Statistical Errors in Software Engineering Experiments: A Preliminary Literature Review

Rolando Reyes, Oscar Dieste, Efraín R. Fonseca C. and Natalia Juristo

July 4, 2018

Statistical Errors in Software Engineering Experiments: *A Preliminary Literature Review*

Rolando P. Reyes Ch.
ORCID: 0000-0002-7554-5182
Universidad de las Fuerzas Armadas ESPE
Sangolquí, Ecuador
rpreyes1@espe.edu.ec

Oscar Dieste
ORCID: 0000-0002-3060-7853
Universidad Politécnica de Madrid
Campus de Montegancedo
Madrid 28660, Spain
odieste@fi.upm.es

Efraín R. Fonseca C.
ORCID: 0000-0001-7129-9335
Universidad de las Fuerzas Armadas ESPE
Sangolquí, Ecuador
erfonseca@espe.edu.ec

Natalia Juristo
ORCID: 0000-0002-5559-7049
Universidad Politécnica de Madrid
Campus de Montegancedo
Madrid 28660, Spain
natalia@fi.upm.es

ABSTRACT

Background: Statistical concepts and techniques are often applied incorrectly, even in mature disciplines such as medicine or psychology. Surprisingly, there are very few works that study statistical problems in software engineering (SE). ***Aim:*** Assess the existence of statistical errors in SE experiments. ***Method:*** Compile the most common statistical errors in experimental disciplines. Survey experiments published in ICSE to assess whether errors occur in high quality SE publications. ***Results:*** The same errors identified in others disciplines were found in ICSE experiments, exhibiting rather large prevalences, over 30% of the reviewed papers in several types of errors such as: a) Missing statistical hypotheses, b) missing sample size calculation, c) failure to assess statistical tests assumptions, and d) uncorrected multiple testing. When experiments restrict to the validation section of a larger research paper, the prevalence of errors increases. The origin of the errors can be traced back to: a) Researchers' inadequate statistical training, and, b) abundance of exploratory research. ***Conclusions:*** This paper provides preliminary evidence that SE research suffers the same statistical problems than other experimental disciplines. However, SE community does not seem to be aware of the existence of shortcomings in their experiments, whereas other disciplines work hard to avoid them. Further research is necessary to find the underlying causes and set corrective measures, but at the outset some actions could be effective: a) Improve the statistical training of SE researchers, and b) enforce quality assessment and reporting guidelines in SE publications.

CCS CONCEPTS

•General and reference →Surveys and overviews;

KEYWORDS

Literature review, Survey, Prevalence, Statistical errors

1 INTRODUCTION

Experimentation makes extensive use of statistics. Several studies warn about the existence of scientific articles with inappropriate statistical procedures [5, 32, 62]. This happens even in mature disciplines, such as the health sciences [6].

In turn, there are very few works studying statistical errors in software engineering (SE) articles. There are works in SE discussing statistical power [23], heterogeneity in meta-analysis [52], and the relative strengths and weaknesses of cross-over designs [41, 81]. This stands in contrast with the other disciplines where it is relatively easy to find works warning about problems in simple statistical concepts such as the definition of hypotheses [16, 58], interpretation of p-values [62], sample size calculation [2, 25] and significance levels [58], to cite a few.

We aim to assess the prevalence of these problems in the SE literature. We have compiled the most common statistical errors in experimental disciplines and surveyed empirical papers published in ICSE between 2006 and 2015, checking whether these papers make or not the compiled errors. Our results point out to SE experiments have the same weaknesses than in other sciences. SE researchers do not use correctly relatively simple concepts such as: hypotheses posing, sample size estimation, inference, and post-hoc testing. These problems seem to be related to inadequate statistical training, and the conduction of exploratory research.

Our contributions are the confirmation of shortcomings in SE experimental research, and the identification of their origin. In our opinion, SE community should improve researchers' statistical training and, more importantly, establish mechanisms (e.g., quality assessment tools, reporting guidelines) to identify and fix statistical problems in SE experiments before they proceed to publication.

The structure of this paper is as follows: Section 2 provides a background to the topic of statistical errors in the sciences and SE. Section 3 presents a short literature review in which several statistical errors are identified. A subset of those errors is screened in experiment articles in Section 4. The origin of those errors is evaluated in Section 5. A critical appraisal of this review is presented in Section 6. Finally, the conclusions are reported in Section 7.

2 BACKGROUND

2.1 Statistical Errors in Experimental Disciplines

Researchers in the sciences and engineering apply statistical techniques to analyze and interpret many of their research results. Hence, statistical techniques have experienced an increase in use, particularly in medicine [2, 63, 83], psychology [5], education [21], and social sciences [25, 58].

There is a relatively large set of publications that provide information about the existence of statistical problems in virtually all disciplines. Not all publications are recent; they have been available since the widespread adoption of experimental research in their respective areas. The reported problems have a broad scope [51], including: the definition of statistical hypotheses [16, 58], interpretation of p-values [62], sample size calculation [2, 25], significance levels [58], and confidence intervals [16], others.

Papers about statistical shortcomings in other disciplines have derived their results from some type of literature review of primary studies from one or more specialized journals. Their conclusions are surprising and perturbing since they report high error rates:

- Welch [83] studied 145 articles from one of the most renowned medical journals, the *American Journal of Obstetrics and Gynecology*, and found that 52.6% of the articles contained inadequate or incomplete statistical descriptions.
- Bakker [5] evaluated 218 articles from high and low impact psychology journals. The author reported that low impact journals exhibit statistical inconsistencies more frequently than high impact journals. Bakker determined that about 15% of all the papers from both high and low impact journals have at least one incorrect statistical conclusion.
- Ercan et al. [25] evaluated 164 and 145 articules in Psychiatry and Obstetrics, respectively. 40% of the psychiatric, and 19% of the obstetrics publications, contained mistakes regarging: Sampling, sample size calculation, and contradictory interpretations of inferential tests.
- Kilkenny et al. [39] assessed the experimental design of 271 papers published in *Medline* and *EMBASE* between 2003 - 2005. More than 60% of the paper exhibit biases during the assembly of the study cohort, weak statistical analysis, missing information, etc.

The origin of the statistical errors can be traced back to several causes:

- According to Castro et al. [75], the analysis and interpretation of empirical results in any scientific discipline depend primarily on how well researchers understand inferential statistics. The authors suggested that researchers in the education community, especially PhD students, are prone to misconceptions, particularly when they are using abstract statistical concepts, such as confidence intervals, sampling distributions with small numbers, sampling variability, different types of distributions, and hypotheses tests.
- Cohen et al. [19] conducted an empirical study with degree students. They found that students lack statistical knowledge, which leads to misinterpret statistical concepts, and bias judgements.
- Brewer [12] evaluated 18 statistical handbooks from renowned publishers, e.g., *Academic Press, Addison-Wesley, McGraw-Hill, Prentice-Hall, John Wiley*, etc. These books contained imprecise statements in topics such as sampling distributions, hypothesis testing, and confidence levels.

2.2 Statistical Errors in SE

The SE community apparently has limited awareness of the existence and impact of statistical shortcomings in its publications. When we searched for SE papers related to statistical problems, we only found the following results: Dybå et al.'s paper regarding statistical power [23], Miller's paper on meta-analysis [52], and two papers by Kitchenham [41] and Vegas et al. [81] that focused on within-subject designs.

Several other papers discuss specific statistical issues. For instance, Kitchenham's paper introduced robust statistical methods [42], while Arcuri and Briand's paper discussed statistical tests for the assessment of randomized algorithms [4]. These works do not assess the weaknesses in current research. They suggest opportunities for improvement in the toolset that SE researchers currently use.

The difference between SE and other experimental disciplines regarding statistical errors is manifest. In medicine and other sciences, statistical problems are routinely identified in publications; this aspect is almost completely overlooked in SE.

The assessment of statistical defects and methodological problems have been addressed in a relatively late period in other disciplines. For instance, while the first formal randomized clinical trial in medicine was conducted in the 1940s [8], the first publication about statistical defects in medicine that we are aware of was published in the 1970s [30]. Given that SE is still in the adoption phase regarding experimental methods and the associated statistical techniques, the little attention paid to the assessment of statistical issues should come as no surprise.

This paper reports an exploratory study aimed to answer the following research questions:

RQ1: What are the most common problems associated with the use of experimental procedures in experimental disciplines?

