Problems with Statistical Practice in Human-Centric Software Engineering Experiments

Barbara Kitchenham
School of Computing and Mathematics, Keele University, Keele Staffordshire, ST5 5BG, UK
b.a.kitchenham@keele.ac.uk

Lech Madeyski*
Faculty of Computer Science and Management, Wroclaw University of Science and Technology
Wyb.Wyspianskiego 27, 50370 Wroclaw, Poland
lech.madeyski@pwr.edu.pl

Pearl Brereton
School of Computing and Mathematics, Keele University, Keele Staffordshire, ST5 5BG, UK
o.p.brereton@keele.ac.uk

ABSTRACT

Background Examples of questionable statistical practice, when published in high quality software engineering (SE) journals, may lead to novice researchers adopting incorrect statistical practices.
Objective Our goal is to highlight issues contributing to poor statistical practice in human-centric SE experiments.
Method We reviewed the statistical analysis practices used in the 13 papers that reported families of human-centric SE experiments and were published in high quality journals.
Results Reviewed papers related to 45 experiments and involved a total of 1303 human participants. We searched for issues that were related to questionable statistical practice that were found in more than one paper. We observed three types of bad practice: incorrect use of terminology, incorrect analysis of repeated measures designs, and post-hoc power testing. We also found two analysis practices (i.e., multiple testing and pre-testing for normality) where statisticians disagree about good practice.
Conclusions Identified issues pose a problem because readers may expect the statistical methods used in papers published in top quality, peer-reviewed journals to be correct. We explain why the practices are problematic and provide recommendations for improved practice.

CCS CONCEPTS

• **General and reference** → **Empirical studies**; **Experimentation**; • **Software and its engineering** → **Software creation and management**.

KEYWORDS

empirical software engineering, statistical practice problems, good practice guidelines, crossover designs, human-centric experiments

*Corresponding author.

ACM Reference Format:
Barbara Kitchenham, Lech Madeyski, and Pearl Brereton. 2019. Problems with Statistical Practice in Human-Centric Software Engineering Experiments. In *Evaluation and Assessment in Software Engineering (EASE '19), April 15–17, 2019, Copenhagen, Denmark*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3319008.3319009

1 INTRODUCTION

As part of a study of the use of meta-analysis in families of human-centric software engineering (SE) experiments[1], we identified 13 papers that had been published in high quality journals [29]. During our assessment of each paper, we reviewed aspects of the statistical design of the individual experiments in the family. We observed a variety of terminology errors and misunderstandings of statistical practice in individual papers that were duplicated in other papers.

We report and discuss these issues for three main reasons:

(1) The issues we discuss occurred in more than one paper after a peer review process. This suggests that they may be examples of issues that are not well-understood by experienced SE researchers and reviewers of top SE journals.
(2) The papers were published in leading SE journals, so there is an expectation that the reported methodology would be correct. Thus, if incorrect terminology or methodology is reported in such papers it may lead to the propagation of bad practice as novice researchers assume the methodology must be correct.
(3) Families of experiments usually involve many human participants. Scientific ethics are clear that we need to be careful to avoid abusing our participants, even if it is just wasting their time and effort by incorrectly analysing their data.

The issues we observed were a mixture of incorrect use of terminology and misunderstanding statistical principles. Specifically, the issues we consider in this paper are:

(1) Misuse of statistical terminology.
(2) Incorrect analysis of repeated measures experiments.
(3) Performing post-hoc power analysis.
(4) The use of pretesting for normality and variance stability.
(5) Testing multiple hypotheses.

After a brief discussion of our methodology in Section 2, we discuss each issue in a separate section. We discuss our findings in Section 8 and present our recommendations in Section 9.

[1]By *human-centric* we mean experiments that depend on human participants to perform skill-based software engineering tasks.

Compared with [29], in this paper, we concentrate on statistical issues related to the analysis of the individual experiments. In [29], we concentrate on issues related to meta-analysis. The contributions of this paper are:

(1) To identify a set of statistical problems and issues that we observed in papers published in high quality software engineering journals.
(2) To explain issues that indicate misunderstanding of statistical principles, and to discuss issues where statistical advice is contradictory.
(3) To provide practical recommendations based on our discussion.

2 MATERIALS AND METHODS

Our search process was done independently of the mapping study by Santos et al. which aimed to classify all software engineering families of experiments (i.e. [50]). However, we used Santos et al.'s mapping as a means of validating the completeness of our search and selection process. Our search was intended to find families of experiments in top ranked software engineering journals that undertook meta-analysis to aggregate the results of each experiment.

We restricted our search to papers in the following five journals that were published between 1999 (when the first paper on SE families of experiments was published [3]) and 2017:

- IEEE Transactions on Software Engineering (TSE).
- Empirical Software Engineering (ESE).
- Journal of Systems and Software (JSS).
- Information and Software Technology (IST).
- ACM Transactions on Software Engineering Methodology (TOSEM).

We searched SCOPUS for studies with the term "family of experiment*" and restricted the search to the five specified journals. This missed a paper that we knew was relevant which did not use the main search term [31]. So then we used the search term, "replicat" AND "meta-analysis" with the same journal restrictions. This found [31] and another relevant paper [44]. Applying our inclusion criteria: families of three our more studies that compared software engineering techniques by asking human participants perform software engineering tasks, we found 13 candidate studies. We repeated these searches on Semantic Scholar and did not find any additional relevant studies. Compared with [50], we found one extra paper [39] and rejected three (two because they were not published in the five selected journal and one because it was observational correlation study correlating team personality and climate to satisfaction and quality [2].)

Overall, the maximum number of experiments per family was five and the 13 papers reported analysis of 45 experiments, that involved a total of 1303 human participants. In all the papers, the authors attempted to meta-analyse the results of the individual experiments comprising the family of experiments.

All three authors read the full text of these papers. The first author extracted detailed data about the experimental methodology from the papers which was validated by the third author. The first author identified possible issues during data extraction, that were reviewed and agreed by the second author.

