




Recommendations for analysing and meta-analysing small sample size software engineering experiments

Barbara Kitchenham¹ · Lech Madeyski² 

Accepted: 24 May 2024
© The Author(s) 2024

Abstract

Context Software engineering (SE) experiments often have small sample sizes. This can result in data sets with non-normal characteristics, which poses problems as standard parametric meta-analysis, using the standardized mean difference (*StdMD*) effect size, assumes normally distributed sample data. Small sample sizes and non-normal data set characteristics can also lead to unreliable estimates of parametric effect sizes. Meta-analysis is even more complicated if experiments use complex experimental designs, such as two-group and four-group cross-over designs, which are popular in SE experiments.

Objective Our objective was to develop a validated and robust meta-analysis method that can help to address the problems of small sample sizes and complex experimental designs without relying upon data samples being normally distributed.

Method To illustrate the challenges, we used real SE data sets. We built upon previous research and developed a robust meta-analysis method able to deal with challenges typical for SE experiments. We validated our method via simulations comparing *StdMD* with two robust alternatives: the probability of superiority (\hat{p}) and Cliff's d .

Results We confirmed that many SE data sets are small and that small experiments run the risk of exhibiting non-normal properties, which can cause problems for analysing families of experiments. For simulations of individual experiments and meta-analyses of families of experiments, \hat{p} and Cliff's d consistently outperformed *StdMD* in terms of negligible small sample bias. They also had better power for log-normal and Laplace samples, although lower power for normal and gamma samples. Tests based on \hat{p} always had better or equal power than tests based on Cliff's d , and across all but one simulation condition, \hat{p} Type 1 error rates were less biased.

Conclusions Using \hat{p} is a low-risk option for analysing and meta-analysing data from small sample-size SE randomized experiments. Parametric methods are only preferable if you have prior knowledge of the data distribution.

Keywords Meta-analysis · Effect size · Non-parametric · Probability of superiority · Small sample sizes · Reproducible research

Communicated by: Carlo A. Furia

✉ Lech Madeyski
lech.madeyski@pwr.edu.pl

¹ School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK

² Wrocław University of Science and Technology, Wyb. Wyspiańskiego 27, Wrocław 50370, Poland

1 Introduction

This article arose from our goal to develop reliable analysis guidelines to allow meta-analysis of families of software engineering (SE) randomized experiments (Basili et al. 1999). Although sometimes criticized for lack of realism, randomized experiments are important because they allow researchers to test causal hypotheses.

The standard methods for meta-analysis of randomized experiments are based either on two-group between-participants experiments or on single-group before-after repeated measures experiments, and are well understood (see, e.g., (Borenstein et al. 2009)). However, software engineering experiments have characteristics that make meta-analysis far more difficult:

- **Complex statistical designs.** Vegas et al. (2016) pointed out that repeated measures crossover-style experiments were extremely popular for organizing software engineering experiments. Subsequently, Santos et al. (2020) confirmed that many families of experiments included cross-over experiments. One of the major advantages of cross-over experiments for small sample-size experiments is that they have higher power than between-group experiments of the same size. However, in the wider statistical literature, we found *only one* paper that considered including repeated measures cross-over style experiments in meta-analysis (Curtin et al. 2002). Furthermore, it only considered the AB/BA cross-over, not the more complex four-group variant common in SE experiments (Kitchenham et al. 2020a, 2022). In addition, we found and corrected existing errors in the published formulas for the variance of the effect sizes (see Kitchenham et al. (2018) and Madeyski and Kitchenham (2018)). However, the estimates obtained by using our revised formulas might themselves be unreliable when calculated from *small* samples, so our corrections to the mathematical formulas cannot address all the meta-analysis problems observed in SE experiments that use cross-over experiments to improve the power of small sample-size experiments.
- **Small sample sizes.** Software engineering experiments (particularly those with human participants) are often criticized for using sample sizes that are too small to give reliable results, and meta-analysis is recommended as a means of addressing the problem (e.g., Shepperd 2018; Jørgensen et al. 2016). However, for meta-analysis to be valid, we need effect sizes and methods to aggregate those effect sizes that are reliable. Unfortunately, small samples are often unrepresentative of the distributions from which they arise. For example, in a recent study of the use of correlations in repeated measures experiments, our simulation studies showed major differences between the estimated variance of different small data sets sampled from the *same* distribution (see Kitchenham et al. 2022). Thus, small data sets are likely to produce unreliable estimates of sample properties from which the standardized mean difference effect sizes are constructed.
- **Data inconsistent with the normal (Gaussian) distribution.** In a study of the methods used to meta-analyse families of SE experiments, we identified several meta-analysis problems (Kitchenham et al. 2020a). A particular problem was that some researchers reported that individual experiments in a family failed normality tests. Still, because of the lack of any alternative meta-analysis method, they used a standard parametric approach to meta-analyse the data from such families (violating the normality assumption).

In order to address these three meta-analysis issues, we concluded that it was necessary to develop a well-defined, validated process for analysis and meta-analysis of SE experiments that properly addressed the problems that arise when analysing families of small sample-size SE experiments.

This article reports our proposed analysis method and the results of our validation of the method. The analysis method was based on three critical concepts that are needed to support robust meta-analysis:

1. **Identifying a Robust Effect Size and a Non-Parametric Analysis Method.** To perform a meta-analysis, it is necessary to select an appropriate effect size to measure each experiment's outcome. To analyse small data sets from unknown distributions, we need both effect sizes that are resilient to anomalous values that can occur as a result of small samples or non-normal distributions, and non-parametric analysis methods that do not require a normal distribution (or any other specific distribution) to calculate them.
2. **Addressing Complex Statistical Designs.** Non-parametric methods are usually restricted to fairly simple experimental designs, but to support SE researchers using cross-over designs to increase the power of small sample-size experiments, we need non-parametric analysis methods that can be applied to such designs.
3. **Specifying the Meta-Analysis Process.** It is necessary to adopt a robust meta-analysis process that can be used to aggregate the chosen effect size.

The statistical analysis proposals presented in this paper are based mainly on three existing statistical analysis proposals that provided methods of addressing the analysis problems facing researchers restricted to using small sample sizes in the context of randomized experiments. Firstly, Kromrey et al. (2005) proposed using the non-parametric effect size Cliff's d for meta-analysis. In this study, we investigated both Cliff's d and the related probability of superiority (referred to as \hat{p}), which are ranked-based effect sizes. Secondly, Brunner and Munzel (2000) and Brunner et al. (2002) addressed the problems associated with variance heterogeneity in rank-based effect sizes by developing test procedures based on Welch's method (Welch 1938) that provide reliable statistical tests after the data is transformed to average ranks. Thirdly, Senn (2002) pointed out that non-parametric methods could be used to analyse a crossover-style experimental design by analysing the *differences* between repeated measures.

However, to develop viable meta-analysis proposals for non-parametric effect sizes, it is essential to *validate* them. Simulation studies are the standard method used to validate statistical analysis proposals (Ripley 2006; García et al. 2010), and it is necessary to validate each element of our proposed method. For our validation studies, we compared the impact of analysing data using our proposal with the usual parametric analysis method based on the standardized mean difference effect size ($StdMD$). We sampled data from four different distributions in order to compare the analyses on normal and non-normal samples. We used normally distributed samples to provide baselines against which to compare the effectiveness of the non-parametric effect sizes and meta-analysis method. To assess the extent to which our methods were robust to non-normal data, we investigated datasets from three distributions with various non-normal characteristics: the Lognormal distribution that produces strongly skewed samples, the Gamma distribution that produces moderately skewed samples, and the Laplace distribution that produces samples that are symmetric but have longer tails, and hence, more outliers than normal samples. Our simulation studies were designed to investigate the power, bias, estimate error, and Type 1 error rates of our analysis method for the different distributions and experimental designs.

In Section 2, we provide an overview of the characteristics of randomized experiments for readers unfamiliar with issues involved in running randomized experiments. Then, we analyse some existing families of randomized experiments that have made their experimental data publicly available, in order to give readers some idea of the range of sample sizes found in SE experiments and the frequency of data sets exhibiting non-normal characteristics. Sample

sizes reported in this section influenced our choice of sample sizes in our simulation studies. Subsequently, we organize our study to introduce our analysis proposals and then validate the different elements of the proposals.

We summarize our effect size proposals in Section 3 and provide a more detailed explanation in our Supplementary Material (Kitchenham and Madeyski 2023). Then, in Section 4, we present simulation results based on samples generated from four different distributions¹ that confirmed the value of non-parametric effect sizes for individual experiments with small sample sizes.

In Section 5, we report our simulations of meta-analysing families of two-group and four-group experiments and confirm the value of non-parametric effect sizes for meta-analysis of small sample size experiments. This was made difficult because there are no well-defined guidelines for using *StdMD* with small samples, and factors such as sample size and experimental design all result in different meta-analysis methods. We illustrate the problem with a small example in Section 5.1 and justify the method we adopted for our meta-analysis simulations.

In Section 6, we summarize our results, identify the limitations of our simulation studies, and present our conclusions.

To assist readers who would like to adopt the use of our analysis methods, we provide a *reproducer* package (Madeyski et al. 2023), written in R, that can be used to reproduce the analyses reported in this paper. Wilcox has provided implementations (in R) of the methods to calculate Cliff's d and \hat{p} , as well as their variances (Wilcox 2012). However, we have amended Wilcox's algorithms to provide consistent estimates of Cliff's d and \hat{p} , as well as an alternative approach to handling extreme values of the effect sizes that lead to estimates of the effect size variance being zero. This is discussed in the Supplementary Material (Kitchenham and Madeyski 2023). Our long-term goal is to promote the reproducibility of research in software engineering (Madeyski and Kitchenham 2017) by supporting our research papers with algorithms and data sets published in the *reproducer* R package (see Kitchenham et al. 2017; Madeyski and Kitchenham 2018; Jureczko and Madeyski 2015; Madeyski and Jureczko 2015).

2 Properties of Data Sets Obtained from Software Engineering Experiments

In this section, we describe the design of formal SE experiments which our analysis proposals are intended to address. Then we use the results of two of our previous investigations of SE experiments to demonstrate the existence of unreliable parameter estimates, small sample sizes, and non-normal residuals. These issues confirm the need for, and the potential value of, robust effect sizes and non-parametric meta-analysis for SE experiments with small sample sizes.

2.1 Characteristics of Randomized Experiments

The goal of a randomized experiment is to formally test causal hypotheses that compare two or more different treatments which address the same condition. The term *treatment* is

¹ We provide a short tutorial on the properties of the four distributions used in our simulations in the Supplementary Material (Kitchenham and Madeyski 2023).

adopted from the health care field, but applies to any method that can be used to achieve a specific objective.

In SE research, randomized experiments aim to investigate different techniques used to perform the same SE task. For example, the most famous series of experiments in SE compared the use of perspective-based code reading for error detection either with checklist-based code reading or with ad-hoc code reading (for a summary of papers investigating this issue, see Ciolkowski (2009)).

SE experiments are usually designed to investigate which of two (or sometimes more) techniques is either most likely to be associated with a correct task outcome, or most likely to ensure a task is completed as quickly as possible (or with the least effort). Thus, in most cases, we require outcome measures related to task output effectiveness and/or task efficiency.