RQ2: What is the prevalence of statistical errors in SE research?

3 STATISTICAL ERRORS IN EXPERIMENTAL DISCIPLINES

3.1 Review Strategy

To answer RQ1, we reviewed several specialized books published on the topic, such as Good et al. [29], Vickers [82], and Huck [34]. These books provide a good starting point for our exploratory study because they are not related to any specific discipline (although there is some bias toward the health sciences) and they focus on serious errors often inspired in real research.

3.2 Collected Data

Two researchers (R. P. Reyes and O. Dieste) reviewed the three aforementioned books. They found that 93 text sections clearly pointing out to some type of error that can be frequently found in the literature. Discrepancies were solved by consensus. The complete listing of paragraphs is available at https://goo.gl/8zb9LU, including links to the reference books and related literature.

3.3 Analysis Method

We applied thematic synthesis to classify the statistical errors. We applied the guidelines by Creswell [21] and Cruzes et al. [22] to avoid biases and achieve methodological rigor in the synthesis and interpretation of results [11]. The analysis consisted of two stages: coding and theme definition. It was conducted by the same two researchers that collected the data.

During the coding stage, both researchers independently assigned low-level codes to each text section, which were later reviewed and harmonized. We created 93 different codes. During the theme definition stage, codes were grouped together by means of higher-level codes. This procedure aligned with our purposes since the high-level themes represent error-prone areas. Both researchers worked collaboratively. They organized the low-level concepts into high-level themes following a directed graph, shown in Fig. 1. Themes and connections between themes and concepts are available at https://goo.gl/8zb9LU.

Nodes represent categories of statistical errors. Categories become progressively more abstract when we traverse the tree from right to left. For instance, the node *study design* (bottom side of Fig. 1) is connected to the nodes *assignment* and *sampling*. This means that the high-level category *study design* contains two types of errors: *assignment* and *sampling* errors. Likewise, *assignment* splits into further lower-level error types, such as *matching*, *randomization*, etc. Notice that Fig. 1 show only a subset of the error types that we have identified to keep the graph within page limits. The graph is available at https://goo.gl/qovXQw.

Codes and high-level themes, as well as the high-level themes themselves, may be connected multiples times because they are mentioned in several books, or multiple times in the same book in different contexts. For instance, *randomization* is discussed twice in terms of the representativeness of the random samples:

> (item #40) *Misconception: If a truly random process is used to select a sample from a population, the resulting sample will turn out to be just like the population, but smaller.* [34, pp. 123]

> (item #41) *Misconception: A sample of individuals drawn from a larger, finite group of people deserves to be called a random sample so long as (1) everyone in the larger group has an equal chance of receiving an invitation to participate in the study and (2) random replacements are found for any of the initial invitees who decline to be involved.* [34, pp. 127]

and once more with regard to the equivalence of experimental groups obtained by random assignment:

> (item #90) *The idea behind randomization is to make the groups as similar as possible [...]. Baseline differences at the beginning of the trial, such as in age o gender, are due to chance. [...] giving a p-value for baseline difference between groups created by randomization is testing a null hypothesis that we know to be true.* [82, pp. 100]

These repeated associations are an indication of relevance, and thus the arcs connecting the corresponding nodes have been made proportionally wider. The number next to the arc indicates the number of times the connection appears in the raw data. Dotted lines represent connections that appear just once.

3.4 Review Results

Statistical errors can be classified in three groups: a) Experimentation, b) meta-analysis, and c) prediction. Most errors are related to *experimentation*. Nevertheless, it is noticeable that *meta-analysis* appears three times in connection with subgroup analysis and the combination of studies with different designs. Prediction appears just once, in relation with linear modeling. In what follows, we will focus in problems associated exclusively with experiments.

Analysis is the experimentation facet most often mentioned in connection with statistical errors. In the three reviewed books, analysis errors appear 63 times. There are two main sources of problems with analysis: the application of *inferential* techniques and the *interpretation* of results:

- The inferential techniques most often used during experimental data analysis are *classical tests*, such as t-tests, and their related concepts, such as p-values and tails. Researchers often make wrong *assumptions* about the tests (e.g., *robustness* of t-test), and they *select* tests in circumstances in which they cannot be applied (e.g., *ordered alternative hypotheses*) or are sub-optimal (e.g., low *powered* tests). All common tests, including t-tests, correlations, and ANOVA, are mentioned in this context.
- Another frequently mentioned inferential technique is *linear modeling*; multiple linear regression is the best known example of linear modeling. The most frequently mentioned problem is the rationale behind the *definition* of the linear model. Other issues, such as the violation of *assumptions* and *usage beyond limits* (e.g., outside the linear phase) are also reported.
- Many supposedly basic concepts, such as *confidence intervals*, *statistical significance*, or *p-values* are frequently misinterpreted.

Study design is second to analysis. Under this theme, we have included methodological issues connected to the management of experimental units, such as *sampling* and *assignment*. In both cases, the origin of the problems comes from inappropriate or missing *randomization* and *sample size calculation*.

Reporting is another problematic aspect, which is mentioned the same number of times (ten) than *study design*. The origin of reporting defects is manifold (e.g., overlooking experimental *incidents* or *multiple testing*), although the absence of *descriptive statistics* (e.g., *means*) is emphasized (tree times) in the reviewed books.

The last prominent theme is *goal definition*. Researchers frequently do not pose *statistical hypotheses*. Failure to explicitly define *null* hypotheses appears three different times in Fig. 1.

4 STATISTICAL ERRORS IN ICSE EXPERIMENTS

The aim of RQ2 is to find out the prevalence of statistical errors and methodological problems in SE research. To answer RQ2, we evaluated the experiments published in 10 editions (2006-2015) of ICSE. ICSE is the flagship conference on SE. We expect that our evaluation yields the prevalence of common statistical errors in the best SE research; lower quality SE research is probably experiencing a worse situation.

4.1 Evaluation Instrument

The complete list of statistical errors that we have compiled contains almost 100 items. Since statistical errors are ubiquitous in the general research literature, it is highly likely that several of those ~ 100 problem types would appear in virtually any SE paper as well. Therefore, an exhaustive review of SE experiments would draw a too pessimistic picture of our field.

We have focused on recurrent types of errors (the ones pointed by wide arrows in Fig. 1. For instance, "null hypothesis" -related problems are referenced multiple times in Fig. 1, as well as test "assumptions" or "central measures". We have selected the most error-prone statistical concepts, developed appropriate questions, and created the 10-question checklist shown in Table 1. All these questions can be easily traced back to Fig. 1 (or the online version at https://goo.gl/qovXQw). Some clarifications are appropriate at this point:

- Q1.1 and Q1.2 may look outdated due to the increasing criticisms to the "Null-Hypothesis and Significance Testing" (NHST), and the recommendations to adopt other statistical approaches such as confidence intervals and effect size indices [10, 71, 77]. However, SE research still falls behind those recommendations. For instance, only 4 out of 21 experiments published in ICSE between 2006-2015 report some measure of effect size, and 2 out of 21 confidence intervals. Nowadays, NHST is still the main statistical approach used in SE.
- Q4 (*Have subjects been randomly assigned to treatments?*) may not be applicable to some types of experiments, e.g., when two defect prediction algorithms are applied to the same code, that is, matched pairs or similar designs. In cases like this, the question is answered as "N/A". Similar action is taken when any question does not make sense for a given experiment, e.g., Q5 (*Have the test assumptions (i.e., normality and heteroskedasticity) been checked or, at least, discussed?*) when an experiment does not use statistical tests.
- Test assumptions vary from test to test. In many cases, reference books present incomplete or even questionable assumptions. Thus, in Q5 (*Have the test assumptions (i.e., normality and heteroskedasticity) been checked or, at least, discussed?*), we will pay attention only to the most usual conditions (normality and heteroskedasticity) that have to be examined before applying virtually any parametric test.
- Q7 (*Have the analysis results been interpreted by making reference to relevant statistical concepts, such as p-values, confidence intervals, and power?*) looks a rather relevant question. Fig. 1 shows that the node "interpretation" is connected by wide arcs with nodes representing relatively simple statistical concepts, such as "power", "confidence interval" and "p-value", among others. However, we doubt that we can answer this question objectively. While authors typically discuss their results at length, during the discussion they may simplify or omit some statistical issues to clearly transmit their message to readers. Thus, we face the risk of making mistakes, e.g., evaluate Q7 negatively due to incomplete reporting. We decided to skip this question (so it is crossed out in Table 1).
- Multiple testing does not appear to be a relevant issue in Fig. 1. However, it was cited three times as a source of problems during both analysis and reporting; note that there are three incoming arcs to this node in Fig. 1. This justifies Q9 (*Is multiple testing reported and accounted for, e.g., Bonferroni?*).
- Q10 (*Are descriptive statistics, such as means and counts, reported?*) is relevant for both analysis and reporting. We will consider it in the context of reporting only, to avoid inflating the number of defects found.