A limitation of this paper is that the set of papers we studied is a subset of the papers identified in [50]. A review of the titles of the studies found by Santos et al. indicates that another 12 papers conformed with all our search criteria except that they did not undertake meta-analysis. Obviously we cannot be sure that the issues we identified applied to the other papers nor that there are other issues in those papers that we have not detected in the papers we reviewed.

Please note that our aim is not to criticize specific papers, but to identify general problems with current statistical practices in software engineering. We note that all the authors reported their procedures in sufficient detail for us to understand how they analyzed their experimental results. There is clearly no intention on the part of the authors to mislead their readers, but it seems that the authors have not benefited as much as we all might hope from the peer review process.

3 INCORRECT TERMINOLOGY

It may be considered unnecessary to worry about the misuse of statistical terminology, but we believe using the right terminology is important because:

- Using the wrong terminology may be indicative of a deeper misunderstanding of statistical methods.
- If researchers do not understand statistical terminology, they will have problems understanding existing statistical text books and new statistical results.
- If junior researchers adopt incorrect terminology, this may cause problems if they move on to work in other disciplines or build their own research teams.

Table 1 reports the names that researchers gave to the methods they used to analyze their experiments. In our opinion only two studies used completely correct terminology to describe their experiment (see [22, 44]). Four papers referred to crossover experiments as "counterbalanced" designs, which although not incorrect, does not fully specify the design. Six papers referred to the design as a factorial which is inconsistent with the usual use of the term factorial in applied statistics (see below).

Most of the studies actually used crossover designs and Vegas et al. [57][2] have already reported problems with the terminology used for crossover experiments. The basic AB/BA crossover design is shown in Table 2. The term AB/BA crossover is used to describe this design in the medical literature (e.g., [52]). In a medical application such as comparing the effectiveness of different drugs, the conditions in the experiment would simply be the drug being used in each time period. In a software engineering context, we ask participants to perform a task such as finding defects or adding a new requirement using one technique, and then perform the same task using another technique. If we asked software engineers to repeat the same task using the same software document, they might simply remember what to do without applying the technique. For that reason, it is necessary to change the materials used in the second time period. However, the experiment we call a 4-group crossover (which is shown in Table 3), does not appear to be used

[2]Our study includes the time period 2015-2017 which was not covered by Vegas's review and three sources that were not searched by Vegas: ESE, IST and JSS.

Table 1: Experimental Designs Reported and Actual

Claimed Design	Actual Design
Counterbalanced design organizing participants into 4-groups [1]	4-Group crossover
Within-participants counter-balanced design [51]	4-Group crossover
2 by 2 Factorial with dependent variables [12]	AB/BA crossover
A balanced within-subject design with a confounding effect [20]	4-Group crossover
A balanced factorial design with group-interaction as a confounding factor [21]	4-Group crossover
Counterbalanced design [24]	4-Group crossover
2 by 2 factorial design with confounded interaction [55]	AB/BA crossover
A balanced within-participants design with a confounding effect [23]	4-Group crossover
Between-subjects balanced design [22]	Between subjects design
A balanced factorial design [13]	4-Group crossover
2 by 2 factorial design with confounded interaction [39]	AB/BA crossover
2 by 2 factorial Counterbalanced Repeated-Measures Design [31]	AB/BA crossover
Pre-test and post-test control group [44]	Pre-test and post-test control group

in other disciplines, so does not have a well-defined name in the statistical literature. Vegas et al. call this a "two-treatment factorial crossover design where the experimental object is a two-level blocking variable". We agree that the design is a two-treatment crossover with an additional blocking variable. However, we do not agree with referring to this sort of design as a "factorial" experiment.

Table 2: AB/BA Crossover Design

Group	Period 1	Period 2
A	Technique 1	Technique 2
	Materials 1	Materials 2
B	Technique 2	Technique 1
	Materials 1	Materials 2

Table 3: 4-Group Crossover Design

Group	Period 1	Period 2
A	Technique 1	Technique 2
	Materials 1	Materials 2
B	Technique 2	Technique 1
	Materials 1	Materials 2
C	Technique 1	Technique 2
	Materials 2	Materials 1
D	Technique 2	Technique 1
	Materials 2	Materials 1

The term *factorial* experiment is used to describe a design where two or more *treatment* factors (usually those that can be applied at different levels such as the weight in tons of different fertilisers, e.g., phosphate and nitrate, applied to plots in an agricultural experiment) are investigated *together* (see for example, [5, 37, 53]). In industrial experiments, this type of design is used with multiple levels of multiple factors to assess the combination of treatment levels that achieve maximum output [4]. Importantly, understanding the interaction between the treatments is the essential goal of the experiment.

Darcy et al. [14] report an example of a software engineering study (which included replication) that was a factorial study. The main hypothesis concerned the impact of coupling and cohesion on maintenance. For this they asked participants to perform a maintenance task on objects with different levels of complexity and cohesion as shown in Table 4. Each of the four experimental conditions involve one of two levels of cohesion: low cohesion ("LCoh") and high cohesion ("HCoh"), and one of two levels of coupling: Low coupling ("LCou") and high coupling ("HCou"), and all four possible combinations are used. Although their experiment involves other conditions (and repeated measures), the basic treatment conditions comprise a factorial experiment. Furthermore the interaction between coupling and cohesion was an important element of the experiment.

Table 4: Factorial Design for Investigation of Code Complexity

		Coupling Level	
		Low	High
Cohesion	Level		
	Low	LCoh, LCou	LCoh, HCou
	High	HCoh, LCou	HCoh, HCou

However, in our opinion, factors such as different software documents used to perform software development tasks in a software experiment are best considered as blocking factors not treatment factors, where blocking factors are defined as sources of variability that are not of primary interest to the experimenter (and are often referred to as nuisance factors). In general, interactions between nuisance blocking factors are usually not included in any analysis. The treatment factor in crossover designs comprises two levels i.e., technique 1 and technique 2 and the levels do not interact (i.e., each *cell*[3] in a crossover design uses only one technique).