Any randomized experiment is comprised of a number of *trials*, where, in the SE context, each trial involves a human subject (or sometimes a team) performing a SE task using one of the techniques under investigation. In order to provide an answer to the question of whether one technique is likely to be better than another technique with respect to effectiveness or efficiency, we need to perform multiple trials of each technique *under the same conditions*. In addition, the trials need to be organized into a valid experimental design.

For a valid experiment, the variation between trials needs to be controlled so that the only difference between trials is due to *random* variation between the individual participants (or teams) in the experiment and the technique being used. In SE, we expect skill differences between participants to impact the effectiveness and efficiency of task performance. This is addressed in three ways:

1. Random assignment of participants to each technique. Random assignment is usually constrained to ensure even numbers of participants per technique for optimal statistical tests. It leads to the simplest statistical design, which statisticians refer to simply as a randomized experiment. It is also referred to as a between-groups experiment, a between-subjects experiment or an A/B experiment. It is effective when sample sizes are large enough to ensure that random allocation will (with a high probability) ensure participant skill differences are spread evenly between each treatment group.
2. Blocking into low- and high-skill groups. If participants can be separated into two or more groups (preferably of the same size), referred to as blocks, on the basis of skill levels, participants in each block can then be assigned randomly to equal-sized groups for each technique. The statistical analysis is more complicated because it must address the impact of the blocking process on the outcome measures. Statisticians refer to this design as a randomized block design. It is sometimes mistakenly referred to as a factorial experiment in SE (see the comment in Kitchenham et al. 2019)
3. Repeated measures designs which measure the results of human participants performing tasks using all the techniques being investigated. This allows the participants to be observed under both treatment conditions and to act as their own control. Formally, such designs make experiments more powerful (i.e., the experiment is more likely to correctly reject the null hypothesis). However, it requires more work from the participants, who have to learn two techniques and undertake two trials, and more work for the experimenters since they need to prepare more experimental materials, which need to be as similar as possible for both tasks (e.g., instead of needing one piece of faulty code, experimenters need two pieces of faulty code and the difficulty of the two programs and their embedded faults need to be as similar as possible), and they need to undertake more complex statistical analysis. There are two main forms of repeated measure design. The simplest design is a within-subject before/after design where all participants perform a task using one

technique and then perform a similar task using the other technique. However, in SE, researchers usually use a more complex form of design called a crossover design. In crossover designs, participants are randomized into two or more *sequence groups*. In the simplest crossover design (which statisticians refer to as an A/B crossover), participants in one sequence group perform a task using technique A first, then later perform a task using technique B. In contrast, participants in the other sequence group use technique B first, then technique A. However, SE researchers have frequently adopted an even more complicated design involving 4-sequence groups for their experimental designs (see Vegas et al. 2016).

All other conditions in an experiment need to be as similar as possible. This means that experimental materials are the same or similar so that the SE tasks are essentially the same for each technique (e.g., the same faulty code module in the case of a code reading experiment or a specification of the same phenomenon in the case of an experiment involving the understandability of different notations) This is obviously more difficult in the case of repeated measures designs.

In addition, it is important that experimenters treat participants the same, irrespective of their assigned treatment. Two important issues in the context of SE is to ensure both that experimenters are not biased towards one of the techniques, and that participant training does not favour one technique more than another.

Finally, the conduct of the experiment must be controlled to ensure that all participants perform their tasks using the allowed materials and methods, all have the same amount of time for their tasks, and that there is no interaction between individual participants (or individual teams). Given the planned duration of the experimental tasks, the tasks need to be designed not to be too difficult for any of the participants to complete, and not to be too simple so that all the participants complete them easily. It is usual for academic researchers to run experiments rather like examinations, so that all participants undertake their tasks at the same time and can be discouraged from copying.

The statistical and meta-analysis methods proposed in this article are intended to support all forms of random experiment discussed in this section, except factorial experiments. They are intended for use when researchers have small sample sizes and unknown distributions.

2.2 Previous Analyses of Experimental Data

The data sets used in this section were obtained from two studies:

1. Kitchenham et al. (2020a), performed a systematic review of methods used to meta-analysis families of experiments. We selected papers that meta-analysed parametric effect sizes and identified 13 primary studies, each of which included between 3 and 5 randomized experiments. The set of studies we used in our experiment overlaps 12 of the 15 papers (Santos et al. 2020) classified as using the aggregated data (AD) technique. We excluded two of the papers found by Santos et al. because they were not published in the five SE journals we selected as having a relatively high impact and excluded another one because it meta-analysed results of correlation studies, not experiments (i.e., Acuña et al. 2015). We also found one paper that Santos et al. missed (i.e., Morales et al. 2016). For more details about the search and selection process (see Kitchenham et al. 2020a). We studied the experiments reported in these papers in some detail and identified them as the type of experiments that the analysis and meta-analysis methods proposed in this article are intended to address. One of the papers reported the results of three team-based

experiments, but the other experiments all reported experiments that involved individual human participants.

2. Kitchenham et al. (2022) investigated the correlation between repeated measures found in cross-over experiments. We did this by re-analyzing the raw data from cross-over experiments. As Santos et al. (2020) reported, most families of experiments did not provide access to their raw data. However, with our collaborators (Scanniello and Gravino), who organized many families of experiments, we obtained raw data from 15 studies (11 of which reported families of experiments). Two of the studies analysed team results and four of the 15 studies overlapped with the data sets used in our previous study (Kitchenham et al. 2020a). These data sets are discussed, in detail, in the Supplementary Material (Kitchenham et al. 2020b) to the paper by Kitchenham et al. (2022) and the raw data are available for the experiments reported in 13 of the studies in the R `reproducer` package (Madeyski et al. 2023).

2.3 Unreliable Variance Estimates

In our study of correlation in crossover experiments (Kitchenham et al. 2022), which was based on both simulation studies and on analysis of software engineering datasets. We analysed data sets from 15 different software engineering papers reporting 36 different software engineering experiments and 69 output metrics.

Our simulation studies confirmed that estimates of sample variances from small samples (e.g., samples of 30 or less) based on normal distributions with equal variances can be very inaccurate and often exhibit large heterogeneity.

Our analysis of the SE data sets showed some large differences between variance estimates from different groups in the same experiment. That is, if we constructed the variance of data from one sequence group in a crossover experiment and the variance of the another sequence group, in the same experiment, they could be very different for small data sets. Thus, with small data sets, we are likely to find within-group variance estimates that are very non-homogeneous, and we cannot tell whether this is due to small data sizes or genuine variance heterogeneity attributable to the experimental conditions. Whatever the reason, standard analysis of variance tools always assumes variance homogeneity for anything more complex than a simple between-two-group experimental design. The implication is that our variance estimates for experiments with small sample sizes are untrustworthy in the sense that we cannot be sure of their accuracy. However, we need reliable variance estimates to construct trustworthy parametric effect sizes, such as the standardized mean difference (*StdMD*) or the point biserial correlation coefficient.

2.4 Sample Sizes in SE Families of Experiments

Although Santos et al. noted that sample sizes for families of experiments were relatively small in their mapping study of 39 families of experiments, we wanted to be sure about the range of sample sizes that had been used by SE researchers for randomized experiments of the type addressed by the methods proposed in this article. To do this, we identified all individual studies and experiments in Kitchenham et al. (2022) and Kitchenham et al. (2020a), excluding those involving team results rather than results obtained from individuals. In Kitchenham et al. (2022), 13 of the 15 papers reported data from 32 experiments that analysed individual participant data. In Kitchenham et al. (2019), we analysed 13 papers

that reported a meta-analysis of families of experiments, of which nine were *not* included in the set of papers discussed in Kitchenham et al. (2022). These nine papers reported 31 experiments that analysed individual participant data. Thus, from a total of 22 papers, we had sample size information on 63 independent experiments. We present a histogram of the sample size data in Fig. 1.

Of the 63 experiments, 53 (i.e., 84%) had 40 or fewer participants. The smallest experiment had 9 participants, and the largest had 178. The median number of participants was 24. This is very close to the value of 23.5 we found for the median of the experiments reports by Santos et al. (2020), excluding the overlapping papers we included in our analysis. These sample sizes confirm that SE randomized experiments usually have small sample sizes. This analysis motivated our choice of sample sizes in the simulation studies reported in this paper.

2.5 Frequency of Non-Normality in SE Experiments

To assess the prevalence of non-normal data, we re-analysed data from 13 of the 15 studies used in Kitchenham et al. (2022), omitting the two studies that analysed data at a team level. The remaining 13 studies reported 32 experiments that used either the standard two-group AB/BA crossover studies or the four-group duplicated crossover design, where duplication was based on the order in which the participants received the software engineering materials needed to perform the required software engineering tasks (Madeyski and Kitchenham 2018). Many of the 32 experiments collected several different metrics from each participant, leading to a total of 64 different datasets.

In order to assess the normality of the data, we analysed the data for each metric in each experiment and investigated the distribution of the residuals. We have argued against *preanalysis* normality testing elsewhere (Kitchenham et al. 2019). In particular, a specific issue in the context of complex experimental designs, such as the crossover designs, is that the experimental design itself can introduce differences between partitions of the data that can make the raw data appear non-normal. However, there is no objection to analysing the distribution of residuals because the systematic differences between experimental conditions are removed by statistical analysis. The experimental designs used by the 32 experiments

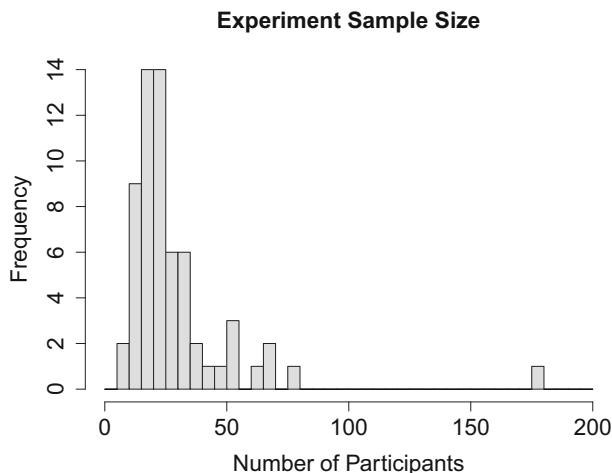


Fig. 1 The number of participants per experiment (bin size = 5)

were all crossover designs, including both two-group and four-group cross-over designs. We discuss the nature of these designs in Kitchenham et al. (2022).

We analysed the AB/BA crossover experiments with the R package `lme4` using the `lmer` formula as shown in Fig. 2. This analysis follows Senn's recommendations for modelling and analysing AB/BA crossover designs by treating Time Period as a simple blocking factor and any interaction between Time Period and the treatment factor as negligible (Senn 2002).

$$\text{Metric} \sim \text{TP} + \text{Treat} + (1|\text{ParticipantID})$$

Fig. 2 `lmer` model function for AB/BA crossover experiments

where:

<code>Metric</code>	identifies the outcome,
<code>TP</code>	is a fixed effect parameter that identifies the time period in which the value was obtained,
<code>Treat</code>	is a fixed effect parameter that identifies the treatment condition that was used to obtain the outcome value,
<code>(1 ParticipantID)</code>	identifies the participant providing the output data and confirms that participant values are treated as random variables.