Table 1: Evaluation checklist

#	Question
Q1.1	Are hypotheses null explicitly defined?
Q1.2	Are hypotheses alternate explicitly defined?
Q2	Has the required sample size been calculated?
Q3	Have subjects been randomly selected?
Q4	Have subjects been randomly assigned to treatments?
Q5	Have the test assumptions (i.e., normality and heteroskedasticity) been checked or, at least, discussed?
Q6	In cases where linear models are applied, has the model definition been discussed?
~~Q7~~	~~Have the analysis results been interpreted by making reference to relevant statistical concepts, such as p-values, confidence intervals, and power?~~
Q8	Do researchers avoid calculating and discussing power post hoc?
Q9	Is multiple testing reported and accounted for, e.g., Bonferroni?
Q10	Are descriptive statistics, such as means and counts, reported?

4.2 Target studies

At the outset, our intention was to survey only experimental papers in the 2006-2015 ICSE editions. However, the decision proved soon to be questionable. We conducted a pilot study using the 2012 ICSE edition to check the feasibility of our study. We immediately realized that the number of full-fledged experiments was quite low; we found only four experiments. In turn, we found many small-scale experiments aimed at evaluating the properties of new techniques or methods, typically reported in "Evaluation" Sections within the same research paper. More concretely, we identified 16 "experiments as evaluations" (18.4% of the total number of papers).

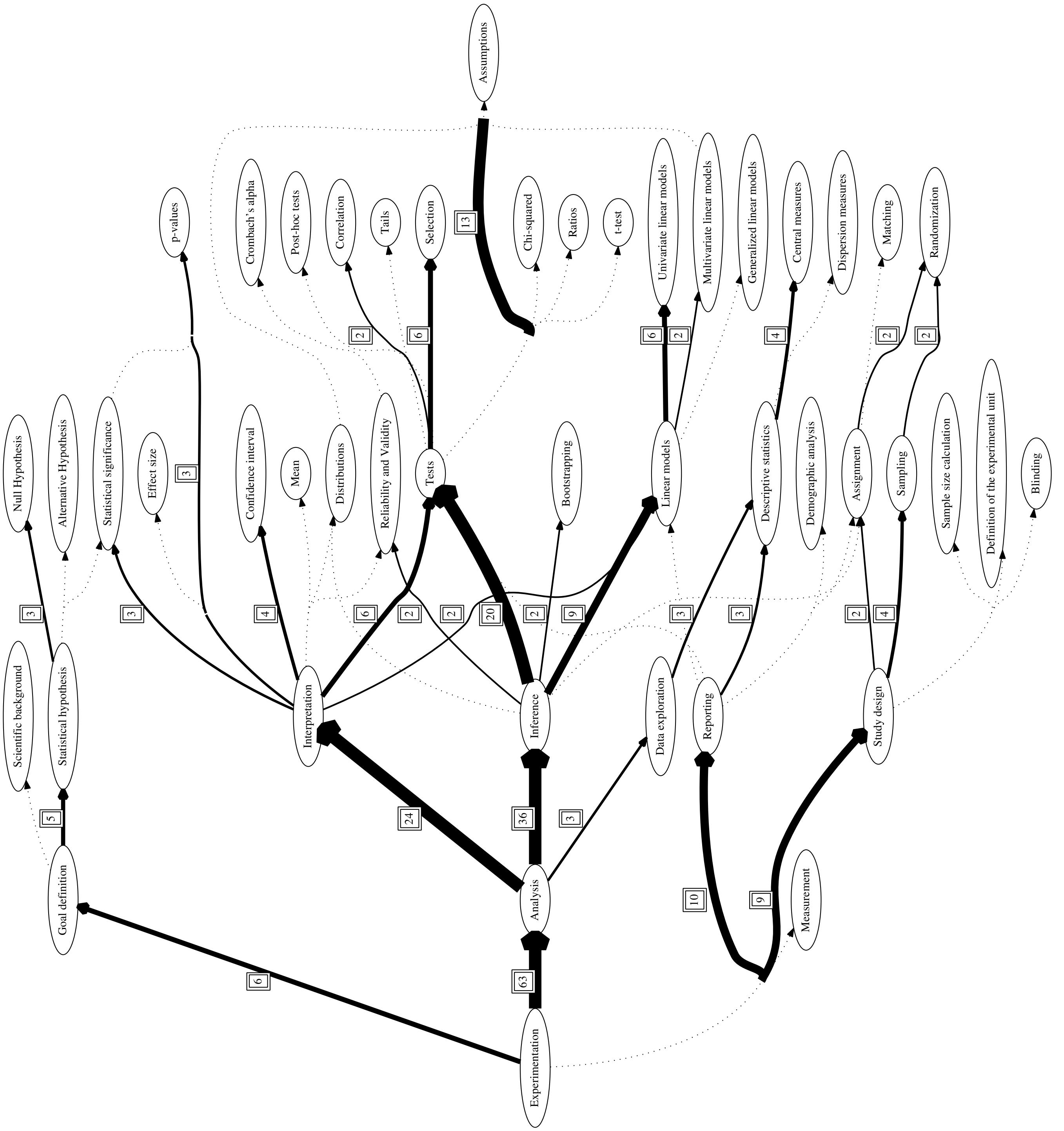


Figure 1: Classification of statistical errors in experimental research papers

The question was wether the survey should be extended to "experiments as evaluations", or restricted to "standalone experiments". "Experiments as evaluations" often apply an experimental methodology, but they have short length, typically 1-3 pages long. The compressed reporting format may lead to writing practices that may be wrongly understood as statistical errors by reviewers. On the other hand, "experiments as evaluations" represent a large share of empirical research; overlooking them implies that this survey's results would be just partial.

We decided to separately evaluate both type of studies. In a first stage, we searched for all "standalone experiments" published in ICSE between 2006-2015. We found 21 papers in total. In a second stage, we collected a similar number[1] of "experiments as evaluations" to avoid over-representation.

4.3 Study selection

Two researchers (O. Dieste and R. P. Reyes) worked separately to screen the tables of contents of the ICSE Technical Track for the 2006-2015 editions. They reviewed the title and abstracts to search for indications that an experiment was reported. In case of doubt, they examined the full text, seeking further evidence of the existence of at least two treatments and the execution of some comparison between treatments (in other words: the minimum conditions that any experiment shall satisfy).

The total number of papers and the papers pre-selected after screening are shown in Table 2. The pre-selection agreement was calculated using Fleiss' κ, as recommended by K. L. Gwet [31, pp. 52]. $\kappa = 0.45$, typically considered as *moderate* [27]. This implies that we may have failed to identify some experiments. Notice that identifying experiments using metadada, such as titles and abstracts, is not straightforward due to missing methodological descriptors.

Three researchers (O. Dieste, E. R. Fonseca, and R. P. Reyes) individually reviewed the pre-selected papers and classified them into the *experiment* and *non-experiment* categories. Disagreement was solved by consensus. 21 papers were classified as "standalone experiments", which represents 2.7% of the total papers published in ICSE. The agreement level for this step of the selection process was Fleiss' $\kappa = 0.52$, typically considered as *moderate* [27, 31]. As it will be reported below, this low agreement is due to the existence of missing information (e.g., hypotheses or randomization procedures) in the manuscripts. Further details are available at https://goo.gl/jHWpq3.

Finally, R. P. Reyes randomly selected three "experiments as evaluations" per ICSE edition from the tables of contents of the ICSE Technical Track. The three researchers independently reviewed these papers, and discrepancies were solved by consensus. The process was repeated until three "experiments as evaluations" were identified for each ICSE 2006-2015 edition.