A type of blocking factor that is not a nuisance factor is a blocking factor that partitions the participant population, for example, skill or experience levels in software engineering, and male or female in medical studies. In studies where skill or experience is an issue, it may be important to investigate whether the impact of a new technique is influenced by the experience of the participants. This type of factor can be incorporated into a crossover style design, but it would lead to additional sequence groups and might be better addressed as a between participants randomized block

[3]A cell is a combination of experimental conditions from which a set of comparable observations are obtained.

experiment, if the factor of interest is categorical, or, as recommend by Maxwell [37], if the factor of interest is better considered as a continuous variable, it can be measured for each participant and used as a covariate in an analysis of covariance (ANCOVA).

In general, experimental design terminology of experiments which do not have repeated measures is relatively straightforward. An experiment without blocks is called a randomized experiment, and experiment with blocking factors is called a randomized blocks experiment. Specific blocks may have more than two levels. Randomized blocks experiments can have more that one type of blocking factor and more than two treatments to compare. A Latin squares design has three or more treatments and three or more different types of block, both of which must have the same number of levels as the number of treatments. A Latin squares design ensures that each treatment is trialled under each combination of block levels. Graeco-Latin squares can be used if there are three types of blocking variable. However, we recommend keeping to comparisons of two treatments, with only the most important blocking factors included in the design.

4 INCORRECT ANALYSIS OF REPEATED MEASURES STUDIES

Vegas et al. [57] reported that crossover designs were sometimes incorrectly analyzed. We observed the same problem. Out of 12 studies that used repeated measures designs, 8 studies did not use a statistical analysis method that catered for repeated measures. They used analysis methods suitable for testing independent groups. As discussed in [36], the impact of this is that researchers:

- Over-estimated their degrees of freedom (basing them incorrectly on the number of observations rather than the number of participants). This would increase the risk of detecting false positives, by incorrectly reducing both the standard error of the mean difference and the critical value of the t-test.
- Obtained an estimate of the variance based on the between-participants variation rather than the within-participants variation. The between-participants variance should be much larger than the within-participants variance if there is a correlation between the outcomes that participants obtained in the first time period and the outcomes they obtained in the second time period. Correlation between outcomes from the same participant is an indication that ability influences the experiment outcomes. Using the between-participant variance will increase the risk of failing to detect genuine effects.

Another impact is that researchers who did not understand how to correctly analyze their study also undertook additional tests to check factors that should have been addressed already by an appropriate analysis:

- 8 papers investigated order and time period effects and interactions for each of their experiments, although, in general, statisticians do not test for interactions among blocking factors.
- 4 papers investigated the impact of different participant ability in each of their experiments, although for appropriate analysis of crossover designs, participants are their own controls removing the impact of ability differences.
- 6 papers that used the 4-group design investigated the difference between materials, although the experimental design was designed to balance such effects and a statistical analysis based on that design would have accounted for any difference between materials.

The main problem is that undertaking extra (and unnecessary) tests compromises the power of experiments and can increase the risk of obtaining spurious positive results. This is discussed further in Section 7.

5 POST-HOC POWER ANALYSIS

Five studies performed post-hoc power analysis, although power analysis should be used during experimental design (i.e., when pre-planning the sample size), not after the experiment has been completed and analyzed (as indeed was mentioned in one of the papers).

Power is the probability of rejecting the null hypothesis when it is false. As explained by Dybå et al. [16], it is determined by three factors:

- The required significance level α and its directionality, where the smaller the value of α, the lower the power. In addition, a non-directional two tailed test will have lower power than a directional one-tailed test at the same α.
- The effect size. The larger the effect size, the higher the power.
- The sample size. The larger the sample size the higher the power.

When planning an experiment, researchers need to determine the sample size needed to ensure an adequate power level, where 0.8 is often considered to be an appropriate level. However, during planning, the effect size is unknown, which has led to researchers mistakenly reporting the power of their experiment from the observed effect size. This is an error because:

(1) Once the experiment has been performed, a post-hoc analysis of power only confirms that if we find a statistically significant effect size, the calculated power is large and if we find a non-significant effect the calculated power is small. Thus, the power analysis has told us nothing of interest about our experiment.
(2) In the event that there is no significant effect, we might wrongly assume this is due to the low power of the experiment, *even if the sample size is large*. In fact, a non-significant effect size with a large power is strong evidence in favour of the null hypothesis.
(3) In the event that there is a statistically significant effect and a small sample size, post-hoc analysis will indicate a high power value, but it is far more likely that the experiment has overestimated the true effect size. Furthermore, subsequent replications using the same sample size on the assumption that it is sufficiently powerful, will then be under-powered leading to further misleading results [25].

See Button et al. [9] for a more detailed explanation of the problems of low power.

As pointed out by Jørgensen et al. [26], what we need to do is decide in advance what *standardized* effect size we think is important, in terms of Cohen's large, medium and small criteria (see [11]), and use a power analysis to determine what sample size is necessary to achieve a power level of 0.8 (i.e., 80% chance of detecting an effect if one genuinely exists). Then, given sufficient power *a priori* to detect, for example, a medium effect size, a non-significant effect is much more likely to indicate a negligible effect size, and a significant effect more likely to provide an unbiased estimate of the true effect size. Only in the case of a relatively large but not statistically significant effect size, should a post-hoc power analysis be used to indicate the sample sizes needed in subsequent experiments.

It should also be noted that the significance tests for repeated measures experiments should be based on the *within-participant* variance, which itself depends on the correlation between the repeated measures, such as crossover designs. The guidelines reported by Cohen do not cover these types of experiment, and researchers need to ensure that any power analysis tools they use can handle such studies.

6 PRE-TESTING FOR NORMALITY AND VARIANCE HETEROGENEITY

Six studies pretested their data for normality before deciding whether to use non-parametric or parametric analyses. Two other studies performed non-parametric and parametric tests and reported the parametric test results, since both tests gave similar results. In addition five studies (including four of the studies that tested their data for normality) reported that they tested for variance stability.

The issue of pre-testing for normality and heterogeneity is complex and there is considerable disagreement in the literature. The theoretical problem with pre-testing is that as with any statistical process that uses multiple testing the Type 1 and Type 2 errors of the analysis may be affected by the multiple tests.

