
Did you ever change a 2-tailed hypothesis into a 1-tailed one?	Yes	21%	Yes	10.5%
			No	10.5%
	No	79%	-	

5.3.4 Survey findings. In our opinion, the main findings of the survey are:

(1) Relationships between research, statistical hypotheses, and tails: Theoretically speaking, **researchers know how to use the statistical concepts**. However, they **do not apply them in practice**. The reason is that experiments do not seem these concepts (research hypothesis, tails) useful to **conduct exploratory research**.
(2) Choice of 1-tailed and 2-tailed tests: Same as before, **experimenters know when to use 1-tailed or 2-tailed tests**. However, some experimenters **make mistakes when proposing usage scenarios**.
(3) Usage of statistical tests: A sizable number of experimenters **make mistakes in the types of tails associated with some tests**. In some cases, **tails are switched**, with the likely intention of increasing the power of the test.

6 THREATS TO VALIDITY

The threats to validity will be reported following Yin [31] and Creswell [3]. Both works, compared to Shadish et al. [26], take into consideration qualitative studies (as this one). According to those authors, threats to validity are classified into four types: internal, external, construct and reliability.

Internal validity: It refers to the inferences made on data. The existence of unknown variables may influence the results of the statistical tests or other analysis procedures [3]. This threat can operate both in the literature review and the survey. To mitigate this threat:

- The literature review was conducted by three researchers. All relevant steps (paper screening, selection, data extraction, and analysis) were performed collaboratively. We set controls, e.g., double-check of the data extraction forms, recalculation of tables, etc. to guarantee the quality of the data.
- The survey was piloted and evaluated by independent researchers. The number and complexity of the questions were limited to avoid respondent's fatigue (the survey could be filled out completely in less than 15 minutes). To improve the veracity/accuracy of the responses, participation was voluntary, and anonymity was secured.

External validity: This threat appears when we generalize the results beyond the context in which they were obtained. The risks are different for the review and the survey:

- The literature review was specifically targeted to ESEM experiments and, consequently, the "community" behind these studies. We do not aim to generalize to other communities; in fact, the specific analysis of the ESEM community is one of the goals of this paper. Further research, e.g., targeting other venues where experimental papers are regularly published, would be necessary to assess the situation of the general SE community.
- The survey has a more general character. For that reason, the sample was not restricted to ESEM researchers. We also include researchers that published experimental papers at EMSE, ICSE, and TSE. We believe that the researchers that publish in these outlets are representative of the general population of experimental researchers.

Construct validity: This threat operates when the study uses variables and metrics that do not represent the underlying theoretical constructs accurately. We have addressed this threat conducting previous research on (1) statistical errors in SE [23] and (2) experimental problems in the sciences (not published yet). These previous studies allowed us to design rigorous instruments (data collection forms, questionnaire); these instruments were also double-checked and/or piloted.

Reliability: It refers to how trustable and repeatable the study results are. To mitigate this threat, we have documented (as much as possible) all methodological steps, e.g., study screening and primary study selection. Documents are available in Google Drive™; the URLs are have been provided throughout the manuscript. Likewise, the raw data and analysis results are publicly available in the same form.

7 DISCUSSION AND CONCLUSIONS

We cannot provide an authoritative answer to **RQ1: Is there evidence of publication bias at ESEM?** The signs are ambiguous. On the one side:

- Hypotheses and tails are left undefined frequently.
- A large percentage of papers perform *post-hoc* tests.
- Multiple testing is the norm. The number of tests conducted on average per experiment is 7.95.
- 1-tailed tests are associated with higher significant test ratios.

On the other side, these practices do not seem oriented to achieve statistically significant results that underlie publication bias:

- Most (80%) of the *post-hoc* tests were unnecessary because the experiment yielded significant results already.
- The ratio of statistically significant tests is 0.37, quite close to the average power of SE experiments, as reported by Jørgensen et al. [12].

In our opinion, **publication bias is not strongly influencing the research agenda, not at least in ESEM**. Experimenters perform some questionable practices: (1) lack of definition of the experimental hypotheses and (2) multiple testing. Both practices probably increase the significant test ratio (0.3 is likely higher than the actual average power of SE experiments) and consequently make some (unintended) pressure on authors, reviewers, and editors.

However, such increment seems to be a collateral effect of how SE experiments are being conducted, but not a major driver. **RQ2: Why do researchers carry out questionable practices?** aimed to find out why researchers perform those questionable practices. The most frequent answer was that they perform **exploratory research**. It is generally agreed that the SE discipline lacks sound scientific knowledge. It is perfectly reasonable, in our opinion, that experimenters perform "reconnaissance" studies, without clear *a priori* hypotheses and using multiple testing to find relationships among variables. In turn, we should adapt our reporting procedures (among other potentially useful measures, such as pre-registration) to acknowledge the exploratory character of the studies and properly qualify the strength of the evidence provided in the experiment.

Finally, we also have observed some weaknesses in the experimenters' statistical knowledge. This problem was pointed out by other researchers, e.g., [17, 28] previously. The ESEM community (and the overall SE community, as well) should establish measures to improve experimenters' statistical skills.

8 ACKNOWLEDGMENTS

We wish to express our appreciation to the experimenters that participated in the survey. This work has been partially supported by the Spanish Ministry of Economy and Competitiveness research grant TIN2014-60490-P, the " Laboratorio Industrial en Ingeniería del Software Empírica (LI2SE)", ESPE research project.