4.4 Execution

The three researchers individually evaluated all papers and gave a *yes/no/not applicable* answer to each checklist question (see Table 7). The level of agreement was *substantial* to *almost perfect* in many cases, which increases the confidence of our results. Details of the evaluation are available at https://goo.gl/3iy9eL ("standalone experiments") and https://goo.gl/qCboSX ("experiments as evaluations").

[1] We rounded up from 21 to 30, i.e., 3 papers per edition × 10 ICSE editions = 30 papers.

Table 2: Summary of the selection process. "Experiments as evaluations" between parentheses

Year	Total papers (TP)	After screening	Selected	%
2006	72	8	2 (3)	2.8% (4.1%)
2007	64	7	2 (3)	3.1% (4.7%)
2008	85	8	1 (3)	1.2% (3.5%)
2009	70	7	0 (3)	0.0% (4.3%)
2010	62	5	1 (3)	1.6% (4.9%)
2011	62	5	1 (3)	1.6% (4.9%)
2012	87	31	4 (3)	4.6% (3.5%)
2013	85	8	1 (3)	1.2% (3.5%)
2014	99	11	5 (3)	5.0% (3.1%)
2015	84	11	4 (3)	4.8% (3.5%)
Total	770	101	21 (30)	2.7% (3.9%)

Table 3: Agreement levels per question

		Standalone Exp.		Exp. as Eval. Sec.	
Stage		κ	**Agree**	κ	**Agree**
Goal definition	Q1.1	0.839	Almost perfect	0.643	Substantial
	Q1.2	0.746	Substantial	0.788	Substantial
Study design	Q2	1,000	Perfect	1.000	Perfect
	Q3	0.092	Slight	0.389	Fair
	Q4	0.541	Moderate	0.585	Moderate
Analysis	Q5	0.752	Substantial	0.662	Substantial
	Q6	1.000	Perfect	0.558	Moderate
	Q8	0.894	Almost perfect	0.803	Almost perfect
Reporting	Q9	0.592	Moderate	0.659	Substantial
	Q10	1.000	Perfect	0.480	Moderate

4.5 Survey Results

Table 4 sumarizes the survey results. Percentages are calculated as $\frac{\{Yes|No|N/A\}}{9}$. The column "No" represents the percentage of papers in the sample that are affected by the error indicated by the corresponding question, i.e., the prevalence of the statistical error. Q1 was split into two parts to differentiate the problems related to the null (Q1.1) and the alternate (Q1.2) hypotheses.

We found clear evidence of the existence of statistical errors in ICSE papers. The prevalence of the different errors vary, but it is substantial in many cases, e.g., Q1, Q2, Q3 and Q5 (hypothesis definition, sample size calculation, random selection and assumption checking, respectively). The results are somewhat different for standalone experiments and "experiments as evaluations". In the later case, the number of "N/A" responses is much higher. Apparently, the reasons are twofold:

(1) Most of the "experiments as evaluations" apply a matched pairs design. Random assignment (Q4) is typically not applicable in this case, e.g., two different bug prediction algorithms are applied to the same code [78, 85].
(2) A large number of studies, e.g., [46, 90] conduct the analysis using descriptive statistics only. Descriptive statistics do not have assumptions to check (Q5). When inferential

Table 4: Percentages of defects

Stage		Standalone Experiments Yes	No	N/A	Experiments as *Evaluation Sections* Yes	No	N/A
Goal definition	Q1.1	66.7%	33.3%	0.0%	13.3%	83.3%	3.3%
	Q1.2	57.1%	42.9%	0.0%	6.7%	90.0%	3.3%
Study design	Q2	0.0%	100.0%	0.0%	3.3%	96.7%	0.0%
	Q3	28.6%	71.4%	0.0%	13.3%	86.7%	0.0%
	Q4	66.7%	28.6%	4.76%	20.0%	0.0%	80.0%
Analysis	Q5	61.9%	33.3%	4.76%	13.3%	20.0%	66.7%
	Q6	4.8%	0.0%	95.24%	3.3%	0.0%	96.7%
	Q8	85.7%	9.5%	4.76%	36.7%	0.0%	63.3%
Reporting	Q9	9.5%	71.4%	19.07%	3.3%	26.7%	70.0%
	Q10	95.2%	0.0%	4.76%	76.7%	13.3%	10.0%

statistics are not used, Q6-9 (linear modelling, power, and post-hoc testing) are not applicable either.

We ran a classification CHAID tree[2] to confirm the above observations. A value "N/A" in Q4 generates a subset containing 80% of all the "experiments as evaluations" studies ($\chi^2 = 29.7, df = 2, p{-}value < 0.001$). The classification tree confirms that the lack of random assignment due to matching is a differential characteristic of the "experiments as evaluations".

Focusing on the prevalence of errors, we found that both standalone experiments and "experiments as evaluations" display similar values[3] when examined using *Question* × *Study type* contingency tables, with the exception of Q1.1 ($\chi^2 = 15.4, df = 1, p - value < 0.001$) and Q1.2 ($\chi^2 = 20.7, df = 1, p - value < 0.001$). In both cases, standalone experiments define null hypotheses (Q1.1) five times more frequently than "experiments as evaluations" (66.7% vs. 13.3%), and alternate hypotheses (Q1.2) eight times (57.1% vs. 6.7%).

Both types of studies do not show statistically significant differences for the remaining questions, although some may be false negatives. The large number of "N/A" in several questions decrease the amount of usable data, thus lowering the power of the tests. However, the low p-values in both the χ^2 and the Fisher's Exact Test suggest that Q3, Q4, Q5, Q10 could achieve statistical significance with larger samples. In all cases, standalone experiments perform random selection (Q3[4], random assignment (Q4), assumption checking (Q5) and reporting of descriptive statistics (Q10) more frequently than "experiments as evaluations". Differences are not so large as in the case of Q1.1 and Q1.2, but still substantial, e.g., 61.9% vs. 13.3% for Q5.

We can also read from Table 4 that:

- The required sample size (Q2) has been calculated in just one study. The definition of the linear model (Q6) has been considered in just two cases.
- Multiple testing (Q9) is a pervasive problem in SE research. Most studies fail to report or correct for multiple testing using adequate, e.g., Bonferroni, methods.
- Random selection (Q3) exhibits high prevalences. Nevertheless, this problem is not easy to solve in human experiments due to the trouble of assembling cohorts. In turn, random selection could be effectively applied in non-human research, e.g., when data is extracted from code repositories.

This survey shows that common statistical errors that are occurring in other sciences happen in SE as well. We have been able to survey a very limited amount of experimental papers in one SE conference. However, both the type and number of problems found suggest that SE is facing the same challenges than other sciences.

5 DISCUSSION

The most likely explanation for the occurrence of the statistical errors associated to Q1-10 is the recent adoption of experimental methods in SE. Many researchers have not taken formal courses on experimental methodology and inferential statistics as part of their Master/PhD training. Self-education tends to bring about wide differences among individuals. In this scenario, two situations could be expected:

(1) The studies conducted by skilled researchers will be higher quality (understanding "quality" as the absence of errors, e.g., $\frac{\text{"Yes" answers}}{\text{All answers}}$) than those conducted by less skilled researchers. We could thus expect that the "quality" values spread from 0% to 100%. As errors are independent, the distribution of "quality" will follow a normal distribution[5].

(2) Those statistical concepts closely related to practice, e.g., random assignment (Q4), assumption checking (Q5), and reporting (Q10) shall exhibit lower error probability than "theoretical" ones, e.g., hypotheses definition (Q1), sample size calculation (Q2), random selection (Q3), linear modeling (Q6), post-hoc power calculation (Q8), and post-hoc testing (Q9).

In order to check the situation 1 above, Fig. 2 shows the histograms for both types of studies. In the case of standalone experiments (Fig. 2a), the histogram matches the assumption: the "quality" scores fill the 0% - 100% interval and the distribution is normal ($Shapiro{-}Wilk = .947, df = 21, p{-}value = .300$). "Experiments as evaluations" (Fig. 2b) display a rather different picture. The distribution is skewed to the left ($skewness = 1.02$), indicating that papers' "quality" gather in the low scores. The distribution is clearly non-normal ($Shapiro - Wilk = .863, df = 30, p - value = .001$).

The previous analysis suggests a different origin, depending on the type of study, for the statistical errors. In the case of standalone experiments, inadequate statistical training may explain the the observed errors.