As reported by Rochon et al. [48], historically, most medical researchers used t-tests for statistical testing. However, during the 1980s and 1990s, there was concern about the number of statistical errors in the medical literature, with violation of assumptions being one of the most common problems. As a result, guidelines for publication in medical journals emphasized the importance of distributional assumptions. To address this issue, researchers started to adopt the process of pre-testing for normality and variance heterogeneity. However, Rochon et al. also point out that other researchers questioned the validity of pre-testing, for a number of theoretical reasons:

(1) Failure to reject the assumption of normality, or the assumption of equal variances does not mean that the assumption is true. This is a particular problem for software engineering experiments since tests of normality have low power to detect significant deviations from normality when sample sizes are small.
(2) Some preliminary tests come with their own assumptions, which may need testing.
(3) Preliminary tests are usually applied to the data to be analysed, which can result in uncontrolled Type 1 and Type 2 error rates.

Arguments for and against pre-testing, have given rise to many simulation studies intended to investigate the practical implications of pretesting. We present the results from three such studies to illustrate the variety of current opinions[4].

Rasch et al. [45] performed simulation studies based on two independent groups design using 5 combinations of equal and unequal group sizes ($n_1 = n_2 = 10, n_1 = n_2 = 30, n_1 = 10$ and $n_2 = 30$, $n_1 = 30$ and $n_2 = 10$, $n_1 = 30$ and $n_2 = 100$) with both equal and unequal variances. They simulated data from five distributions, one of them was the standard normal distribution, while the others were selected from different combinations of skewness and kurtosis. They tested for normality first using the Kolmogoroff-Smirnov test, and then for heterogeneity using the Levene test [33]. If the data rejected normality, they used the Wilcoxon-Mann-Whitney (WMW) U-test. If the data passed the test for normality, but failed the test for heterogeneity, they used Welch's test ([58]), otherwise they used a t-test. They simulated 100,000 data sets for each combination of distribution, size, variance and effect size

They showed that such pre-testing leads to unknown final Type 1- and Type-2 risks if the respective statistical tests are performed using the same set of observations. They concluded that:

- It is preferable to apply no pretests, and to use the Welch-test as a standard.
- The WMW U test could not be recommended for most cases.

Rochon et al. [48] simulated an independent groups experiment, using equal sample sizes, three different underlying distributions and two different pretest strategies. Strategy 1 involved testing data from each of the groups for normality using the Shapiro-Wilk test and then using a t-test if both tests suggested the data as normal, and the WMW U test if either test indicated non-normality. Strategy 2 involved calculating the residual from the mean in each group and then pooling the data to test it for normality. They did not test for unequal variances nor did they consider unequal variances in their simulations. They considered only equal sample sizes of 10, 20 30, 40, and 50 *per group* and three underlying distributions: the normal, uniform and exponential distributions. They used the Shapiro-Wilk test to check for normality. They simulated 100,000 data sets for each combination of sample size, distribution and effect size.

They reported "dramatic effects of preliminary testing for normality on the *conditional* Type 1 error rate of the main test" which although similar for the two strategies were more pronounced for Strategy 2. However, the effect of the *unconditional* decision strategy was very small for both Type 1 and Type 2 errors. The definition of a conditional probability of a Type 1 error using a t-test, in the context of pre-testing for normality, is the probability that a data set that has passed a test for normality subsequently reports a significant effect from a t-test when none exits. Basically it is the probability of a t-test reporting a false positive after the data set has passed a test for normality. A similar definition applies to the conditional probability of a Type 1 error using a non-parametric test. The unconditional probability of a Type 1 error using a pre-testing process is the probability of Type 1 error using the statistical test decided by the pre-test results.

[4] All the tests for normality discussed in this section are described in [46].

Operationally, the process used to estimate the conditional probability of Type 1 error for a t-test, is obtain N samples of an independent groups experiment with the mean difference of zero that pass the test for normality (irrespective of the true underlying distribution). This means N is set in advance (to say 100000) and sampling continues until N data set samples pass the test for normality, then each of the N samples is tested using a t-test and the number of false positives at a pre-decided alpha level are counted, then:

$$P_{Cond}(Type1Error) = \frac{NumFalsePositives}{N} \quad (1)$$

A similar process is used to calculate the probability of a Type 1 error given the use of a non-parametric test.

In contrast, to estimate the unconditional probability of a Type 1 error given a pre-test for normality, N samples (with mean difference=0) are obtained and each is tested for normality. Samples that pass are subjected to a t-test, samples that do not are subjected to a non-parametric test, then:

$$P_{Uncond}(Type1Error) = \frac{Nt_{FalsePositives} + Nnp_{FalsePositives}}{N} \quad (2)$$

where $Nt_{FalsePositives}$ is the number of false positives obtained from t-tests and $Nnp_{FalsePositives}$ is the number of false positives obtained from the non-parametric tests.

They concluded that:

- Although theoretically incorrect, pre-testing did not cause much harm.
- At worst it was unnecessary for large samples, e.g., samples with more that 40 observations in each group, since the $t-$test was sufficiently robust.
- In the case of small samples, the Shapiro-Wilk test is insufficiently powerful to detect deviations from normality, so non-parametric methods should be preferred.

Lantz et al. [32] used simulation to investigate pre-testing of *three* independent groups design. They used equal sample sizes in each group with three sizes: n=15, n=30 and n=60 per group. They assessed each group for normality using the Shapiro-Wilk test, and if any one of the groups failed the normality test, they used the Kruskal-Wallis test, otherwise they used a one-way ANOVA. They did not test for heterogeneity, neither did they simulate unequal variances. They simulated data from 4 distributions: normal, exponential, Laplace and uniform. They simulated data for five different standardized mean difference effect sizes: 0, 0.10, 0.25, 0.40, 0.65. They also considered four different significance levels for the Shapiro-Wilk test. They simulated 100000 data sets for each set of conditions. They mentioned that Rochon et al. found differences between the conditional error rates and the unconditional errors rates, and used the unconditional error rates to assess the overall two-stage process.