For the four-group duplicated crossover, described in Kitchenham et al. (2020b, Section 3), we introduced two new blocking factors, as shown in Fig. 3, specifically:

<code>COID</code>	(i.e., <code>CrossOverID</code>) to identify which group of participants belonged to the same AB/BA crossover, and
<code>System</code>	to specify which software system materials (i.e., code or documents) were used in each experimental condition.

$$\text{Metric} \sim \text{TP} + \text{Treat} + \text{System} + \text{COID} + (1|\text{ParticipantID})$$

Fig. 3 `lmer` model function for four-group crossover experiments

We treat `System` and `COID` (which identifies which groups are matched together) as simple blocking factors and assume interactions among such factors are negligible.

After analysing data for each metric, we analysed the distribution of the standardized residuals. Specifically, we calculated the mean, median, variance, skewness, and kurtosis values, identified the number of outliers for each output metric using the R `boxplot` function, and tested the residuals for normality using the Anderson-Darling (AD) test, which we found was preferred to the Shapiro-Wilk test by several recent studies (see Kitchenham et al. 2019).

Table 1 Data sets with non-normal properties

Study	Exp	Metrics	N	AD P-Val	Number of outliers
S1	USB2	Time	24	0.045	1
S4	PoliTo2	Comprehension	17	0.013	0
S4	UniGe	Comprehension	66	0.002	2
S6	EUBAS	Comprehension	24	0.012	5
S6	R1UGOT1	Time	63	0.000	12
S8	UniBas2	Comprehension	31	0.040	0
S8	UniBas1	Efficiency	33	0.049	3
S8	UniBas2	Efficiency	31	0.048	1
S9	UniBZ	Comprehension	26	0.009	1
S9	UniBZ	Time	26	0.008	5
S10	P2007	NATPPH	22	0.000	5
S12	PROF	Efficiency	16	0.007	2
S12	UNIBAS	Efficiency	49	0.000	6
S12	UNINA	Efficiency	19	0.003	4
S13	EXP1	Fc	55	0.020	1
S14	CSI2010	Effectiveness	32	0.012	2

2.6 Analysis Results and Implications

At the $p = 0.05$ level, residuals from 16 (of 64) data sets (i.e., 25%) failed the normality test (see details in Table 1). The expected false positive rate given 64 data sets is 3 (with a 95% upper bound of 6), which suggests that the proportion of data sets with non-normal properties is excessive, even if the three experiments with p -values only just below 0.05 are ignored. Five of the 64 data sets exhibited more than five outliers (based on a boxplot of the residuals), and *all* of those five data sets also failed the normality test.

It is important to note that:

1. Data sets with non-normal residuals were found in nine of 13 families (i.e., in more than 69% of families of experiments).
2. In only two cases were data with non-normal residuals from the same experiment.
3. Among the data sets with non-normal properties are all eight experiments with sample sizes larger than 30. Larger datasets are usually more trustworthy than small ones, so it is possible that the smaller samples could have included false negatives as a result of the lack of power usually associated with normality tests (Kitchenham et al. 2019).

Hence, not only did we find a much larger number of failed tests than would be expected, but these were not restricted to specific families, specific metrics, or specific experiments. Thus, our results confirm the following:

- SE experiments are often relatively small.
- Families of experiments are quite likely to include at least one experiment with residuals that will fail tests of normality².

² We do not claim that the data sets are samples from non-normal distributions, only that the characteristics of the data sets are likely to cast doubts on the validity of analyses that rely on normally distributed data.

3 Non-parametric Methods and Robust Effect Sizes

Robust effect sizes aim to provide summary sample statistics that are less influenced by outliers than the usual mean and variance. As discussed by Derrick et al. (2017), outliers are sample values that differ substantially from other values in a sample. Outliers increase the variability of a sample and can result in unreliable estimates of the mean and variance. Outliers can occur if the data is inherently non-normal, but can also occur by chance in normal samples, and are most likely to occur in small samples.

Early suggestions for robust effect sizes were based on finding robust measures of central location and dispersion that could be used to calculate robust equivalents of standardized mean difference effect sizes. For example, Kraemer and Andrews (1982) developed an effect size catering for pre-test and post-test studies where the post-test had both a treatment and a control group. The effect size was based on the proportion of observations in the treatment group that were greater than the median of the observations in the control group.

Hedges and Olkin (1983) extended this work by investigating a number of different experimental designs. Their study makes it clear that effect sizes depend on experimental design.

Recent research has been more influenced by rank-based robust statistics (also known as order statistics). In particular, many researchers have considered effect sizes based on the probability that a random observation obtained from one group is greater than a random observation from another group. This has been given a variety of different names in different disciplines, such as the Common Language effect size (McGraw and Wong 1992), the *A* measure of Stochastic Superiority (Varga and Delany 2000; Arcuri and Briand 2014; Madeyski et al. 2014), the Probability Index (Acion et al. 2006), Probability of Benefit (Faraone 2008), and the Mann-Whitney probability of superiority (Rahlf's et al. 2013), and which we refer to simply as the probability of superiority \hat{p} .

Cliff's *d* (Cliff 1993) is another non-parametric statistic closely related to the probability of superiority. This is also referred to as the Mann-Whitney difference by Rahlf's et al. (2013).

In this study, we investigate the probability of superiority (which we refer to as \hat{p}) and Cliff's *d*, both of which we have advocated as non-parametric effect sizes in a previous study (Kitchenham et al. 2017). We have concentrated on these two non-parametric effect sizes because prior research suggests they support our analysis requirements:

1. Kromrey et al. (2005) have proposed using Cliff's *d* as an alternative to the standardized mean difference *StdMD*. They reported that unweighted Cliff's *d* had lower bias than weighted Cliff's *d*, Cohen's δ or Hedges *g* under all experimental conditions. Even under conditions of severe variance heterogeneity together with a large population effect size, Cliff's *d* exhibited only minimum bias.
2. Brunner et al. have confirmed that \hat{p} can be used in the context of different experimental designs, including two-way designs and randomized block designs (Brunner and Munzel 2000; Brunner et al. 2002; Wilcox 2012). The method Brunner et al. used for statistical tests of \hat{p} is designed to cater for variance heterogeneity, thus improving the reliability of statistical tests of hypothesis.

Other types of robust statistics, such as trimmed means, do not have the benefit of previous existing research to clarify whether they could be used to construct valid effect sizes, nor how they could be used in the context of meta-analysis.

We show below that Cliff's *d* and \hat{p} effect sizes are functionally related, which means they should behave very similarly. However, \hat{p} is defined as a probability and has the expected range of values between 0 and 1, which is slightly easier to understand than Cliff's *d*, which is the difference between two probabilities and has values between -1 and 1. We decided to

investigate both effect sizes because, if there was no practical difference between the two effect sizes, researchers could use the effect size they felt most comfortable with.

3.1 The Meaning and Derivation of Cliff's d and \hat{p}

The effect sizes \hat{p} , and Cliff's d can both be derived from estimates of the probabilities that observations from one group are greater than (p_1), less than (p_3) or equal to (p_2) observations in another group, where $p_1 + p_2 + p_3 = 1$.

For \hat{p} :

$$\hat{p}_{X>Y} = p_1 + \frac{p_2}{2} \quad (1)$$

and

$$\hat{p}_{X<Y} = p_3 + \frac{p_2}{2} \quad (2)$$

The values of \hat{p} vary from 0 to 1, with values close to 0.5 suggesting that there is no significant difference between the groups. Formal statistical tests are based on analysing the average ranks of data using a method that allows for variance heterogeneity (Brunner and Munzel 2000). Values of $\hat{p}_{X>Y}$ significantly greater than 0.5 imply that condition X has increased values of the outcome variable. In contrast, values significantly less than 0.5 suggest condition X has decreased the values of the outcome variable.

For Cliff's d :

$$d_{X>Y} = p_1 - p_3 \quad (3)$$

Cliff's d values range from -1 to 1, with a value close to zero suggesting that there is no significant difference between the groups. Formal statistical tests are based on identifying whether the confidence interval for d includes zero (Long and Cliff 1997). It is clear from (1) and (3) that there is a functional relationship between the effect sizes.

Values of $d_{X>Y}$ significantly greater than 0 imply that condition X has increased values of the outcome variable. In contrast, values significantly less than 0 suggest condition X has decreased the values of the outcome variable.

We provide a more detailed discussion of these effect sizes and the formulas we use to calculate their variance in Section 2 of our Supplementary Material (Kitchenham and Madeyski 2023). Section 2 also explains how the effect sizes and their variances can be derived from the *superiority matrix*, which defines the relationship between the values in each group and is the basis of the algorithms Wilcox developed to calculate and test Cliff's d and \hat{p} (see <https://dornsife.usc.edu/rwilcox/>).

We also provide algorithms to calculate these effect sizes in our reproducer R package (see `reproducer::Cliffd.test` and `reproducer::PHat.test` functions). Assuming data from a two-group experiment, the algorithms deliver estimates of the relevant effect size and its variance. The algorithms also perform statistical tests of the effect sizes (either two-sided or one-sided) and calculate the 95% confidence intervals (see Section 8.3.3 of the Supplementary Materials).

In the statistical literature, \hat{p} and Cliff's d are not usually subscripted, but it is important to note that they have a direction that needs to be respected, particularly if you are comparing values from different experiments.

3.2 Extracting Robust Effect Sizes from Different Experimental Designs

Although we have described Cliff's d and \hat{p} in terms of probabilities, they are also closely related to rank order statistics and, in particular, are functionally related to non-parametric tests such as the Mann-Whitney test. However, the Mann-Whitney test is inappropriate for complex experimental designs because the variance of rank averages is heteroscedastic if there is a significant difference between groups. Brunner and his colleagues (Brunner and Munzel 2000; Brunner et al. 2002) developed an analysis method for \hat{p} that avoids variance inequality problems by adapting Welch's method that allows for unequal variances in t -tests to rank-transformed data. Their research is discussed by Wilcox (2012), who also provides implementations of their analysis algorithms. As discussed below and by Wilcox, Brunner et al.'s method supports most standard experimental designs and, in particular, randomized block experiments and cross-over experiments.

3.2.1 Calculating Non-parametric Effect Sizes for Randomized Block Designs

Randomized block designs are used to increase the generality of results and/or to reduce spurious variability. For example, in SE experiments, participants usually perform software engineering tasks using some specific software engineering materials. To avoid comparing SE methods using only a single application (e.g., program, set of modules, or documentation), we might want to use documents related to several different applications to increase the generality of conclusions we can draw from our experiment. Alternatively, we might be concerned that a technique is very dependent on the skill of its user, and we might want to assign our participants to different skill groups with the aim of controlling the variability among individuals. Organizing participants into groups with similar skills or using similar SE documents is intended to control spurious variability and is the basis of the randomized block designs. Once participants are separated into similar blocks, they are randomly assigned to the different experimental treatments.

Randomized block experiments are always analysed using a within-block analysis. If we design an experiment with k blocks (where $k \geq 2$) and two treatments³, we can calculate the value of the effect size for each block and then calculate the effect size of the experiment as a whole as the average of the effect size for each block:

$$\overline{NPES} = \frac{\sum_{i=1}^k NPES_i}{k} \quad (4)$$

$NPES_i$ estimates a specific NP effect size for group i . Thus, we have treated the randomized block experiment as a group of k independent two-group randomized experiments and have isolated the treatment difference from the block effect by comparing the individual treatment differences under the same blocking condition.