In the case of "experiments as evaluations", training alone cannot explain the data. In our opinion, the low scores point to the secondary role of statistics and experimental methodology in these

[2] The response variable was the study type ("standalone experiments" and "experiments as evaluations") and the predictors the questions Q1-10. We used the default CHAID parameters in SPSS, with the exception of the parent and child nodes, that were set to 10 and 5 cases respectively due to the small number of cases.

[3] Notice that the "N/A" values may suggest misleading relations. For instance, Q9 gives *Yes/No* values 9.5% and 71.4% for standalone experiments, and 3.3% and 26.7% for "experiments as evaluations". Values differ greatly, but the odds $\frac{71.4}{9.5} = 7.5 \sim \frac{26.7}{3.3} = 8.1$ are rather similar.

[4] Notice that Q3 yields $\kappa = 0.09$ and $\kappa = 0.39$ for standalone experiments and "experiments as evaluations", respectively. Random sampling is a controversial issue in SE. Q3 results should be taken with caution.

[5] Statistical errors are probably dependent. When a researcher learns a statistical topic, e.g., sample size calculation, such knowledge may probably lead to avoid other errors, e.g., post-hoc power calculation. However, the errors underlying Q1-10 are rather diverse as to appear strongly clustered in papers.

works. Not only "experiments as evaluations" take a relatively short space in manuscripts (giving a justification for summarizing "unnecessary stuff"), but also statistical rigor is probably second to the authors' objectives (they are probably more interested in providing a convincing case for their proposals).

To check the situation 2 above, Table 5 contains the odds of making an error. The odds are the same concept introduced in footnote 3; they represent the probability that an event happen (answering negatively the Q_i question, i.e., an statistical error is present in a paper) rather than another (answering Q_i positively, i.e., there is not such an error).

Table 5: Odds of making the error indicated in Q1-10

		Odds ($No \div Yes$)	
Concept type	Q	Standalone experiments	Experiments as evaluations
Practical concepts	Q4	0.4	0.0
	Q5	0.5	1.4
	Q10	0.0	0.2
Theoretical concepts	Q1.1	0.5	5.0
	Q1.2	0.8	10.0
	Q2	+Inf	+Inf
	Q3	2.5	5.0
	~~Q6~~	~~0.0~~	~~0.0~~
	~~Q8~~	~~0.1~~	~~0.0~~
	Q9	10.0	10.0

In the case of "experiments as evaluations", the data matches our assumption[6] *exactly*. All "theoretical" concepts have large odds ratios ($\geq$ 5.0), whereas "practical" ones have small ones ($\leq$ 1.4). For standalone experiments, the situation is *almost* the same. For the "theoretical" concepts, odds ratios are smaller than in the case of "experiments as evaluations", with the only exception of Q9. This is coherent with the higher error rate of the "experiments as evaluations" studies. However, Q1.1 and Q1.2 odds ratios are much smaller (0.5 and 0.8, respectively) and comparable to the odds ratios that appear in the group of "practical" concepts.

The previous analysis confirms that inadequate training is the most likely explanation for the presence of statistical errors in experiments. In the case of "experiments as evaluations" studies, a more casual usage of statistics increases the error rate, but the final outcome is the same.

One anomaly in Table 5 is the large odds ratio that Q9 exhibits for standalone experiments. It has the same value than in "experiments as evaluations". Such value is even less plausible given the small odds ratios for Q1.1 and Q1.2: any researcher with a good knowledge of statistical hypotheses should be aware of the impact of multiple testing on α levels. The most likely reason is that, in addition to testing the statistical hypotheses, standalone experiments also perform exploratory research (manifested in a large number of uncorrected post-hoc tests). Exploratory research can be easily found in many SE experiments, e.g., [7, 86].

Post-hoc testing is associated to *p-hacking*, that is, the acceptance of outcomes that fit expectations [55]. p-hacking leads to publication bias. Jørgensen et al. [38] evaluated the existence of publication bias in SE publications following Ioannidis' critical perspective for medicine [35]. Both papers came to a similar conclusion: the likelihood of publication bias is rather high. More importantly for our purposes, both papers report that the underlying reasons for publication bias are statistical: many inference tests and predilection for statistically significant results, among others. Our data supports Jørgensen et al.'s observations: post-hoc testing increases the number of tests, and the lack of correction for multiple testing probably inflates the number of false positives, thus leading to publication bias.

6 THREATS TO VALIDITY

This study applied two research protocols: a literature review and a paper survey. Both protocols have a great degree of similarity. They shall meet some criteria regarding the relevance of the primary studies to answer the research questions, as well as the agreement among studies. Table 6 shows an evaluation[7] according to Thompson et al. [80]. Evaluation was positive overall. We can say with relative confidence that the literature review and the results from the survey are trustable. However, they are not complete due to the limited number of the primary sources used; three well-known books about statistical errors and experimental papers from one SE conference were used in the study. The external validity of this research is thus limited. Additionally, the literature review followed a simplified, but well-defined protocol. We preserved the page numbers of the books from which we extracted information about statistical errors. We disclosed the entire thematic analysis, including codes and high-level themes. All decisions have been made by at least two researchers. These precautions increased the validity of the literature review.

With regards to the paper survey, we have taken reasonable precautions to avoid biases. Three researchers participated in the paper selection and evaluation. All decisions have been recorded and made public for review. Agreement levels (using Fleiss' κ) have been calculated and disclosed.

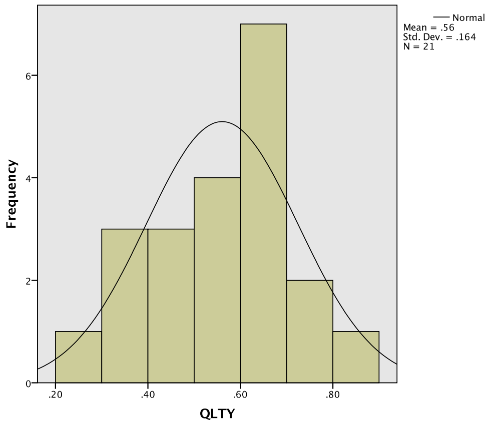
However, the precautions taken did not mean that we performed a correct assessment in all cases. The selection process produced low Fleiss κ, which suggests that we may have skipped some experiments and, thus, potentially biased the results. Even if this is the case, it looks clear that statistical errors are present in SE; only the prevalences, or percentages of appearance, could be affected.

We do not claim that the reported prevalences are representative of all types of SE research. Actually, the prevalences reported in this paper probably represent the best practice in SE research, with the possible exceptions of the ESEM and EASE conferences, and maybe some journals, such as Empirical Software Engineering. As we move away from outlets of repute, the number and severity of statistical errors likely increases.

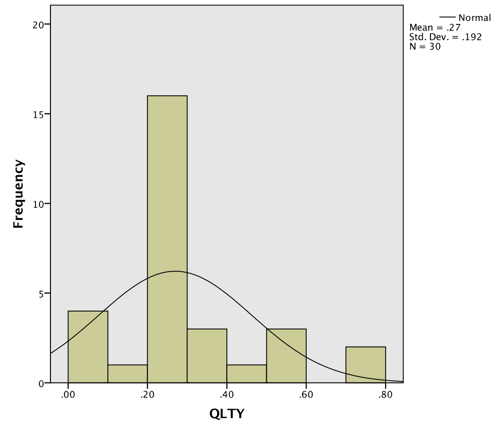
Finally, we wish to point out that the results obtained in the Discussion Section are somewhat speculative. We cannot rule out that there are alternative explanations for the distribution of quality scores and the odds ratios. As usual, further research will be necessary to confirm our deductions.

[6] We are crossing out Q6 and Q8 because: a) Q6 was applicable only in 2 out of 51 studies, and b) post-hoc power analysis (Q8) is a commission, not omission, error; authors may perform correctly simply by not conducting a power analysis. Their inclusion would not have challenged our conclusions.

[7] There are many appraisal procedures; we have chosen [80] because it is rather simple and domain-independent.