They reported that the two stage process seemed to be a better choice for testing the difference between means of three groups:

- Simply using either ANOVA or the Kruskal-Wallis test as a one stage process did not perform noticeably better than the two stage procedure.
- For strong non-normality the two stage process was much better than simply using ANOVA.

Thus, three fairly recent simulation studies managed to come to rather conflicting results. The studies all used different simulation methods. For example only one considered unequal variances as well as normality, and while two studies used the Shapiro-Wilk test for normality, the other used the Kolmogoroff-Smirnov test. They used different distributions to simulate non-normal data. The only commonality between the studies is that they all used 100,000 simulations for each condition they investigated.

In terms of their relevance to SE experiments, the experimental conditions investigated were simple compared with crossover designs and did not include any repeated measures studies. The researchers also used sample sizes rather larger than we usually use in SE research. In the 13 papers we studied, there were 45 different experiments with the average number of participants per experimental group being 10.7 and the median being 6.4. For the seven papers that used the 4-group crossover design, there were 25 separate experiments with a mean group size of 6.1 and a median of 6.

With respect to tests for non-normality, many researchers have reported that the Shapiro-Wilk test has more power to detect departures from normality than the Kolmogoroff-Smirnov test, thus the former is more reliable than the latter [34, 38, 46, 56]. As a result, it is possible to have a situation when the Shapiro-Wilk test is significant but the Kolmogoroff-Smirnov test is not [34]. Furthermore, the larger the samples, the easier it is to obtain significant results from even small departures from normality. Hence, a significant test result in such a situation should be interpreted with an understanding of this effect. However, in spite of many studies favouring the Shapiro-Wilk test, recently, several other studies have suggested that the Anderson-Darling test is to be preferred [27, 28, 43].

Finally, the three simulation studies, [32, 45, 48], did not explicitly consider the impact of outliers, although a recent paper by Derrick et al. [15] identified problems for paired studies with extreme observations. They demonstrated that for small sample sizes a single extreme outlier could reduce the power of the standard t test, *even if the outlier is in the same direction as the effect size of the other observations*. This occurs because the estimated variance can be inflated enough to undermine the ability of the t test to detect a genuine effect. After a simulation study, they reported that Yuen's paired samples t-test (i.e., a test performed on trimmed data) and the Wilcoxon signed rank sum test both exhibited robust behaviour in the presence of a single outlying observation.

Rietveld and van Hout's tutorial [47] provides a good discussion of the problems associated with testing assumptions, and the advantages and disadvantages of various analysis methods in the presence of non-normality. They conclude that we should:

- Report more information about our data (means, medians, variance, skewness, tailedness (i.e., kurtosis), outliers etc.).
- *Not* routinely use conventional non-parametric tests like the WMW test in case one or more of the assumptions of t tests are not met.
- Consider using less conventional, but robust statistics which have been developed and tested in the last decades, for example, the probability of superiority [6] or Cliff's d [10]..

We agree with these recommendations and, additionally, are strong advocates of reproducible research [35] and robust statistical methods [30] in empirical software engineering research. If we reported more detailed information about human-centric experiments in software engineering research, we would be in a better position to choose appropriate analysis methods.[5]

Also, there is considerable evidence suggesting conventional analysis of ranked data such as the WMW test and the Kruskall-Wallis test, are not robust, see for example [17–19, 60]. So, if we have small non-normal data sets we need to use robust tests such as those developed by Brunner and his colleagues (see [7] or [8]) or Cliff [10].

7 TESTING MULTIPLE HYPOTHESES

All but two of the 13 papers we studied included multiple main hypotheses. The largest number of main hypotheses tested was 6 (which happened twice). The multiple tests occurred when researchers investigated several different outcome measures. However, most studies performed other ancillary tests, such as testing for normality and variance equality and testing for factors such as participant skill, the impact of blocking factors such as time period and materials and interactions between such factors.

The problem with multiple tests on the same data set is that multiple testing increases the risk of detecting spurious significant results. However, as Nakagawa points out [41], using adjustments such as the Bonferroni adjustment or its less stringent sequential variants decrease the power of experiments and increase the risk of missing significant results. The problem with multiple comparisons, the Bonferroni correction and interesting alternatives to the Bonferroni correction are further discussed by Madeyski [34] and Maxwell [37].

There are several valid methods to address this issue in general:

- We can state in advance our main hypothesis and report all other tests as exploratory. However, for auditability, this depends on publishing our detailed analysis plan before performing our experiments.
- As proposed by Nakagawa, we should report effect sizes and their confidence intervals, rather than just the results of significance tests.
- We can apply less stringent adjustment methods to maintain an experiment-wide control of Type 1 errors, such as the method proposed by Rom [49]. Rom's method was judged the best of five methods (in terms of power) in a comparative study [42].
- Nakagawa also mentions the possibility of controlling the false discovery (FDR), which is intended to balance the probability of a Type 1 error, while maintaining a reasonable power level (see [37, 54])[6].

What we should never do is to report only statistically significant results. It is encouraging to note that there is *no* evidence that any of the authors of the 13 papers we reviewed failed to report any results. Authors reported non-significant results as well as significant results. In terms, of individual hypothesis tests, we identified 130 hypothesis tests across all the primary studies of which 80 (i.e., 61%) were reported to be non-significant.