Based on the following two standard statistical results for independent variables x and y and a constant c :

$$var(x + y) = var(x) + var(y)$$

$$var(cx) = c^2 var(x)$$

³ Throughout this paper, we assume that there are only two treatment options; meta-analysis is not well-defined for experiments with multiple treatments.

we can use the variance of the NP effect size calculated in each block to estimate the variance of \overline{NPES} :

$$\text{var}(\overline{NPES}) = \frac{\sum_{i=1}^k [\text{var}(NPES_i)]}{k^2} \quad (5)$$

Since there is no restriction on the number of blocks in a randomized blocks design, (4) and (5) mean that we can extract the overall estimate of \hat{p} , Cliff's d and their respective variances for any experimental design that can be broken down into a set of two-group randomized experiments.

We provide an algorithm in our `reproducer` R package (Madeyski et al. 2023) to analyse randomized block experiments comprising two treatment conditions and two blocking conditions (see `reproducer::Calc4GroupNPStats`).

3.2.2 Calculating Non-parametric Effect Sizes for Cross-over Designs

Although (Senn 2002) was mainly interested in parametric analysis, he did point out that non-parametric analysis can be based on applying the Mann-Whitney test to the period difference values.

In our SupplementaryMaterial (Kitchenham and Madeyski 2023), we explain how Senn's suggestion can be applied to the AB/BA cross-over design. Specifically, the time period *difference values* are equivalent to data from a two-group randomized experiment (between groups). In addition, the time period difference values from a four-group cross-over design are equivalent to data from a randomized block experiment with two blocks and two treatments. An important issue is that the difference values within the same block represent effect sizes in opposite directions. For example, in the case of a two-block A/B cross-over, assuming the participants in block A use treatment T1 in period 1 and treatment T2 in period 2, and the participant difference values (i.e. the value of each participant in period 1 subtracted from the participant's value in period 2), the values of participants ($x_i S$), are modelled as:

$$x_i = \mu_i + t_2 - (p + \mu_i + t_1) = p + t_1 - t_2 \quad (6)$$

where t_1 is the change in outcome caused by using T1 and t_2 is the change in outcome caused by using T2, p is the period effect, which is assumed to be the same for both groups, μ_i is the hypothetical overall mean of participant i performing the specific SE task. The effect of using difference values is to remove the individual participants' effect.

The difference values for participants (y_i) in group B that use treatment T2 in period 1 are modelled as :

$$y_i = \mu_i + t_2 - (p + \mu_i + t_1) = p + t_2 - t_1 \quad (7)$$

The effect of using difference values is to remove the effect due to individual participants and to leave a period effect which is the same for both groups, and to have difference values that (ignoring the common period effect) are modelled by $t_1 - t_2$ in one group and $-(t_1 - t_2)$ in the other. Thus, any difference between T1 and T2, will be strongly emphasized. In practice, it is likely that the estimates of \hat{p} and Cliff's d obtained from crossover designs will be systematically larger than estimates from between-groups designs. This is exactly what happens with the standardized mean difference effect sizes (Madeyski and Kitchenham 2018). We consider methods of meta-analysis suitable for robust effect sizes from different types of experimental design in more detail in Section 3.4.3.

3.2.3 Limitations

Senn's approach to non-parametric analysis of crossover designs assumes that outcome variables are ratio-scale numbers. Short ordinal-scale outcomes or binary outcomes would not be suitable.

Brunner's analysis method allows us to calculate \hat{p} for any experiments that can be decomposed into independent two-group experiments, including two-group and four-group cross-over models. However, we do not claim that the method can be used to meta-analyse any experimental design.

In particular, the analysis method does not directly support meta-analysis of genuine factorial experiments, where researchers are investigating the joint impact of two different treatments (for example, the impact of using both design inspections and code inspections). However, it should be noted that currently, there is no well-defined meta-analysis method for factorial experiments.

3.3 Meta-Analysis of Non-parametric Effect Sizes

Meta-analysis is a means of obtaining a *summary of a set effect sizes* from a series of independent randomized experiments. The summary is usually a weighted average. Effect sizes suitable for meta-analysis of randomized SE experiments have several characteristics:

1. They provide a measure of the difference between the techniques being compared.
2. They are usually unit-free. Although the mean difference between groups is an effect size, it is seldom used for meta-analysis in SE because it presupposes that the value being measured has some objective interpretation scale.
3. They are not functions of sample size like test statistics. This is because test statistics, such as a t -test values, can be influenced by changing the sample size without changing the difference between the treatment groups.

Meta-analysis is usually applied to parametric effect sizes, particularly the standardized mean difference or the point bi-serial correlation coefficient, but Cliff's d and \hat{p} both conform with the three criteria reported above. In addition, Kromrey et al. (2005) proposed applying meta-analysis to Cliff's d .

For parametric effect sizes, the weighting factor is usually the inverse of the effect size variance. However, Kromrey et al. found that applying a weighted average to Cliff's d values gave biased results⁴. Thus, they recommended using the unweighted mean of the estimate of Cliff's d from each experiment and the average of the individual experiment variances of each estimate of Cliff's d to compute the standard error and confidence intervals. This approach is exactly the same method that we suggested using to find the overall value of Cliff's d or \hat{p} from a complex experiment comprising a series of independent two-group experiments.

⁴ This happens because extreme values of Cliff's d result in small variances. As explained in the Supplementary Material (Kitchenham and Madeyski 2023), the variance of Cliff's d is based on the variability of values in a *superiority matrix*, which compares each observation in one group (e.g., Group A) with each observation in the other group (e.g., Group B), allocating the value 1 if a value in Group A is greater than a value in Group B, -1 if it is less than a value in Group B, and 0 if the two values are equal. The more the superiority matrix values tend to be the same value (either 1 or -1), the less variability there is in the superiority matrix, and the smaller the calculated variance becomes. Therefore, if a specific sample delivers an inflated estimate for d , it will also produce a very small variance, and a standard meta-analysis process will give additional weight to the inflated d value, which, in turn, will inflate the weighted mean.

Thus, if we have values of Cliff's d or \hat{p} from a series of independent experiments, we can use (4) to estimate the overall effect size and (5) to estimate its variance. In addition, for \hat{p} , we propose using Brunner et al.'s method for statistical tests of significance and the construction of confidence intervals for the overall mean that allow for any variance heterogeneity.

3.4 Other Aspects of Meta-analysis

In addition to providing estimates of the overall mean and its variance, meta-analysis procedures also encourage analysts to assess issues such as heterogeneity. Our analysis proposal supports such analyses.

3.4.1 Heterogeneity Analysis

To investigate heterogeneity among individual experiments, we can compare the estimate of the overall variance with the effect size variance, using a method similar to heterogeneity analysis for meta-analysis but assuming equal weights for each study. Using this approach, we can perform a homogeneity test based on the Q statistic where:

$$Q = \frac{\sum_{i=1}^k (NPES_i - \overline{NPES})^2}{\sigma^2} \quad (8)$$

where $NPES_i$ is the i th non-parametric effect size estimate (either Cliff's d or \hat{p}), k is the number of effect sizes being aggregated, \overline{NPES} is the average the k non-parametric effect sizes, and σ^2 is the average variance of the k $NPES_i$ values:

$$\sigma^2 = \frac{\sum_{i=1}^k \text{var}(NPES_i)}{k} \quad (9)$$

The Q statistic is distributed as a chi-squared with $k - 1$ degrees of freedom.

In addition, it is also usual to measure the extent of heterogeneity using the I^2 statistics (Higgins et al. 2003):

$$I^2 = 100 \frac{(Q - k + 1)}{Q} \quad (10)$$

where the negative values of I^2 are set to zero. I^2 values less than 25% are interpreted as indicating low heterogeneity, with values in the range 25-50% indicating moderate heterogeneity and values greater than 50% indicating high heterogeneity.

3.4.2 Fixed Effects and Random Effects Meta-Analysis

Meta-analysis methods and tools often refer to the need to distinguish between fixed and random effects analysis (Viechtbauer 2010). Fixed effects are recommended if the different experiments can all be assumed to have been based on data sets sampled from the same distribution. If such an assumption cannot be justified, analysts are recommended to use random effects analysis. If we adopt a fixed effects analysis, we should expect heterogeneity analysis to confirm low heterogeneity levels⁵.

Fixed effects meta-analysis takes advantage of the assumption that all data comes from the same distribution to give maximum weight to experiments with the lowest variance estimates.

⁵ To avoid multiple testing, the best statistical practice is to choose fixed or random effects prior to any data analysis.

If analysts use random effects analysis, they cannot assume that all variance estimates are measuring the same parameter, so assuming that low variances are indicators of more reliable results is invalid. If analysts choose a random effects analysis, the analysis method reduces the importance of the weights. In practice, large weights are reduced, and low weights are increased, making all weights closer to the same value.

In the context of meta-analysis using Cliff's d and \hat{p} , Kromey's proposal to base meta-analysis on unweighted means is equivalent to always selecting a random effects analysis. Given that SE activities in the industry are always performed on different software materials with practitioners of different skills, this seems a reasonable default.

3.4.3 Meta-analysis Using Results from Different Experimental Designs

As we discussed above, analysis of repeated measures designs (which include before-after designs, two-group AB/BA crossover designs, and four-group crossover designs) is likely to have systematically larger robust effect sizes (and power) than the equivalent between-groups effect sizes in just the same way as they have for the *StdMD* effect size. However, in the case of the robust effect sizes, there is no method for converting the effect size found in crossover designs to the equivalent effect sizes likely to be found in between-group designs.

Thus, for meta-analysis using Cliff's d and \hat{p} , we recommend the following approaches:

- Meta-analyse effect sizes calculated from repeated measures designs separately from between-groups designs to assess the probability of personal improvement.
- To calculate repeated measures effect sizes equivalent to between-groups design, *reanalyse* the experimental results considering the first time period (since the design of a between-groups experiment is identical to the design of the first part of an AB/BA crossover design). Then, all effect sizes can be meta-analysed together to assess the probability of one technique outperforming the other.

4 Effect Size Simulation Studies

In this section, we describe the simulations we undertook to evaluate Cliff's d and \hat{p} . The goal of our evaluation was to assess the value of Cliff's d and \hat{p} as alternatives to the standardized mean difference *StdMD* in situations where sample sizes are small, and the underlying data distributions are unknown. We also wanted to assess whether there was any significant difference between the effectiveness of the two non-parametric methods.

To evaluate how well Cliff's d and \hat{p} addressed our three meta-analysis requirements, i.e., being robust to non-normality, addressing different statistical designs, and supporting reliable meta-analysis, we needed to undertake a wide range of simulation studies. The scope of the studies and the evaluation criteria we used are identified in Table 2.

The simulations were organized into four main categories:

1. Simulations of two-group experiments comparing a treatment group and a control group, which are formally referred to as randomized experiments (or informally as between-groups experiments)
2. Simulations of four-group experiments comparing two treatment groups and two control groups, organized in two blocks, each containing one treatment and one control group. In our study, block effects are simulated as a fixed value to one of the parameters in one block before simulating the experiment data.