(a) Standalone experiments



(b) "Experiments as evaluations"

Figure 2: Histograms

Table 6: Appraisal criteria used for review

Appraisal criteria	RQ1 assessment	RQ2 assessment
Comprehensive literature search?	No	No
Appropriate criteria used to select articles for inclusion?	Partially: The selected books were appropriate, but they have a general orientation. The books may have skipped the discussion of specific statistical errors that probably appear in other sources, such as research papers.	Partially: ICSE is the flagship conference in SE. Other conferences may publish lower quality experiments.
Included studies that were sufficiently valid for the type of question asked?	Yes: The three books specifically addressed the topic of statistical errors.	Yes: Experiments published in ICSE represent the best practice in ESE.
Were the results similar from study to study?	Yes: There was a great degree of coincidence. Several errors were identified by two or three books simultaneously. The same high-level themes were synthetized from the different books.	Yes: Statistical problems repeated across experiments.

7 CONCLUSIONS

The results of this preliminary review point out to the existence in SE of the same type of statistical errors that are found in other scientific disciplines. These problems are not of a complicated or sophisticated nature. They are surprisingly simple and include undefined hypotheses, missing sample size calculations, randomization, and multiple testing, among others. The lack of information about the existence of such problems in SE is rather surprising. The SE methodological literature has not widely addressed this topic; only some works [23, 41, 52, 81] have scratched the surface. Researchers may not be aware of the existence of statistical errors, much less their prevalence and potential impact.

There are two reasons that seem to explain the presence of statistical errors in SE research: a) The recent widespread adoption of experimentation in SE, and b) the frequent conduction of exploratory research. In our opinion, the sudden exigence to apply experimental methods in SE research has led researchers to self-training in statistics. Additionally, it is rather unlikely that SE research teams contain or may count on statistical consultants. In this situation, it is relatively easy that errors slip into designs and ultimately published papers. This situation matches other sciences that have a long experimental tradition, such as medicine and ecology, which only recently have paid attention to statistical errors.

As empirical research in SE approaches a mature stage, there will be a greater awareness about statistical errors and the need to avoid them. However, it does not look wise that the SE community passively wait to reach such state. Besides the establishment of formal training courses in universities and professional societies (something that is happening nowadays), the SE community shall enforce good practices, such as reporting guidelines and quality standards, that have demonstrated useful in other sciences, e.g., medicine [72], psychology [74] and education [28], among others. Furthermore, these good practices can be easily enforced by journal Editors and conference PC Chairs with relatively little cost and effort.

Exploratory research is another source of problems. From the viewpoint of this research, exploratory research manifests as the absence of statistical hypotheses, and the conduction of multiple uncorrected tests. However, these errors lead to publication bias, already detected in SE [38]. Experiment pre-registration is probably the best way to fight against publication bias [15], but it is not easy to set up and enforce. To the best of our knowledge, pre-registration has not been discussed so far in SE. Further research is needed to find out effective ways to battle publication bias in SE. In the meantime, the establishment of reporting guidelines and quality standards may improve the situation.

8 ACKNOWLEDGMENTS

This work was 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] Saba Alimadadi, Sheldon Sequeira, Ali Mesbah, and Karthik Pattabiraman. 2014. Understanding JavaScript event-based interactions. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 367–377.

[2] Douglas G Altman. 1998. Statistical reviewing for medical journals. *Statistics in medicine* 17, 23 (1998), 2661–2674.

[3] Paul V Anderson, Sarah Heckman, Mladen Vouk, David Wright, Michael Carter, Janet E Burge, and Gerald C Gannod. 2015. CS/SE instructors can improve student writing without reducing class time devoted to technical content: experimental results. In *Proceedings of the 37th International Conference on Software Engineering-Volume 2*. IEEE Press, 455–464.

[4] Andrea Arcuri and Lionel Briand. 2014. A Hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering. *Software Testing, Verification and Reliability* 24, 3 (2014), 219–250. DOI:http://dx.doi.org/10.1002/stvr.1486

[5] Marjan Bakker and Jelte M Wicherts. 2011. The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods* 43, 3 (2011), 666–678.

[6] Kirk R Baumgardner. 1997. A review of key research design and statistical analysis issues. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology* 84, 5 (1997), 550–556.

[7] Gabriele Bavota, Bogdan Dit, Rocco Oliveto, Massimiliano Di Penta, Denys Poshyvanyk, and Andrea De Lucia. 2013. An empirical study on the developers' perception of software coupling. In *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 692–701.

[8] A Bhatt. 2010. Evolution of Clinical Research: A History Before and Beyond James Lind. *Perspectives in Clinical Research* 1, 1 (March 2010), 6–10.

[9] Christian Bird, Nachiappan Nagappan, Premkumar Devanbu, Harald Gall, and Brendan Murphy. 2009. Does distributed development affect software quality?: an empirical case study of windows vista. *Commun. ACM* 52, 8 (2009), 85–93.

[10] Marc Branch. 2014. Malignant side effects of null-hypothesis significance testing. *Theory & Psychology* 24, 2 (2014), 256–277.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] James K Brewer. 1985. Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational and Behavioral Statistics* 10, 3 (1985), 252–268.

[13] Yan Cai and WK Chan. 2012. MagicFuzzer: scalable deadlock detection for large-scale applications. In *Proceedings of the 34th International Conference on Software Engineering*. IEEE Press, 606–616.

[14] Mariano Ceccato, Alessandro Marchetto, Leonardo Mariani, Cu D Nguyen, and Paolo Tonella. 2012. An empirical study about the effectiveness of debugging when random test cases are used. In *Proceedings of the 34th International Conference on Software Engineering*. IEEE Press, 452–462.

[15] Chris Chambers, Marcus Munafo, and more than 80 signatories. 2013. Trust in science would be improved by study pre-registration. *The Guardian*, 5 June 2013. Available: https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration [Last accessed: 16 August 2017]. (2013).

[16] 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.

[17] Ilinca Ciupa, Andreas Leitner, Manuel Oriol, and Bertrand Meyer. 2008. ARTOO: adaptive random testing for object-oriented software. In *Proceedings of the 30th international conference on Software engineering*. ACM, 71–80.

[18] James Clause and Alessandro Orso. 2010. LEAKPOINT: pinpointing the causes of memory leaks. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*. ACM, 515–524.

[19] Steve Cohen, George Smith, Richard A Chechile, Glen Burns, and Frank Tsai. 1996. Identifying impediments to learning probability and statistics from an assessment of instructional software. *Journal of Educational and Behavioral Statistics* 21, 1 (1996), 35–54.

[20] Lucas Cordeiro and Bernd Fischer. 2011. Verifying multi-threaded software using smt-based context-bounded model checking. In *Proceedings of the 33rd International Conference on Software Engineering*. ACM, 331–340.

[21] John W Creswell. 2002. *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall.

[22] Daniela S Cruzes and Tore Dybå. 2011. Recommended steps for thematic synthesis in software engineering. In *Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on*. IEEE, 275–284.

[23] Tore Dybå, Vigdis By Kampenes, and Dag IK Sjøberg. 2006. A systematic review of statistical power in software engineering experiments. *Information and Software Technology* 48, 8 (2006), 745–755.

[24] Stefan Endrikat, Stefan Hanenberg, Romain Robbes, and Andreas Stefik. 2014. How do api documentation and static typing affect api usability?. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 632–642.

[25] Ilker Ercan, Yaning Yang, Guven Özkaya, Sengul Cangur, Bulent Ediz, Ismet Kan, and others. 2008. Misusage of statistics in medical research. (2008).

[26] Filomena Ferrucci, Mark Harman, Jian Ren, and Federica Sarro. 2013. Not going to take this anymore: multi-objective overtime planning for software engineering projects. In *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 462–471.

[27] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.

[28] Christine A Franklin. 2007. Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework. American Statistical Association.

[29] Phillip I Good and James W Hardin. 2012. *Common errors in statistics (and how to avoid them)*. John Wiley & Sons.

[30] Sheila M Gore, Ian G Jones, and Eilif C Rytter. 1977. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. *BMJ* 1, 6053 (1977), 85–87.

[31] K.L. Gwet. 2014. *Handbook of Inter-Rater Reliability. The Definitive Guide to Measuring the Extent of Agreement Among Raters* (4 ed.). Advanced Analytics, LLC.

[32] M Sayeed Haque and Sanju George. 2007. Use of statistics in the Psychiatric Bulletin: author guidelines. *The Psychiatrist* 31, 7 (2007), 265–267.

[33] Hwa-You Hsu and Alessandro Orso. 2009. MINTS: A general framework and tool for supporting test-suite minimization. In *Software Engineering, 2009. ICSE 2009. IEEE 31st International Conference on*. IEEE, 419–429.