REFERENCES

[1] [n.d.]. http://eseiw2019.com/call-for-papers/
[2] Hyun-Chul Cho and Shuzo Abe. 2013. Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research* 66, 9 (2013), 1261–1266.
[3] J.W. Creswell. 2002. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications. http://books.google.es/books?id=nSVxmN2KWeYC
[4] Jose Luis de la Vara, Markus Borg, Krzysztof Wnuk, and Leon Moonen. 2014. Survey on Safety Evidence Change Impact Analysis in Practice: Detailed Description and Analysis. (2014).
[5] Nicholas J DeVito and Ben Goldacre. 2019. Catalogue of bias: publication bias. *BMJ Evidence-Based Medicine* 24, 2 (2019), 53–54. https://doi.org/10.1136/bmjebm-2018-111107 arXiv:https://ebm.bmj.com/content/24/2/53.full.pdf
[6] Kerry Dwan, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, Carrol Gamble, Davina Ghersi, John P. A. Ioannidis, John Simes, and Paula R. Williamson. 2008. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLOS ONE* 3, 8 (08 2008), 1–31. https://doi.org/10.1371/journal.pone.0003081
[7] Phillipa J Easterbrook, Ramana Gopalan, JA Berlin, and David R Matthews. 1991. Publication bias in clinical research. *The Lancet* 337, 8746 (1991), 867–872.
[8] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.
[9] A. Indrayan and M.P. Holt. 2016. *Concise Encyclopedia of Biostatistics for Medical Professionals*. CRC Press. https://books.google.es/books?id=mBkNDgAAQBAJ
[10] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8 (2005), e124.
[11] Andreas Jedlitschka and Dietmar Pfahl. 2005. Reporting Guidelines for Controlled Experiments in Software Engineering. *4th International Symposium on Empirical Software Engineering (ISESE 2005)* (November 2005), 95–104.
[12] Magne Jørgensen, Tore Dybå, Knut Liestøl, and Dag IK Sjøberg. 2016. Incorrect results in software engineering experiments: How to improve research practices. *Journal of Systems and Software* 116 (2016), 133–145.
[13] Jessica Jung, Kai Hoefig, Dominik Domis, Andreas Jedlitschka, and Martin Hiller. 2013. Experimental comparison of two safety analysis methods and its replication. In *2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. IEEE, 223–232.
[14] Natalia Juristo and Ana M Moreno. 2013. *Basics of software engineering experimentation*. Springer Science & Business Media.
[15] Norbert L Kerr. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2, 3 (1998), 196–217.
[16] Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for performing Systematic Literature Review in Software Engineering*. Technical Report EBSE-2007-01. Keele University and University of Durham.
[17] Barbara Kitchenham, Lech Madeyski, and Pearl Brereton. 2019. Problems with Statistical Practice in Human-Centric Software Engineering Experiments. In *Proceedings of the Evaluation and Assessment on Software Engineering (EASE '19)*. ACM, New York, NY, USA, 134–143. https://doi.org/10.1145/3319008.3319009
[18] Barbara A Kitchenham and Shari L Pfleeger. 2008. Personal opinion surveys. In *Guide to Advanced Empirical Software Engineering*. Springer, 63–92.
[19] Guy R McPherson. 2001. Teaching & learning the scientific method. *The American Biology Teacher* 63, 4 (2001), 242–245.
[20] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1 (2017), 0021.
[21] Regina Nuzzo et al. 2014. Statistical errors. *Nature* 506, 7487 (2014), 150–152.
[22] Teade Punter, Marcus Ciolkowski, Bernd Freimut, and Isabel John. 2003. Conducting on-line surveys in software engineering. In *Empirical Software Engineering, 2003. ISESE 2003. Proceedings. 2003 International Symposium on*. IEEE, 80–88.
[23] Rolando P. Reyes, Oscar Dieste, Efraín R. Fonseca, and Natalia Juristo. 2018. Statistical Errors in Software Engineering Experiments: A Preliminary Literature Review. In *Proceedings of the 40th International Conference on Software Engineering (ICSE '18)*. ACM, New York, NY, USA, 1195–1206. https://doi.org/10.1145/3180155.3180161
[24] J. Sayyad Shirabad and T.J. Menzies. 2005. The PROMISE Repository of Software Engineering Databases. School of Information Technology and Engineering, University of Ottawa, Canada. http://promise.site.uottawa.ca/SERepository
[25] Frank L. Schmidt and John E. Hunter. 2015. *Methods of Meta-Analysis* (3rd ed.). SAGE Publications.
[26] W.R. Shadish, T.D. Cook, and D.T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company. 510 pages.
[27] Fujian Song, Lee Hooper, and Yoon Loke. 2013. Publication bias: what is it? How do we measure it? How do we avoid it? *Open Access Journal of Clinical Trials* 2013, 5 (2013), 71–81.
[28] S. Vegas, C. Apa, and N. Juristo. 2016. Crossover Designs in Software Engineering Experiments: Benefits and Perils. *IEEE Transactions on Software Engineering* 42, 2 (February 2016), 120–135. https://doi.org/10.1109/TSE.2015.2467378
[29] Andrew Vickers. 2010. *What is a P-value anyway?: 34 stories to help you actually understand statistics*. Addison-Wesley Longman.
[30] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*. Springer Science & Business Media.
[31] Robert K Yin. 2013. *Case study research: Design and methods*. Sage publications.