3. Simulations of families of five two-group experiments. Differences between families are simulated by adding a small random value to one of the parameters before simulating the five related experiments.
4. Simulations of families of five four-group experiments. Differences between families are simulated in a similar way to the two-group families.

In each category, simulations were split into studies investigating power, estimate error and small sample bias and studies investigating Type 1 Error rates. More details of our simulation process are presented in the following sections.

4.1 Data Distributions

Our simulation studies are based on obtaining random samples of different sizes representing two- and four-group experiments drawn from four different distributions:

1. The normal distribution (more formally referred to as the Gaussian distribution). This was selected because most data analysis methods assume normally distributed data.
2. The log-normal distribution, which is strongly skewed.
3. The gamma distribution, which is moderately skewed.
4. The Laplace distribution, which is symmetric but has more outliers than a normal distribution.

The three non-normal distributions were chosen because they exhibit properties that vary from the normal distribution properties in different ways. We discuss the properties of the data distributions in our Supplementary Material (Kitchenham and Madeyski 2023). An important issue is that the two parameters of the log-normal distribution are functionally related to one another, as are the two parameters of the gamma distribution. In both cases, this means that changes to one of the parameters that are intended to change the mean of a sample, also *cause* changes to the sample variance.

Our simulations include examples of experiments exhibiting variance heterogeneity. We investigate the impact of variance heterogeneity by increasing the variance of the treatment group data. We undertake this investigation only for the normal and Laplace data because, for these distributions, changes to the variance (spread) parameter are independent of the

Table 2 Evaluation process overview

Evaluation issue	Simulation features
Scope of simulations	<p>Data generated from normal and three non-normal probability distributions.</p> <p>Samples with variance heterogeneity.</p> <p>A range of relatively small sample sizes.</p> <p>Simulations of single two-group and four-group experiments.</p> <p>Simulations of families of 5 two-group and families of 5 four-group experiments</p>
Evaluation criteria	<ol style="list-style-type: none"> 1. Power, the probability that a statistical test <i>correctly</i> rejects the null hypothesis. 2. Type 1 error rate (α level), the probability that a test <i>incorrectly</i> rejects the null hypothesis. 3. Small sample bias, i.e., any systematic bias from the true effect size observed when averaging effect size estimates obtained from small samples. 4. Effect size error, i.e., the expected deviation between the true effect size and individual effect size estimate for small samples.

mean (μ). In fact, introducing heterogeneity into log-normally distributed data provides a good example of the extent to which non-normality can invalidate the standard parametric analysis (see Section 4.6).

Heterogeneity is an important issue because many tools supporting parametric statistical methods of formal experiments, such as analysis of variance, assume that the treatment or process under examination changes the mean of the outcome values but leaves the variance of the outcome values unchanged. This is a rather dangerous assumption in the context of software engineering methods that rely on human expertise where alternative scenarios are possible. For example, a new SE method may improve the performance of less able software engineers more than that of the more able software engineers, reducing the variance among experiment participants. Therefore, we wanted to investigate how resilient non-parametric effect sizes were to variance heterogeneity.

4.2 Evaluation Criteria and their Measurement

In this section, we explain the importance of our four evaluation criteria (see Table 2) and how we used simulation results to measure them.

We expected *StdMD* to be the most powerful effect size for normally distributed samples, but we also anticipated that the non-parametric effect sizes would perform better than *StdMD* for non-normal data. It was also possible that the non-parametric effect sizes would outperform *StdMD* for some criteria, even for normally distributed samples. In particular, even with normal samples, *StdMD* is known to exhibit small sample bias (Hedges and Olkin 1985), but given the results of Kromrey et al.'s study, we expected the non-parametric effect sizes to exhibit less small sample bias. Small sample bias is an important factor in meta-analysis because aggregating effect sizes that suffer from small sample bias will result in biased meta-analysis estimates.

We discuss each of the four criteria in the following sections

4.2.1 Power

Power is usually considered a critical factor when comparing the effectiveness of alternative statistical methods. We estimated power for a specific combination of non-zero effect size, data distribution and sample size as the proportion of samples for which the difference between control and treatment samples were correctly assessed as being statistically different from zero.

We wanted to investigate whether the benefits of using Cliff's d or \hat{p} for small sample sizes with non-normal data were sufficient to make up for the expected loss of power should data be normally distributed.

Since we were interested both in whether the non-parametric effect sizes were more powerful for non-normal distributions and whether there was any general difference between Cliff's d and \hat{p} , we report the power difference between *StdMD* and each of the non-parametric effect size as:

$$PowerDiff = 100 \times (NPESPower - StdMDPower) \quad (11)$$

where *NPESPower* is the power for each non-parametric effect size, and *StdMDPower* is the power of the standardized mean difference calculated for two-group and four-group samples, for different samples and each data distribution and various non-zero effect size differences. A positive *PowerDiff* value implies that the non-parametric effect size has

out-performed *StdMD*; in contrast, a negative *PowerDiff* value implies that *StdMD* has out-performed the non-parametric effect size.

4.2.2 Small Sample Size Bias

A parameter is said to exhibit small sample bias if the estimate of the parameter obtained by taking the average of parameter estimates obtained from many small samples differs from the expected value parameter. We report the extent of small sample bias as a percentage relative error:

$$\text{PercentageRelativeError} = \frac{100 \times (\text{ExpectedES} - \text{ObservedES})}{\text{ExpectedES}} \quad (12)$$

where

$$\text{ObservedES} = \frac{\sum_{i=1}^{10000} (\text{ObservedES}_i)}{10000} \quad (13)$$

ObservedES_i is the effect size for the i th of the 10000 effect size estimates obtained from a specific sample size, effect size and data distribution. *ExpectedES* is the true population effect size for the specific simulation conditions. For *StdMD*, *ExpectedES* is the theoretical effect size, calculated by substituting the parameter values used in our simulation studies into the relevant probability distributions. For the non-parametric effect sizes, *ExpectedES* is the effect size found from a single ultra-large experiment.

Comparisons of relative bias are always problematic when the values of the divisors are expected to differ. In our case, to avoid misleading results, it is necessary to use the *centralised* \hat{p} (i.e., subtract 0.5 from each estimate of \hat{p}). The 0.5 value is a standard adjustment *constant*; it is not an element of \hat{p} that is subject to any estimation uncertainty, but it does inflate the bias divisor. The important point is that using the centralised version of \hat{p} , the relative bias values for \hat{p} and Cliff's d are always *exactly* equal because \hat{p} and Cliff's d are directly functionally related to each other. Even so, the relative bias values can be slightly misleading because for each effect size, the expected value of Cliff's d and centralized \hat{p} is less than the theoretical value *StdMD*, so we would expect our measure of relative bias to slightly favour *StdMD*.

4.2.3 Effect Size Estimate Error

Since the non-parametric effect sizes are intended to be robust estimators, it was also possible that estimates would be more accurate than *StdMD*. Rather than accuracy, we estimated the *relative error* of the effect size estimates for samples of different sizes. For each of the 10000 simulations in each simulation condition, we calculated the magnitude (absolute) relative error (MRE) as:

$$\text{MRE}_i = \frac{|\text{ExpES} - \text{ObsES}_i|}{|\text{ExpES}|} \quad (14)$$

where *ExpES* is the expected effect size for a particular effect size (parametric or non-parametric), for the specific simulation condition and ObsES_i is the observed value of the effect size for a specific simulation, and $i = 1, \dots, 10000$.

For each condition, we calculate the median of the MRE_i values, giving the median magnitude relative error (*MdMRE*) for the specific effect size and the specific simulation

condition. We multiplied $MdMRE$ by 100 to give the percentage median magnitude error ($PMdMRE$):

$$PMdMRE = 100 \times MdMRE = 100 \times \text{median}(MRE_i) \quad (15)$$

Like relative bias, $MdMRE$ must be based on the centralised \hat{p} , which means that the $MdMRE$ values are exactly the same for \hat{p} and Cliff's d . Again, theoretically, the fact that the magnitude of Cliff's d and centralized \hat{p} are less than the corresponding theoretical magnitude of $StdMD$ means that $MdMRE$ slightly favours $StdMD$.

4.2.4 Type 1 Error Rates

The Type 1 error rate is the probability that a statistical test will incorrectly reject the null hypothesis. A well-constructed test process will ensure that the Type 1 error rate is close to the α -level of the tests. We expected the robust effect sizes and $StdMD$ to behave similarly with respect to Type 1 error rates, i.e., tests of significance performed at the $\alpha = 0.05$ level should lead to a Type 1 error rate of approximately 0.05. The Type 1 error rate was estimated as the proportion of a set of simulations, with a mean difference of zero and the same sample size and data distribution, that incorrectly found the mean difference significantly different from zero. The test was based on the t -tests for $StdMD$ and the confidence intervals for Cliff's d and \hat{p} . Maintaining the expected Type 1 error rate is critical for hypothesis testing and constructing valid confidence intervals.

4.3 Two-Group Experiment Simulation Details

In this section, we specify the experimental conditions we used to simulate two-group experiments.

All our two-group simulations were based on simple between-group experiments. We simulated experiments with 10, 20, 30, 40, and 80 observations (i.e., 5, 10, 15, 20, and 40 observations per group) for normal, log-normal, gamma, and Laplace distributions. We chose to simulate samples of 10, 20, 30, and 40 participants because study sizes between 10 and 40 participants are typical of the small sizes we see in SE experiments.

The extreme values for our simulations were chosen for different reasons:

- We simulated two-group samples with 5 observations per group because Wilcox reported concerns about \hat{p} when sample sizes were very small. So, we wanted to investigate whether \hat{p} was significantly flawed.
- We simulated two-group samples with 40 observations per group since a more straightforward way of addressing small sample sizes is to use bigger samples, and we wanted to investigate whether larger samples are sufficient to avoid the need for non-parametric effect sizes.

For the normal distribution, for the control group, we used a distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. For the treatment group, we used four different mean values (0, 0.2, 0.5, 0.8). The nonzero values correspond to the values that Cohen (1992) identified as small, medium and large standardized effect sizes in the context of psychological studies. These are usually considered acceptable ranges for simulation studies. For the other distributions,

Table 3 Simulation parameter values

Distribution	Parameter 1 Values	Parameter 2 Value
Normal	Mean $\mu = (0, 0.2, 0.5, 0.8)$	Variance $\sigma^2 = 1$
Log-normal	Mean $\mu = (0, 0.266, 0.72375, 1.43633)$	Variance $\sigma^2 = 1$
Gamma	Rate $\beta = (1, 1.1225, 1.3415, 1.6224)$	Shape $\alpha = 3$
Laplace	Mean $\mu = (0, 0.283, 0.707104, 1.131374)$	Shape $\beta = 1$

we chose parameter values that would have theoretical standardized effect sizes magnitudes of $(0, 0.2, 0.5, 0.8)$ on the raw data scale. The parameter values used in the simulations are shown in Table 3. In the case of the gamma distribution, for the control group, we used the value 1 for the rate parameter and 3 for the shape parameter since the gamma distribution must have a rate parameter greater than 0. These values were chosen for convenience. For the treatment effect, we used the parameter values that generated negative standardized mean differences of $(-0.2, -0.5, -0.8)$ because increasing the rate parameter decreases the mean on the raw data scale. We kept the expected magnitude of *StdMD* values the same for the different probability distributions to make the simulation outcomes more comparable.