[34] Schuyler W Huck. 2009. *Statistical misconceptions*. Routledge.

[35] John P.A. Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine* 2, 8 (2005), 696–701. DOI:http://dx.doi.org/10.1002/stvr.1486

[36] David S Janzen, John Clements, and Michael Hilton. 2013. An evaluation of interactive test-driven labs with WebIDE in CS0. In *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 1090–1098.

[37] Lingxiao Jiang, Ghassan Misherghi, Zhendong Su, and Stephane Glondu. 2007. Deckard: Scalable and accurate tree-based detection of code clones. In *Proceedings of the 29th international conference on Software Engineering*. IEEE Computer Society, 96–105.

[38] 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.

[39] Carol Kilkenny, Nick Parsons, Ed Kadyszewski, Michael FW Festing, Innes C Cuthill, Derek Fry, Jane Hutton, and Douglas G Altman. 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PloS one* 4, 11 (2009), e7824.

[40] Andrew King, Sam Procter, Dan Andresen, John Hatcliff, Steve Warren, William Spees, Raoul Jetley, Paul Jones, and Sandy Weininger. 2009. An open test bed for medical device integration and coordination. In *Software Engineering-Companion Volume, 2009. ICSE-Companion 2009. 31st International Conference on*. IEEE, 141–151.

[41] B. Kitchenham, J. Fry, and S. Linkman. 2003. The case against cross-over designs in software engineering. In *Software Technology and Engineering Practice, 2003. Eleventh Annual International Workshop on*. 65–67.

[42] Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong. 2016. Robust Statistical Methods for Empirical Software Engineering. *Empirical Software Engineering* (2016), 1–52. DOI:http://dx.doi.org/10.1007/s10664-016-9437-5

[43] Fredrik Kjolstad, Danny Dig, Gabriel Acevedo, and Marc Snir. 2011. Transformation for class immutability. In *Proceedings of the 33rd International Conference on Software Engineering*. ACM, 61–70.

[44] Christian FJ Lange and Michel RV Chaudron. 2006. Effects of defects in UML models: an experimental investigation. In *Proceedings of the 28th international conference on Software engineering*. ACM, 401–411.

[45] Otávio Augusto Lazzarini Lemos, Fabiano Cutigi Ferrari, Fábio Fagundes Silveira, and Alessandro Garcia. 2012. Development of auxiliary functions: should you be agile? an empirical assessment of pair programming and test-first programming. In *Proceedings of the 34th International Conference on Software Engineering*. IEEE Press, 529–539.

[46] Rupak Majumdar and Koushik Sen. 2007. Hybrid concolic testing. In *Software Engineering, 2007. ICSE 2007. 29th International Conference on*. IEEE, 416–426.

[47] David Mandelin, Doug Kimelman, and Daniel Yellin. 2006. A Bayesian approach to diagram matching with application to architectural models. In *Proceedings of the 28th international conference on Software engineering*. ACM, 222–231.

[48] Mika V Mäntylä, Kai Petersen, Timo OA Lehtinen, and Casper Lassenius. 2014. Time pressure: a controlled experiment of test case development and requirements review. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 83–94.

[49] Collin McMillan, Mark Grechanik, Denys Poshyvanyk, Qing Xie, and Chen Fu. 2011. Portfolio: finding relevant functions and their usage. In *Proceedings of the 33rd International Conference on Software Engineering*. ACM, 111–120.

[50] Lijun Mei, WK Chan, and TH Tse. 2008. Data flow testing of service-oriented workflow applications. In *Proceedings of the 30th international conference on Software engineering*. ACM, 371–380.

[51] Habshah Midi, AHM Rahmatullah Imon, and Azmi Jaafar. 2012. The Misconceptions of Some Statistical Techniques In Research. *Jurnal Teknologi* 47, 1 (2012), 21–36.

[52] James Miller. 1999. Can results from software engineering experiments be safely combined?. In *Software Metrics Symposium, 1999. Proceedings. Sixth International*. IEEE, 152–158.

[53] Rahul Mohanani, Paul Ralph, and Ben Shreeve. 2014. Requirements fixation. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 895–906.

[54] Sebastian C Müller and Thomas Fritz. 2015. Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, Vol. 1. IEEE, 688–699.

[55] 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.

[56] Noboru Nakamichi, Kazuyuki Shima, Makoto Sakai, and Ken-ichi Matsumoto. 2006. Detecting low usability web pages using quantitative data of users' behavior. In *Proceedings of the 28th international conference on Software engineering*. ACM, 569–576.

[57] TH Ng, Shing Chi Cheung, WK Chan, and Yuen-Tak Yu. 2007. Do maintainers utilize deployed design patterns effectively?. In *Proceedings of the 29th international conference on Software Engineering*. IEEE Computer Society, 168–177.

[58] Raymond S Nickerson. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods* 5, 2 (2000), 241.

[59] Adrian Nistor, Qingzhou Luo, Michael Pradel, Thomas R Gross, and Darko Marinov. 2012. Ballerina: Automatic generation and clustering of efficient random unit tests for multithreaded code. In *Proceedings of the 34th International Conference on Software Engineering*. IEEE Press, 727–737.

[60] Aditya V Nori and Sriram K Rajamani. 2010. An empirical study of optimizations in YOGI. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*. ACM, 355–364.

[61] Renato Novais, Camila Nunes, Caio Lima, Elder Cirilo, Francisco Dantas, Alessandro Garcia, and Manoel Mendonça. 2012. On the proactive and interactive visualization for feature evolution comprehension: An industrial investigation. In *Proceedings of the 34th International Conference on Software Engineering*. IEEE Press, 1044–1053.

[62] Regina Nuzzo and others. 2014. Statistical errors. *Nature* 506, 7487 (2014), 150–152.

[63] Cara H Olsen. 2003. Review of the use of statistics in infection and immunity. *Infection and immunity* 71, 12 (2003), 6689–6692.

[64] Sangmin Park, Richard W Vuduc, and Mary Jean Harrold. 2010. Falcon: fault localization in concurrent programs. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*. ACM, 245–254.

[65] Fayola Peters, Tim Menzies, and Lucas Layman. 2015. LACE2: Better privacy-preserving data sharing for cross project defect prediction. In *Proceedings of the 37th International Conference on Software Engineering-Volume 1*. IEEE Press, 801–811.

[66] Yuhua Qi, Xiaoguang Mao, Yan Lei, Ziying Dai, and Chengsong Wang. 2014. The strength of random search on automated program repair. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 254–265.

[67] Steven P Reiss. 2008. Tracking source locations. In *Proceedings of the 30th international conference on Software engineering*. ACM, 11–20.

[68] Filippo Ricca, Massimiliano Di Penta, Marco Torchiano, Paolo Tonella, and Mariano Ceccato. 2007. The role of experience and ability in comprehension tasks supported by UML stereotypes. In *ICSE*, Vol. 7. 375–384.

[69] Paige Rodeghero, Collin McMillan, Paul W McBurney, Nigel Bosch, and Sidney D'Mello. 2014. Improving automated source code summarization via an eye-tracking study of programmers. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 390–401.

[70] Norsaremah Salleh, Emilia Mendes, John Grundy, and Giles St J Burch. 2010. An empirical study of the effects of conscientiousness in pair programming using the five-factor personality model. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*. ACM, 577–586.

[71] Jesper W Schneider. 2015. Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics* 102, 1 (2015), 411–432.

[72] Kenneth F Schulz, Douglas G Altman, and David Moher. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine* 8, 1 (2010), 18.

[73] Janet Siegmund, Christian Kästner, Sven Apel, Chris Parnin, Anja Bethmann, Thomas Leich, Gunter Saake, and André Brechmann. 2014. Understanding understanding source code with functional magnetic resonance imaging. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 378–389.

[74] Janice Singer. 1999. Using the American Psychological Association (APA) style guidelines to report experimental results. In *Proceedings of workshop on empirical studies in software maintenance*. 71–75.

[75] Ana Elisa Castro Sotos, Stijn Vanhoof, Wim Van den Noortgate, and Patrick Onghena. 2007. Students misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review* 2, 2 (2007), 98–113.

[76] Matt Staats, Gregory Gay, and Mats PE Heimdahl. 2012. Automated oracle creation support, or: how I learned to stop worrying about fault propagation and love mutation testing. In *Proceedings of the 34th International Conference on Software Engineering*. IEEE Press, 870–880.