However, we, as researchers, need to take pains to minimize the number of tests we perform on individual data sets. We should always analyze our data in a manner that is appropriate given the basic experimental design, to avoid performing unnecessary tests of factors that are addressed by, or balanced by, the design. We should also try to avoid the use of multiple outcome measures. For example:

(1) If we have been collecting data about different aspects of a characteristic, such as measuring understandability and modifiability as aspects of maintainability, it may be preferable to analyze the average of the two outcomes. This is particularly important if the aspects are likely to be correlated where it may be misleading to analyze the results separately. For example, in the case of understandability, there is a strong *a priori* probability that a participant's score for an understandability task would be correlated to their score for a modifiability test. So obtaining positive results for tests for understandability and modifiablity does not give additional support for a new technique compared with a single test of the average outcome.
(2) If we have been collecting data about effectiveness and efficiency, we would suggest nominating effectiveness as the main outcome criteria. There are several reasons. Firstly, if a technique is new to the participants, we would not expect them to use it as efficiently (i.e., quickly) as possible, the first few times they use it. Secondly, if efficiency is measured as the number of correct answers or correctly performed tasks per unit time, it does not properly penalize incorrect answers. The impact of an mistake in an experiment is not the same as the impact of a mistake while maintaining a real software product, where, if software tasks are incorrectly performed, there can be substantial rework costs. So the ratio of correct answers to total time could be very misleading, since it would not distinguish between a participant who only attempted to answer eight of ten questions in the allotted time period and got all of them correct, from a participant who answered ten questions in the allotted time period but only got eight correct.[7]

8 DISCUSSION

Based on a study of 13 papers published in high quality journals, which reported using families of experiments, we found a number of issues related to statistical analysis of human-centric SE experiments. Some of the issues related to misunderstanding statistical methods (i.e., wrong terminology, failure to consider the impact of

[5] If researchers report more information about the properties of their data sets, they must make it clear whether they are reporting statistics obtained from raw data or from residuals.

[6] Maxwell provides a very detailed discussion of the concepts underlying FDR and identifies that a number of researchers believe it to be a useful concept. However, he does not include FDR in the flow chat he presents to indicate the most appropriate methods of adjusting for multiple tests under different circumstances.

[7] This argument could also be applied to effectiveness, where it might be preferable to measure effectiveness in terms of correct answers minus incorrect answers.

repeated measures during analysis, performing unnecessary statistical tests, and post-hoc power analysis), others related to issues where there is disagreement among statisticians as to what constitutes best practice (i.e., pre-testing normality assumptions and performing multiple tests on the same data set). In the case of incorrect analysis we have proposed more valid methods, in the case of statistical controversies, we have discussed the issues and provided some advice.

We believe one of the underlying problems is the adoption of complicated statistical designs, without fully appreciating that complicated designs will imply complicated analysis methods. In particular, in the case of the 4-group crossover design, the number of participants per group is likely to be small, which leads to low powered experiments. In the case of families of experiments, the intention is to use the replications to counter the small sample sizes in individual experiments, but the small sample size per group means that the results from each experiment may be unreliable and aggregation of individual experimental results may likewise be unreliable.

We would also draw attention to one experimental design that addresses the issues that crossover designs are intended to address with fewer analysis complications. This is the pretest-posttest design with a control group which was used by one paper we studied (i.e., [44]). This design is shown in Table 5. Like a crossover, it is a repeated measures design intended to cater for skill differences, however, there is no crossover. All participants use the same technique in the first time period using the same materials. In the second time period all participants use the second set of materials. However, half the participants use the control technique, while the other half use the alternative technique (after appropriate training).

Table 5: PreTest-PostTest Design

Group	Period 1	Period 2
A	Control Technique Materials 1	Control Technique Materials 2
B	Control Technique Materials 1	Alternative Technique Materials 2

Skill differences are catered for, either by using a covariance analysis, or by analysing the difference data (i.e., the outcome for each participant in time period 2 minus the outcome for the same participant in time period 1) as an independent groups design. Furthermore if we have two new techniques, in the second period Group A can use alternative technique 1 and Group B can use the alternative technique 2. This design increases the number of participants within each group compared with a 4-group crossover, and does not introduce a possible time period by method interaction like all crossover designs [44]. This design is also particularly well-suited to covariance analysis.

9 RECOMMENDATIONS

Overall we recommend the following guidelines for human-centric SE experiments:

- Use, whenever possible, simple experimental designs which are easier to understand and may also encourage larger group sizes. Such designs are also easier for the reader to understand.
- Avoid unnecessary tests and post-hoc power analysis.
- With multiple outcome measures, if the different outcomes measure different aspects of a specific characteristic using the same unit of measurement (e.g. percentage of correct answers), use the average value. Give preference to measuring effectiveness over efficiency in formal experiments (assuming efficiency can be measured on a long ordinal scale, rather than a binary scale of correct or not correct).
- When dealing with software engineering tasks, where the skills of individual participants will significantly affect experimental outcomes, consider the pretest-posttest control design. Information concerning effect sizes and their variances for this design can be found in [40].
- If the number of participants per group is small (perhaps less than 15), and there is no *a priori* reason to assume normality, use *robust* non-parametric methods (e.g., [7, 8, 10]) without pre-testing [30].
- For moderate to large samples, use the Yuen-Welch method [59], which is based on trimmed data and assumes unequal variances, without pre-testing.
- Report more statistics about data sets at an appropriate level of granularity, usually based on raw data for each group, and, in the case of crossover designs, based on difference data for each sequence group. Include means, medians, variances, kurtosis and skewness statistics, number of outliers, and correlations between repeated measures.

ACKNOWLEDGMENTS

Lech Madeyski was partially supported by the Polish Ministry of Science and Higher Education under Wroclaw University of Science and Technology Grant 0401/0201/18.

REFERENCES

[1] S. Abrahão, C. Gravino, E. Insfran Pelozo, G Scanniello, and G. Tortora. 2013. Assessing the Effectiveness of Sequence Diagrams in the Comprehension of Functional Requirements: Results from a Family of Five Experiments. *IEEE Transactions on Software Engineering* 39, 3 (2013), 327–342.

[2] Silvia T. Acuña, Marta N. Gómez, Jo E. Hannay, Natalia Juristo, and Dietmar Pfahl. 2015. Are team personality and climate related to satisfaction and software quality? Aggregating results from a twice replicated experiment. *Information and Software Technology* 57, 1 (2015), 141–156.

[3] V.R. Basili, F.Shull, and E. Lanubile. 1999. Building knowledge through families of experiments. *IEEE Transactions on Software Engineering* 25, 4 (1999), 456–473.