For each *non-zero* effect size and each distribution, the simulation process was:

1. Establish the population values of \hat{p} and Cliff's d (see the next paragraph).
2. For each required sample size (i.e., 10, 20, 30, 40, and 80), generate data for 10000 two-group experiments and, for each experiment, obtain the sample estimates of *StdMD*, \hat{p} , and Cliff's d . For each of the 10000 experiments, return the value of each effect size and three output values that identify whether each analysed effect size was significant, based on *one-sided* tests with $\alpha = 0.05$. We use one-sided tests because we set the direction of the effect for each simulation. If a simulation produces a significant result in the wrong direction, it is equivalent to a Type 1 error because it has incorrectly rejected the null hypothesis.
3. From the values returned from the 10000 simulated experiments, calculate the percentage magnitude relative error (i.e., *PMdMRE*), the percentage relative bias, the power of the 10000 simulated experiments, and the *PowerDiff* value for each non-parametric effect size.
4. We simulated normal and Laplace experiments with variance heterogeneity by setting the standard deviation for the treatment group to 1.5. This change implies changes to the theoretical *StdMD* estimates and the large sample estimates of the non-parametric effect sizes, as shown in Table 4.

To obtain an estimate of the population values of \hat{p} and Cliff's d , for each data type and effect size, we simulated ultra-large experiments with 10,000,000 observations per group. Then, we calculated *StdMD*, \hat{p} , and Cliff's d from the sample. For *StdMD*, the theoretical

Table 4 Expected non-parametric effect sizes values

Distribution	\hat{p}	Cliff's d
Normal	$(0, 0.556, 0.638, 0.714)$	$(0, 0.112, 0.276, 0.428)$
Log-normal	$(0, 0.575, 0.696, 0.845)$	$(0, 0.149, 0.391, 0.69)$
Gamma	$(0, 0.446, 0.365, 0.286)$	$(0, -0.108, -0.269, -0.428)$
Laplace	$(0, 0.57, 0.666, 0.747)$	$(0, 0.14, 0.332, 0.495)$

values can be calculated from the relevant probability density distribution, so if the calculated *StdMD* values are very close to the theoretical values, we can assume that the sample is a good representation of the population. Then, we assume that the estimates of \hat{p} and Cliff's d from the ultra-large sample are also close estimates of the population effect sizes. This process was undertaken for:

- Two-group experiments, using samples for each data type. For the gamma distribution, we also investigated the impact of both increasing and decreasing the value of the rate parameter.
- Two-group experiments, with an increase in variance for the treatment group, for normal, log-normal and Laplace data.
- Four-group experiments, with and without a fixed block effect for each data type.
- Four group experiments with an increase in variance for the treatment group, for normal, log-normal and Laplace data.

The results of this process are shown in tables in Section 4 of the Supplementary Material (Kitchenham and Madeyski 2023). The non-parametric effect size estimates used as input parameters in our two-experiment simulations are shown in Tables 4 and 5.⁶

We investigated the null hypothesis error rates using simulations where the difference between the control and treatment means (or rate parameter) was zero. In this case, the theoretical values of *StdMD* and Cliff's d are both zero and the theoretical value of \hat{p} is 0.5. This means we cannot construct a reliable measure of relative error. For Type 1 error rate assessments, all statistical tests were *two-sided tests* because significant effects in either direction indicate a Type 1 error.

4.4 Two Group Simulation Results

For simulations of two-group experiments, relative estimate error, power, and relative bias results for non-zero effect sizes are shown in Table 6 and the Type 1 error rates are shown in Table 7. In both tables, the "Type" column defines the data type used in the simulations reported in each row: N for normal data, N-H for normal data with extra heterogeneity, Lap for Laplace data, Lap-H for Laplace data with extra heterogeneity, L for log-normal data and G for gamma data. We present only one *PMdMRE* value and one relative bias value for the non-parametric effect sizes in Table 6 because centralised \hat{p} and Cliff's d , *MdMRE* values and relative bias are identical. The column labelled *GrpSize* indicates the number of observations in each group. The column labelled *Diff* in Table 6 has values labelled S, M, and H corresponding to the small, medium and large differences between the control and treatment groups.

The results tables are quite long and detailed, so we provide the set of graphs shown in Fig. 4 and additional summary statistics derived from the tables to identify the most important results.

The graphs were constructed from the outcome values for power, bias, effect size estimate error and type 1 error rates. The graphs allow us to summarize the effectiveness of the effect sizes across the different sample distributions, effect sizes and sample sizes:

⁶ There are some slight differences between the values reported in the Supplementary Material and the values in this paper (of the order of .001). This is intentional. If there was a difference between the two-group estimates and the four-group estimates of the non-parametric effect sizes when the values should be identical, we gave preference to the four-group estimates (since they were based on more observations) unless there was any inconsistency between the \hat{p} and Cliff's d estimates, in which case we took the average estimate or the most consistent estimate.

Table 5 Effect of variance heterogeneity on normal and Laplace effect size values

Distribution	Theoretical <i>StdMD</i>	Large sample \hat{p}	Large sample Cliff's <i>d</i>
Normal	(0, 0.157, 0.392, 0.628)	(0, 0.544, 0.609, 0.671)	(0, 0.088, 0.219, 0.343)
Laplace	(0, 0.157, 0.392, 0.628)	(0, 0.556, 0.6355, 0.706)	(0, 0.112, 0.271, 0.411)

- The graph in the top row of Fig. 4 shows the relationship between power, experiment size and effect size for each of the effect size estimates. The experiment size for two-group experiments is twice the group size reported in Table 6. Each boxplot is based on 18 outcome variables of the specified experiment size for each of the six different data distributions (i.e., the normal and Laplace data samples both with and without variance heterogeneity, and the gamma and log-normal data samples) and each of the three different non-zero effect size differences. The variation within each boxplot is mainly due to the size of the mean difference, with the power being greatest for the largest mean difference. Since power levels of 0.8 are usually recommended for reliable experiments, the results show clearly that the power of experiments is unacceptably low for experiments of 40 or fewer observations for both parametric and non-parametric effect sizes. For experiments of 80 observations, power is still low when the difference between the groups is small.
- The lefthand graph on the second row of Fig. 4 reports the power difference between each of the non-parametric effect sizes and *StdMD* (multiplied by 100). Each boxplot is based on 90 observations corresponding to each of the six different data types, five different sample sizes, and three different effect size differences. Positive values indicate that the non-parametric effect size has better power than *StdMD*, and negative values indicate that they have worse power than *StdMD*. The boxplots confirm that the non-parametric effect sizes are frequently more powerful than *StdMD*. Specifically, the power of Cliff's *d* is better than the power of *StdMD* in 37 of the 90 (i.e., more than 41%), while the power of \hat{p} is better in 55 of the 90 (i.e., more than 61% of) cases. Reference to Table 6 confirms that the non-parametric effect sizes are usually more powerful for Laplace and log-normal samples, while *StdMD* is usually more powerful for normal and gamma samples.
- The righthand graph on the second row of Fig. 4 is also based on 90 observations. It shows very clearly that the non-parametric effect sizes do not exhibit any systematic small sample bias, whereas estimates of *StdMD* systematically overestimate the true value of the effect size, sometimes by very large amounts. Reference to Table 6 confirms that the particularly large overestimates for *StdMD* correspond to estimates derived from log-normal samples. This result is important because it means we can trust aggregates of small sample estimates for Cliff's *d* and \hat{p} but we cannot trust aggregate estimates of *StdMD* from small samples.
- The lefthand graph on the bottom row indicates that there is little difference between the *PMdMRE* values for the non-parametric effect sizes and *StdMD*. Specifically and bearing in mind that small *PMdMRE* values indicate better accuracy, *PMdMRE* values for the non-parametric effect size were less than the *PMdMRE* values for the corresponding *StdMD* effect size for 56 of the 90 simulations (62%).
- The righthand graph on the bottom row of Fig. 4 is based on the 30 Type 1 error rates reported in Table 7. All tests were performed at the 0.05 α level, so the expected outcome of the simulations should be approximately 0.05. In fact, \hat{p} systematically overestimates with a median value of 0.0539, whereas Cliff's *d* and *StdMD* systematically underesti-

Table 6 Relative error, bias and power for two group experiments

Type	Grp Size	Diff	NP Bias	StdMD Bias	PMdMRE NP	PMdMRE StdMD	Obs PHat	Obs ClfId	Obs StdMD	Power PHat	Power ClfId	Power StdMD
N	5	S	-3.83	6.03	221.43	219.32	0.554	0.108	0.212	0.08	0.06	0.08
N	5	M	-1.93	8.87	88.41	89.64	0.635	0.271	0.544	0.17	0.12	0.16
N	5	L	-1.15	9.59	53.27	57.73	0.712	0.423	0.877	0.30	0.21	0.29
N	10	S	1.81	5.77	167.86	155.57	0.557	0.114	0.212	0.12	0.10	0.11
N	10	M	0.79	5.03	63.77	62.52	0.639	0.278	0.525	0.28	0.25	0.29
N	10	L	0.41	4.85	39.25	40.37	0.715	0.430	0.839	0.52	0.47	0.53
N	15	S	1.64	3.64	127.78	125.97	0.557	0.114	0.207	0.14	0.12	0.14
N	15	M	0.49	3.19	50.08	50.69	0.639	0.277	0.516	0.37	0.34	0.38
N	15	L	0.24	3.08	30.43	32.53	0.715	0.429	0.825	0.67	0.65	0.69
N	20	S	-0.26	0.68	109.82	107.32	0.556	0.112	0.201	0.15	0.13	0.15
N	20	M	-0.22	1.42	43.12	43.11	0.638	0.275	0.507	0.45	0.43	0.46
N	20	L	-0.13	1.61	26.17	27.84	0.714	0.427	0.813	0.78	0.77	0.80
N	40	S	-0.01	0.86	77.46	75.50	0.556	0.112	0.202	0.22	0.21	0.22
N	40	M	-0.07	0.93	30.25	30.47	0.638	0.276	0.505	0.70	0.69	0.72
N	40	L	-0.04	0.94	18.22	19.80	0.714	0.428	0.808	0.96	0.96	0.97
N-H	5	S	-0.69	11.17	309.09	285.73	0.544	0.087	0.175	0.09	0.06	0.08
N-H	5	M	-0.12	12.04	118.35	116.03	0.609	0.218	0.439	0.14	0.10	0.14
N-H	5	L	-0.51	12.08	75.44	74.20	0.670	0.340	0.704	0.22	0.15	0.22
N-H	10	S	-3.79	1.70	213.64	195.15	0.542	0.085	0.160	0.09	0.08	0.09
N-H	10	M	-1.47	3.84	81.65	79.42	0.607	0.215	0.407	0.20	0.18	0.21
N-H	10	L	-0.67	4.21	52.05	50.80	0.670	0.340	0.654	0.36	0.33	0.38
N-H	15	S	1.81	3.63	167.68	159.41	0.545	0.090	0.163	0.10	0.09	0.11
N-H	15	M	0.26	3.43	65.34	63.72	0.609	0.219	0.405	0.26	0.24	0.27
N-H	15	L	0.04	3.22	39.05	40.47	0.671	0.342	0.648	0.48	0.46	0.51
N-H	20	S	0.56	2.99	145.45	138.61	0.544	0.088	0.162	0.12	0.11	0.12