[77] Denes Szucs and John Ioannidis. 2017. When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in Human Neuroscience* 11 (2017), 390.

[78] Jianbin Tan, George S Avrunin, and Lori A Clarke. 2006. Managing space for finite-state verification. In *Proceedings of the 28th international conference on Software engineering*. ACM, 152–161.

[79] Shin Hwei Tan and Abhik Roychoudhury. 2015. relifix: Automated repair of software regressions. In *Proceedings of the 37th International Conference on Software Engineering-Volume 1*. IEEE Press, 471–482.

[80] Matthew Thompson, Arpita Tiwari, Rongwei Fu, Esther Moe, and David I Buckley. 2012. A Framework To Facilitate the Use of Systematic Reviews and Meta-Analyses in the Design of Primary Research Studies. (2012).

[81] 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.

[82] Andrew Vickers. 2010. *What is a P-value anyway?: 34 stories to help you actually understand statistics*. Addison-Wesley Longman.

[83] Gerald E Welch and Steven G Gabbe. 1996. Review of statistics usage in the American Journal of Obstetrics and Gynecology. *American journal of obstetrics and gynecology* 175, 5 (1996), 1138–1141.

[84] Richard Wettel, Michele Lanza, and Romain Robbes. 2011. Software systems as cities: A controlled experiment. In *Proceedings of the 33rd International Conference on Software Engineering*. ACM, 551–560.

[85] Michael W Whalen, Suzette Person, Neha Rungta, Matt Staats, and Daniela Grijincu. 2015. A flexible and non-intrusive approach for computing complex structural coverage metrics. In *Proceedings of the 37th International Conference on Software Engineering-Volume 1*. IEEE Press, 506–516.

[86] Stefan Winter, Oliver Schwahn, Roberto Natella, Neeraj Suri, and Domenico Cotroneo. 2015. No PAIN, no gain?: the utility of PArallel fault INjections. In *Proceedings of the 37th International Conference on Software Engineering-Volume 1*. IEEE Press, 494–505.

[87] Chang Xu, Shing-Chi Cheung, and Wing-Kwong Chan. 2006. Incremental consistency checking for pervasive context. In *Proceedings of the 28th international conference on Software engineering*. ACM, 292–301.

[88] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. 2012. Does organizing security patterns focus architectural choices?. In *Proceedings of the 34th International Conference on Software Engineering*. IEEE Press, 617–627.

[89] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. 2015. Do security patterns really help designers?. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, Vol. 1. IEEE, 292–302.

[90] Yanbing Yu, James A Jones, and Mary Jean Harrold. 2008. An empirical study of the effects of test-suite reduction on fault localization. In *Proceedings of the 30th international conference on Software engineering*. ACM, 201–210.

[91] Fadi Zaraket, Adnan Aziz, and Sarfraz Khurshid. 2007. Sequential circuits for relational analysis. In *Software Engineering, 2007. ICSE 2007. 29th International Conference on*. IEEE, 13–22.

[92] Dina Zayan, Micha l Antkiewicz, and Krzysztof Czarnecki. 2014. Effects of using examples on structural model comprehension: a controlled experiment. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 955–966.

[93] Lingming Zhang, Dan Hao, Lu Zhang, Gregg Rothermel, and Hong Mei. 2013. Bridging the gap between the total and additional test-case prioritization strategies. In *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 192–201.

Table 7: Problems found in standaline experiments and ”experiments as evaluations” published in ICSE between 2006-2015

Empirical Studies	Code	Goal definition	Study design			Analysis			Reporting	
		(Q1) *Null hipothesis*	(Q2) *Sample size calculation*	(Q3) *Random sampling*	(Q4) *Random assignment*	(Q5) *Assumptions*	(Q6) *Model definition*	(Q8) *Post-hoc power*	(Q9) *Multiple testing*	(Q10) *Means*
Standalone Experiments	2006-EX01 [56]	No/Yes	No	No	Yes	No	N/A	No	No	Yes
	2006-EX02 [44]	Yes/Yes	No	Yes	No	Yes	N/A	Yes	No	Yes
	2007-EX03 [57]	Yes/No	No	No	No	No	N/A	Yes	No	Yes
	2007-EX04 [68]	Yes/Yes	No	No	Yes	No	N/A	Yes	No	Yes
	2008-EX05 [90]	No/No	No	Yes	Yes	N/A	N/A	N/A	N/A	Yes
	2010-EX06 [70]	Yes/Yes	No	No	Yes	No	N/A	Yes	No	Yes
	2011-EX07 [84]	Yes/Yes	No	No	Yes	Yes	N/A	Yes	No	Yes
	2012-EX08 [14]	Yes/Yes	No	No	Yes	Yes	N/A	Yes	No	Yes
	2012-EX09 [88]	Yes/Yes	No	No	Yes	Yes	N/A	Yes	No	Yes
	2012-EX10 [61]	Yes/Yes	No	No	No	Yes	N/A	Yes	N/A	Yes
	2012-EX11 [45]	Yes/Yes	No	No	Yes	Yes	N/A	Yes	No	Yes
	2013-EX12 [7]	No/No	No	No	No	Yes	N/A	Yes	Yes	Yes
	2014-EX13 [92]	Yes/Yes	No	No	Yes	Yes	N/A	Yes	No	Yes
	2014-EX14 [48]	Yes/Yes	No	Yes	Yes	Yes	N/A	Yes	No	Yes
	2014-EX15 [73]	No/No	No	Yes	No	No	N/A	Yes	N/A	N/A
	2014-EX16 [24]	No/No	No	No	Yes	Yes	N/A	Yes	No	Yes
	2014-EX17 [53]	Yes/Yes	No	Yes	Yes	Yes	N/A	Yes	N/A	Yes
	2015-EX18 [86]	Yes/Yes	No	Yes	N/A	No	N/A	Yes	Yes	Yes
	2015-EX19 [89]	Yes/No	No	No	Yes	Yes	N/A	Yes	No	Yes
	2015-EX20 [54]	No/No	No	No	Yes	Yes	Yes	No	No	Yes
	2015-EX21 [3]	No/Yes	No	No	No	No	N/A	Yes	No	Yes
”Experiments as Evaluations”	2006-CM01 [78]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	N/A
	2006-CM02 [87]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	No
	2006-CM03 [47]	No/No	No	No	Yes	N/A	N/A	N/A	N/A	No
	2007-CM04 [91]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2007-CM05 [37]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	N/A
	2007-CM06 [46]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2008-CM07 [50]	No/No	No	Yes	N/A	N/A	N/A	N/A	N/A	Yes
	2008-CM08 [67]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2008-CM09 [17]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2009-CM10 [40]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2009-CM11 [33]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2009-CM12 [9]	Yes/No	No	No	N/A	Yes	Yes	Yes	No	Yes
	2010-CM13 [60]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2010-CM14 [18]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2010-CM15 [64]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2011-CM16 [43]	N/A/N/A	No	No	N/A	N/A	N/A	N/A	N/A	N/A
	2011-CM17 [49]	Yes/Yes	No	No	Yes	No	N/A	Yes	No	Yes
	2011-CM18 [20]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2012-CM19 [13]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2012-CM20 [59]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2012-CM21 [76]	Yes/Yes	No	Yes	N/A	N/A	N/A	Yes	No	Yes
	2013-CM22 [26]	No/No	No	No	N/A	Yes	N/A	Yes	Yes	Yes
	2013-CM23 [36]	No/No	No	No	Yes	No	N/A	Yes	No	Yes
	2013-CM24 [93]	No/No	No	Yes	N/A	No	N/A	Yes	No	No
	2014-CM25 [66]	No/No	No	No	N/A	No	N/A	Yes	No	Yes
	2014-CM26 [1]	No/No	No	No	Yes	Yes	N/A	Yes	No	No
	2014-CM27 [69]	Yes/No	Yes	Yes	Yes	Yes	N/A	Yes	No	Yes
	2015-CM28 [79]	No/No	No	No	N/A	N/A	N/A	N/A	N/A	Yes
	2015-CM29 [85]	No/No	No	No	Yes	No	N/A	Yes	N/A	Yes
	2015-CM30 [65]	No/No	No	No	N/A	No	N/A	Yes	N/A	Yes