[4] G.E.P. Box and K.B. Wilson. 1951. On the Experimental Attainment of Optimum Conditions (with discussion). *Journal of the Royal Statistical Society Series B* 13, 1 (1951), 1–45.

[5] George E.P. Box, J. Stuart Hunter, and William G. Hunter. 2005. *Statistics for Experimenters Design, Innovation and Discovery* (second edition ed.). Wiley-InterScience, Hoboken, NJ, USA.

[6] Edgar Brunner and Ullrich Munzel. 2000. The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal* 42, 1 (2000), 17–25. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U

[7] Edgar Brunner and Ullrich Munzel. 2000. The multivariate nonparametric Behrens–Fisher problem: Asymptotic theory and and small sample approximation. *Biometrical Journal* 42 (2000), 17 – 25. https://doi.org/10.1016/S0378-3758(02)00269-0

[8] Edgar Brunner, Ullrich Munzel, and Madan L. Puri. 2002. The multivariate nonparametric Behrens–Fisher problem. *Journal of Statistical Planning and Inference* 108, 1–2 (2002), 37 – 53. https://doi.org/10.1016/S0378-3758(02)00269-0

[9] Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafo. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14 (2013), 365–376.

[10] Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin* 114 (1993), 494—509.

[11] Jacob Cohen. 1992. A power primer. *Psychological Bulletin* 112, 1 (1992), 155–159.

[12] J.A. Cruz-Lemus, M. Genero, M.E. Manso, S. Morasca, and M. Piattini. 2009. Assessing the understandability of UML statechart diagrams with composite states—A family of empirical studies. *Empirical Software Engineering* 14, 6 (2009), 685–719. https://doi.org/10.1007/s10664-009-9106-z cited By 35.

[13] José A. Cruz-Lemus, Marcela Genero, Danilo Caivano, Silvia Abrahão, Emilio Insfrán, and José A. Carsí. 2011. Assessing the influence of stereotypes on the comprehension of UML sequence diagrams: A family of experiments. *Information and Software Technology* 53, 12 (2011), 1391–1403.

[14] David P. Darcy, Chris F. Kemerer, Sandra A. Slaughter, and James E. Tomayko. 2005. The Structural Complexity of Software: An Experimental Test. *IEEE Transactions on Software Engineering* 31, 11 (Nov. 2005), 982–995. https://doi.org/10.1109/TSE.2005.130

[15] Ben Derrick, A. Broad, D. Toher, and P. White. 2017. The impact of an extreme observation in a paired samples design. *Metodoloki Zvezki - Advances in Methodology and Statistics* 14, 2 (2017), 1–17.

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

[17] Morten W. Fagerland. 2012. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMC Medical Research Methodology* 12, 1 (2012), 1–7.

[18] Morten W. Fagerland and Leiv Sandvik. 2009. Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary Clinical Trials* 30 (2009), 490–496.

[19] Morten W. Fagerland and Leiv Sandvik. 2009. The Wilcoxon–Mann–Whitney test under scrutiny. *Statistics in Medicine* 28, 10 (2009), 1487–1497.

[20] A. Fernandez, S. Abrahão, and E. Insfran. 2013. Empirical validation of a usability inspection method for model-driven Web development. *Journal of Systems and Software* 86, 1 (2013), 161–186. https://doi.org/10.1016/j.jss.2012.07.043

[21] Ana M. Fernández-Sáez, Marcela Genero, Danilo Caivano, and Michel R. V. Chaudron. 2016. Does the level of detail of UML diagrams affect the maintainability of source code?: A family of experiments. *Empirical Software Engineering* 21, 1 (2016), 212–259.

[22] Ana M. Fernández-Sáez, Marcela Genero, Michel R. V. Chaudronand, Danilo Caivano, and Isabel Ramos. 2015. Are Forward Design or Reverse-Engineered UML diagrams more helpful for code maintenance?: A family of experiments. *Information and Software Technology* 57 (2015), 644–663.

[23] Javier Gonzalez-Huerta, Emilio Insfrán, Silvia Mara Abrahão, and Giuseppe Scanniello. 2015. Validating a model-driven software architecture evaluation and improvement method: A family of experiments. *Information and Software Technology* 57 (2015), 405–429.

[24] I. Hadar, I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, and A. Susi. 2013. Comparing the comprehensibility of requirements models expressed in Use Case and Tropos: Results from a family of experiments. *Information and Software Technology* 55, 10 (2013), 1823–1843. https://doi.org/10.1016/j.infsof.2013.05.003

[25] John P. A. Ioannidis and Claire Mokrysz. 2008. Why Most Discovered True Associations are Inflated. *Epidemiology* 19, 5 (2008), 640–648.

[26] Magne Jørgensen, Tore Dybå, Knut Liestøl, and Dag I.K. Sjøberg. 2016. Incorrect results in software engineering experiments: How to improve research practices. *The Journal of Systems and Software* 116 (2016), 133–145.

[27] H. J. Keselman, Abdul R. Othman, and Rand Wilcox. 2013. Preliminary testing for normality: Is it a good practice? *Journal of Modern Applied Statistical Methods* 12, 2 (2013), 53–65.

[28] H. J. Keselman, Abdul R. Othman, and Rand Wilcox. 2014. Testing normality in the multi-group problem: Is it a good practice? *Clinical Dermatology* 2, 1 (2014), 53–65.

[29] Barbara Kitchenham, Lech Madeyski, and Pearl Brereton. [n. d.]. Meta-analysis for Families of Experiments in Software Engineering: A Systematic Review and Reproducibility and Validity Assessment. *Empirical Software Engineering* ([n. d.]). In Review.

[30] Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong. 2017. Robust Statistical Methods for Empirical Software Engineering. *Empirical Software Engineering* 22, 2 (2017), 579–630. https://doi.org/10.1007/s10664-016-9437-5

[31] Oliver Laitenberger, Khaled El Emam, and Thomas G. Harbich. 2001. An internally replicated quasi-experimental comparison of checklist and perspective-based reading of code documents. *IEEE Transactions on Software Engineering* 27, 5 (2001), 387–418.