Table 6 continued

Type	Grp Size	Diff	NP Bias	StdMD Bias	PMdMRE NP	PMdMRE StdMD	Obs PHat	Obs Clifflid	Obs StdMD	Power PHat	Power Clifflid	Power StdMD
N-H	20	M	0.25	2.73	56.42	56.10	0.609	0.219	0.403	0.32	0.30	0.34
N-H	20	L	0.15	2.50	34.50	35.44	0.671	0.343	0.644	0.59	0.57	0.62
N-H	40	S	2.06	2.41	98.86	95.95	0.545	0.090	0.161	0.16	0.16	0.17
N-H	40	M	1.00	1.76	39.22	38.33	0.610	0.220	0.399	0.52	0.51	0.54
N-H	40	L	0.70	1.44	23.61	24.29	0.672	0.344	0.637	0.85	0.85	0.88
Lap	5	S	2.60	25.59	185.71	234.68	0.572	0.144	0.251	0.10	0.07	0.09
Lap	5	M	1.50	23.00	80.72	96.59	0.668	0.337	0.615	0.21	0.15	0.19
Lap	5	L	1.13	22.38	43.32	63.88	0.750	0.500	0.979	0.35	0.23	0.35
Lap	10	S	-0.01	10.13	128.57	159.36	0.570	0.140	0.220	0.14	0.11	0.12
Lap	10	M	0.49	10.01	51.81	66.37	0.667	0.334	0.550	0.35	0.32	0.31
Lap	10	L	0.26	10.00	33.60	43.68	0.748	0.495	0.880	0.61	0.56	0.56
Lap	15	S	2.87	9.95	103.17	129.02	0.572	0.144	0.220	0.16	0.15	0.14
Lap	15	M	1.38	7.94	40.56	53.08	0.668	0.337	0.540	0.48	0.45	0.41
Lap	15	L	1.04	7.46	25.33	35.53	0.750	0.499	0.860	0.78	0.76	0.71
Lap	20	S	-0.10	4.37	89.29	109.81	0.570	0.140	0.209	0.18	0.17	0.16
Lap	20	M	0.18	4.51	35.54	45.52	0.666	0.333	0.523	0.56	0.55	0.48
Lap	20	L	0.16	4.56	22.06	29.97	0.747	0.495	0.836	0.88	0.86	0.80
Lap	40	S	2.19	5.03	62.50	76.15	0.572	0.143	0.210	0.29	0.28	0.23
Lap	40	M	1.12	3.47	25.08	31.25	0.668	0.336	0.517	0.83	0.83	0.74
Lap	40	L	0.76	3.10	15.49	20.94	0.749	0.498	0.825	0.99	0.99	0.97
Lap-H	5	S	-1.06	22.89	221.43	296.88	0.555	0.111	0.193	0.09	0.06	0.07
Lap-H	5	M	-0.47	22.53	91.88	120.47	0.635	0.270	0.480	0.16	0.11	0.15
Lap-H	5	L	-0.71	22.26	65.85	78.27	0.705	0.409	0.768	0.25	0.17	0.25
Lap-H	10	S	0.71	10.09	167.86	205.13	0.556	0.113	0.173	0.12	0.10	0.10
Lap-H	10	M	-0.10	10.55	69.74	83.51	0.635	0.271	0.433	0.27	0.23	0.23

Table 6 continued

Type	Grp Size	Diff	NP Bias	StdMD Bias	PMdMRE NP	PMdMRE StdMD	Obs PHat	Obs ClifId	Obs StdMD	Power PHat	Power ClifId	Power StdMD
Lap-H	10	L	-0.31	10.51	41.46	54.07	0.705	0.411	0.694	0.46	0.42	0.41
Lap-H	15	S	-4.03	2.62	134.13	161.97	0.554	0.107	0.161	0.12	0.11	0.11
Lap-H	15	M	-1.82	5.46	52.52	66.10	0.633	0.266	0.413	0.34	0.31	0.29
Lap-H	15	L	-1.19	6.01	33.33	42.51	0.704	0.407	0.666	0.60	0.57	0.52
Lap-H	20	S	-2.72	1.16	109.82	141.03	0.554	0.109	0.159	0.14	0.13	0.12
Lap-H	20	M	-1.20	3.56	43.91	57.40	0.634	0.268	0.406	0.42	0.40	0.35
Lap-H	20	L	-0.89	4.02	28.05	37.16	0.704	0.408	0.653	0.73	0.71	0.63
Lap-H	40	S	-0.67	1.29	81.03	100.43	0.556	0.111	0.159	0.22	0.21	0.18
Lap-H	40	M	-0.42	2.07	32.66	40.88	0.635	0.270	0.400	0.66	0.65	0.55
Lap-H	40	L	-0.46	2.12	20.12	26.16	0.705	0.410	0.641	0.94	0.94	0.86
L	5	S	3.31	35.27	180.00	249.78	0.577	0.155	0.271	0.10	0.07	0.06
L	5	M	0.64	38.44	69.39	90.35	0.697	0.395	0.692	0.27	0.19	0.16
L	5	L	0.50	53.12	24.64	53.87	0.847	0.693	1.225	0.63	0.41	0.39
L	10	S	0.64	20.60	113.33	168.25	0.575	0.151	0.241	0.14	0.12	0.10
L	10	M	0.12	24.55	42.86	59.27	0.696	0.392	0.623	0.45	0.41	0.34
L	10	L	0.27	34.95	18.84	36.47	0.846	0.692	1.080	0.91	0.88	0.77
L	15	S	-2.26	14.87	98.52	132.36	0.573	0.147	0.230	0.17	0.15	0.13
L	15	M	-0.79	19.30	34.92	46.63	0.694	0.389	0.597	0.59	0.56	0.47
L	15	L	-0.06	27.78	14.01	30.47	0.845	0.690	1.022	0.98	0.97	0.90
L	20	S	-0.55	12.37	83.33	114.99	0.575	0.149	0.225	0.20	0.19	0.16
L	20	M	-0.23	16.22	29.85	40.32	0.696	0.391	0.581	0.70	0.69	0.58
L	20	L	0.07	23.73	12.32	27.35	0.845	0.690	0.990	1.00	1.00	0.96
L	40	S	-0.16	7.89	58.56	77.03	0.574	0.149	0.216	0.31	0.30	0.25
L	40	M	-0.25	11.36	19.96	27.84	0.696	0.391	0.557	0.93	0.93	0.83
L	40	L	0.10	16.53	8.33	20.67	0.845	0.691	0.932	1.00	1.00	1.00

Table 6 continued

Type	Grp Size	Diff	NP Bias	StdMD Bias	PMdMRE NP	PMdMRE StdMD	Obs PHat	Obs ClifId	Obs StdMD	Power PHat	Power ClifId	Power StdMD
G	5	S	4.12	8.30	233.33	232.20	0.444	-0.112	-0.217	0.09	0.06	0.07
G	5	M	1.13	8.82	92.59	91.11	0.363	-0.273	-0.544	0.17	0.13	0.15
G	5	L	0.43	10.10	58.88	54.92	0.285	-0.430	-0.881	0.30	0.21	0.28
G	10	S	0.70	3.71	174.07	156.00	0.446	-0.109	-0.207	0.11	0.09	0.11
G	10	M	0.29	3.84	62.96	60.61	0.365	-0.271	-0.519	0.27	0.24	0.28
G	10	L	0.32	4.46	39.25	36.84	0.285	-0.429	-0.836	0.52	0.47	0.54
G	15	S	0.31	1.23	134.57	124.55	0.446	-0.108	-0.202	0.12	0.11	0.13
G	15	M	-0.13	2.20	52.26	49.07	0.365	-0.270	-0.511	0.35	0.33	0.37
G	15	L	0.22	2.86	30.43	29.67	0.286	-0.429	-0.823	0.67	0.64	0.71
G	20	S	0.18	2.02	117.59	106.80	0.446	-0.108	-0.204	0.14	0.13	0.15
G	20	M	0.19	2.01	44.44	41.84	0.365	-0.271	-0.510	0.44	0.42	0.47
G	20	L	0.39	2.31	26.17	25.20	0.285	-0.430	-0.818	0.78	0.77	0.82
G	40	S	-0.75	-0.22	81.60	76.81	0.446	-0.107	-0.200	0.21	0.20	0.22
G	40	M	-0.43	0.53	31.48	29.91	0.366	-0.269	-0.503	0.67	0.66	0.72
G	40	L	0.02	0.88	18.28	18.13	0.286	-0.428	-0.807	0.96	0.96	0.98

Table 7 Type 1 error rates for two group experiments

Design Type	Grp SizeObserved:.....		Type 1 Error Rate:.....		
		\hat{p}	Cliff's d	$StdMD$	\hat{p}	Cliff's d	$StdMD$
N	5	0.5027	0.0054	0.0066	0.0777	0.0324	0.0436
N	10	0.5005	0.0009	0.0011	0.0597	0.0435	0.0527
N	15	0.4993	-0.0013	-0.0025	0.0520	0.0410	0.0484
N	20	0.4995	-0.0009	-0.0018	0.0532	0.0437	0.0497
N	40	0.4992	-0.0017	-0.0038	0.0534	0.0480	0.0519
N-H	5	0.5011	0.0021	0.0061	0.0804	0.0329	0.0439
N-H	10	0.4990	-0.0020	-0.0025	0.0550	0.0393	0.0491
N-H	15	0.5018	0.0036	0.0062	0.0568	0.0449	0.0515
N-H	20	0.4999	-0.0002	0.0004	0.0514	0.0423	0.0475
N-H	40	0.5000	0.0000	-0.0005	0.0539	0.0499	0.0537
Lap	5	0.4999	-0.0001	-0.0009	0.0803	0.0312	0.0345
Lap	10	0.5004	0.0008	0.0000	0.0562	0.0402	0.0458
Lap	15	0.5007	0.0014	0.0022	0.0512	0.0388	0.0459
Lap	20	0.5007	0.0013	0.0033	0.0547	0.0456	0.0479
Lap	40	0.4998	-0.0003	-0.0008	0.0485	0.0426	0.0479
Lap-H	5	0.4985	-0.0030	-0.0025	0.0731	0.0315	0.0332
Lap-H	10	0.4975	-0.0049	-0.0076	0.0558	0.0385	0.0437
Lap-H	15	0.4998	-0.0004	0.0000	0.0507	0.0395	0.0447
Lap-H	20	0.5008	0.0016	0.0046	0.0499	0.0428	0.0470
Lap-H	40	0.5003	0.0005	0.0020	0.0539	0.0489	0.0478
L	5	0.5014	0.0027	0.0023	0.0811	0.0336	0.0181
L	10	0.4976	-0.0049	-0.0124	0.0569	0.0383	0.0256
L	15	0.4990	-0.0020	-0.0028	0.0492	0.0381	0.0307
L	20	0.4986	-0.0028	-0.0067	0.0494	0.0392	0.0353
L	40	0.4994	-0.0012	-0.0032	0.0513	0.0467	0.0431
G	5	0.4979	-0.0041	-0.0065	0.0778	0.0318	0.0372
G	10	0.4999	-0.0002	0.0019	0.0581	0.0408	0.0450
G	15	0.4992	-0.0016	-0.0017	0.0519	0.0390	0.0440
G	20	0.4996	-0.0009	-0.0006	0.0483	0.0410	0.0442
G	40	0.4999	-0.0003	-0.0003	0.0517	0.0468	0.0489

mate with median values of 0.0405 and 0.0454, respectively. The large outliers shown for \hat{p} correspond to the smallest sample size, which is consistent with the research reported by Wilcox (2012). The small outliers shown for $StdMD$ correspond to the log-normal samples.