[32] Björn Lantz, Roy Andersson, and Peter Manfredsson. 2016. Preliminary Tests of Normality When Comparing Three Independent Samples. *Journal of Modern Applied Statistical Methods* 15, 2 (2016), 135–148.

[33] H. Levene. 1960. Robust tests for equality of variances. In *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, I. Olkin (Ed.). University Press Stanford, USA, 279–292.

[34] Lech Madeyski. 2010. *Test-Driven Development: An Empirical Evaluation of Agile Practice.* Springer, (Heidelberg, London, New York). https://doi.org/10.1007/978-3-642-04288-1

[35] Lech Madeyski and Barbara Kitchenham. 2017. Would Wider Adoption of Reproducible Research be Beneficial for Empirical Software Engineering Research? *Journal of Intelligent & Fuzzy Systems* 32 (2017), 1509–1521. https://doi.org/10.3233/JIFS-169146

[36] Lech Madeyski and Barbara Kitchenham. 2018. Effect Sizes and their Variance for AB/BA Crossover Design Studies. *Empirical Software Engineering* 23, 4 (2018), 1982–2017. https://doi.org/10.1007/s10664-017-9574-5

[37] Scott E. Maxwell, Harold D.Delany, and Ken Kelley. 2018. *Designing Experiments and Analyzing Data A model Comparison Perspective* (third edition ed.). Routledge, New York, NY, USA.

[38] Rupert G. Miller. 1997. *Beyond ANOVA: Basics of Applied Statistics.* CRC Press, Boca Raton, FL, USA.

[39] José Miguel Morales, Elena Navarro, Pedro Sánchez-Palma, and Diego Alonso. 2016. A family of experiments to evaluate the understandability of TRiStar and i* for modeling teleo-reactive systems. *Journal of Systems and Software* 114 (2016), 82–100.

[40] Scott B. Morris and Richard P. DeShon. 2002. Combining Effect Size Estimates in Meta-Analysis With Repeated Measures and Independent-Groups Designs. *Psychological Methods* 7, 1 (2002), 105–125. https://doi.org/10.1037//1082-989X.7.1.105

[41] Shinichi Nakagawa. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology* 15, 6 (2004), 1044–1045.

[42] Stephen Olejnik, Jianmin Li, Suchada Supattathum, and Carl J. Huberty. 1997. Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioural Statistics* 22 (1997), 389–406.

[43] Abdul R. Othman, H. J. Keselman, and Rand Wilcox. 2015. Assessing Normality: Applications in Multi-Group Designs. *Malaysian Journal of Mathematical Sciences* 9, 1 (2015), 53–65.

[44] Dietmar Pfahl, Oliver Laitenberger, Günther Ruhe, Jörg Dorsch, and Tatyana Krivobokova. 2004. Evaluating the learning effectiveness of using simulations in software project management education: results from a twice replicated experiment. *Information and Software Technology* 46, 2 (2004), 127–147. https://doi.org/10.1016/S0950-5849(03)00115-0

[45] Dieter Rasch and Klaus D. Kubinger Âůand Karl Moder. 2011. The two-sample t test: pre-testing its assumptions does not pay off. *Stats Papers* 52 (2011), 219–2–31.

[46] Nornadiah Mohd Razali and Yap Bee Wah. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics* 2, 1 (2011), 21–33.

[47] Toni Rietveld and Roeland van Hout. 2015. The t test and beyond: Recommendations for testing the central tendencies of two independent samples in research on speech, language and hearing pathology. *Journal of Communication Disorders* 58 (2015), 158–168.

[48] Justine Rochon, Matthias Gondan, and Meinhard Kieser. 2012. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology* 12, 81 (2012), 1471–2288.

[49] D.M. Rom. 1990. A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika),* 77 (1990), 663–666.

[50] Adrian Santos, Omar S. Gómez, and Natalia Juristo. 2018. Analyzing Families of Experiments in SE: A Systematic Mapping Study. *CoRR* abs/1805.09009 (2018), 1–18. arXiv:1805.09009 http://arxiv.org/abs/1805.09009

[51] Guiseppi Scanniello, Carmine Gravino, Marcela Genero, José A. Cruz-Lemus, and Genovetta Tortora. 2014. On the impact of UML analysis models on source-code comprehensibility and modifiability. *ACM Transactions on Software Engineering and Methodology* 23, 2, Article 13 (2014), 26 pages.

[52] Stephen Senn. 2002. *Cross-over Trials in Clinical Research* (2nd ed.). John Wiley and Sons, Ltd., Indianapolis, Indiana, USA.

[53] George W. Snedecor and William G. Cochran. 1980. *Statistical Methods.* The Iowa State University Press, Ames, Iowa, USA.

[54] John D. Storey. 2002. A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* 64, 3 (2002), 1479–498.

[55] M.A. Teruel, E. Navarro, V. López-Jaquero, F. Montero, J. Jaen, and P. González. 2012. Analyzing the understandability of Requirements Engineering languages for CSW systems: A family of experiments. *Information and Software Technology* 54, 11 (2012), 1215–1228. https://doi.org/10.1016/j.infsof.2012.06.001

[56] H. C. Thode. 2002. *Testing for normality.* Marcel Dekker, New York, NY, USA.

[57] Sira Vegas, Cecilia Apa, and Natalia Juristo. 2016. Crossover Designs in Software Engineering Experiments: Benefits and Perils. *IEEE Transactions on Software Engineering* 42, 2 (2016), 120–135. https://doi.org/10.1109/TSE.2015.2467378

[58] B. L. Welch. 1938. The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika* 29, 3-4 (1938), 350–362. https://doi.org/10.1093/biomet/29.3-4.350

[59] Rand R. Wilcox. 2012. *Introduction to Robust Estimation & Hypothesis Testing* (3rd edition ed.). Elsevier, Amsterdam, The Netherlands.
[60] Donald W. Zimmerman. 2003. A Warning about the Large-Sample Wilcoxon-Mann-Whitney Test. *Understanding Statistics* 2, 4 (2003), 267–280.