4.5 Simulations of Four Group Randomized Blocks Experiments

The basic parameters of the four-group simulations were similar to those of the two-group simulations in terms of the choice of underlying distributions and their parameter values and numbers of replications. However, we used group sizes of 5, 10, 15, 20 and 40 to simulate experiments with total sample sizes of 20, 40, 60, 80 and 120 participants. In addition, for

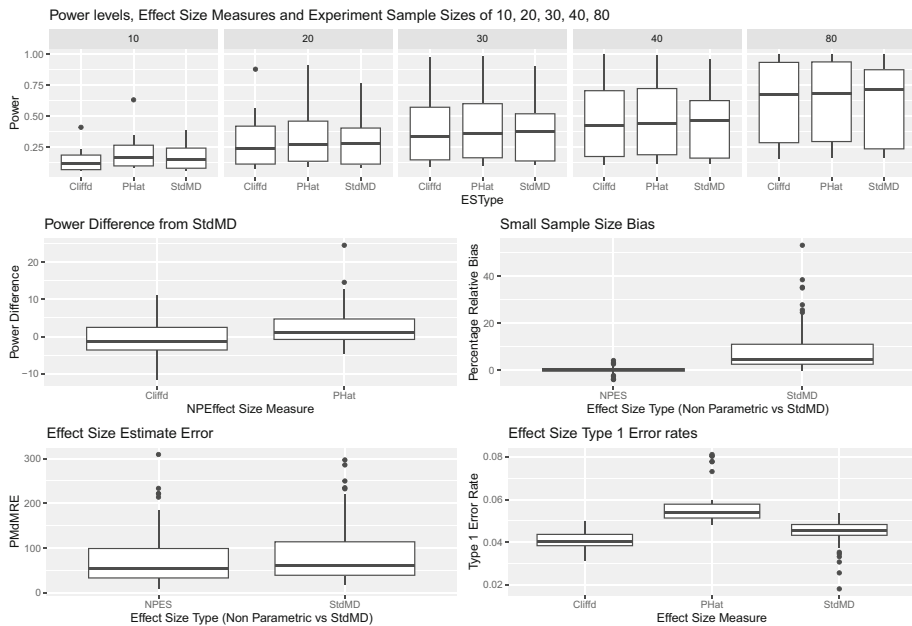


Fig. 4 Two-group simulation experiments results summary (Boxplots defined by the x-axis and obtained from all four data distributions)

each experiment, two groups (one corresponding to each treatment and control technique) were given an adjustment to simulate a difference between the blocks.

For the normal, log-normal and Laplace distributions, the block effect was modelled as a fixed 0.5 increase to the mean effect; for the Gamma distributions, the block effect was modelled as a fixed 0.5 increase to the shape parameter.

For normal and log-normal samples, we simulated experiments with and without variance heterogeneity, and all samples included the block effect. For both these distributions, the change to four-group experiments with a block effect left the theoretical *StdMD* effect sizes and the large sample non-parametric effect sizes unchanged compared with the two-group values.

For the log-normal and Gamma distributions, we simulated samples with and without the block effect. For both these distributions, the change to four-group experiments without the block effect should leave the theoretical *StdMD* effect sizes and the large sample non-parametric effect sizes unchanged compared with the two-group values. However, the inclusion of the block effect slightly altered the theoretical effect sizes for *StdMD* and the large sample non-parametric effect sizes as shown in Table 8. For simulated experiments that

Table 8 Effect of block effects on log-normal and gamma effect sizes values

Distribution	Theoretical <i>StdMD</i>	Large sample \hat{p}	Large sample Cliff's <i>d</i>
Log-normal	(0.194, 0.486, 0.777)	(0.575, 0.696, 0.845)	(0.149, 0.391, 0.69))
Gamma	(−0.208, −0.52, −0.833)	(0.444, 0.359, 0.277)	(−0.113, −0.281, −0.445)

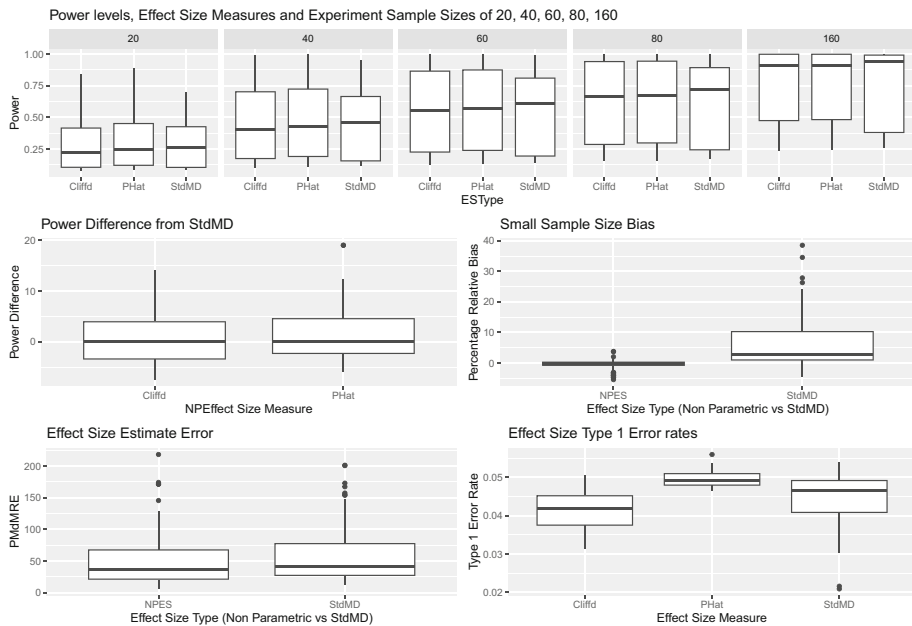


Fig. 5 Four-group simulation experiments results summary (Boxplots constructed from outputs defined by the x-axis and obtained from all four data distributions)

included the block effect, bias and error were assessed against the revised theoretical and large sample effect sizes.

The full result tables for the simulations of four-group randomized block experiments can be found in the Supplementary Material (Kitchenham and Madeyski 2023) in Section 5 and are summarized in Fig. 5. The results were very similar to those obtained from the two-group experiments:

- Power increases for the non-parametric effect sizes and *StdMD* as the experiment sizes increase and as the effect size difference increases. In addition, the power diagrams for experiment sizes of 20, 40 and 80 for the two-group and four-group graphs are virtually identical.
- The *PowerDiff* results for Cliff's d and \hat{p} were more similar than for the two-group simulations. These results are discussed in Section 4.7
- The small sample size bias is again negligible for the non-parametric effect sizes but substantial for *StdMD*.
- The *PMdMRE* values for the non-parametric effect size were less than the *PMdMRE* values for the corresponding *StdMD* effect size for 71 of the 120 simulations (59%).
- The main change was that the Type 1 error rates were less biased with medians of 0.042, 0.0493, and 0.0465 for Cliff's d , \hat{p} , and *StdMD* respectively.

4.6 Log-normal Sample with Additional Variance Heterogeneity

Under most of the conditions we simulated, *StdMD* estimates from non-normal samples provided what appeared to be reasonable analysis results. As we show in this section, this

is not the case for *StdMD* estimates from log-normal samples when variance heterogeneity is introduced. Table 9 shows the results of analysing simulations of two-group experiments based on log-normal data. Each entry in the table is based on an ultra-large experiment with 10000000 observations in each group that is intended to deliver effect size estimates close to the population values. Each group of three rows use the same mean and variance parameters. In all cases, the variance parameter of the control group was set 1, and the variance of the treatment group was set to $1.5^2 = 2.25$. Rows 1-3 simulate the situation when there is no difference between the treatment and control group means. The other groups of 3 rows use the Small (S), Medium (M) and Large (L) mean differences for log-normal data shown in Table 3.

For each group of three rows with the same mean difference, the top row shows the effect of calculating *StdMD* directly from the mean and variance of the simulated log-normal data. In the absence of variance heterogeneity, *StdMD* values for 0, S, M, and L mean differences should be 0, 0.2, 0.5 and 0.8, respectively, and clearly, the variance heterogeneity has significantly impacted the estimates. However, results shown in row 1 make it clear that \hat{p} and Cliff's *d* analysis leads to estimates that are unaffected by variance heterogeneity. The subsequent two rows show, respectively, the results of analysing the data after transforming log-normal data and the results of analysing data simulated from a normal distribution with the same mean and variance as the log-normal simulations. The second two rows all have the same *StdMD*, \hat{p} and Cliff's *d* estimate values. The same pattern of results is seen for each set of three related analyses.

The *StdMD* values from the first row in each group of three are consistent with the theoretical values found by substituting the parameter values into the log-normal probability distribution. However, the *StdMD* results based on raw log-normal data are invalid because the correct analysis process requires transforming the data prior to analysis. This is an example of a case when using a standard parametric analysis because the data was unknown, which would lead to grossly invalid results irrespective of sample size. It is also a good example of how trustworthy the non-parametric effect sizes can be.

Table 9 Large sample results for log-normal samples of two-group experiments with additional variance heterogeneity

	Sample Type	Mean Difference	Data Analysed	Phat Estimate	Cliff's <i>d</i> Estimate	<i>StdMD</i> Estimate
1	L-Het	0	Raw	0.50	0.00	0.22
2	L-Het	0	Transformed	0.50	0.00	0.00
3	N-Het	0	Raw	0.50	0.00	0.00
4	L-Het	S	Raw	0.56	0.12	0.28
5	L-Het	S	Transformed	0.56	0.12	0.21
6	N-Het	S	Raw	0.56	0.12	0.21
7	L-Het	M	Raw	0.66	0.31	0.36
8	L-Het	M	Transformed	0.66	0.31	0.57
9	N-Het	M	Raw	0.66	0.31	0.57
10	L-Het	L	Raw	0.79	0.57	0.43
11	L-Het	L	Transformed	0.79	0.57	1.13
12	N-Het	L	Raw	0.79	0.57	1.13

4.7 Comparing Non-parametric Effect Sizes Cliff's d and \hat{p}

Table 10 compares the power levels for \hat{p} and Cliff's d observed for each sample size, effect size and data type. For both two-group and four-group experiment simulations, the \hat{p} power level is greater than or equal to the Cliff's d power level.

Wilcox reported that Cliff's d Type 1 error rates were better than \hat{p} Type 1 error rates for very small sample sizes (Wilcox 2012). We also observed this effect in our two-group experiment simulations. Table 11 compares the median Type 1 error rates and power difference values for Cliff's d and \hat{p} over the full range of sample sizes and distributions in our two- and four-group simulations. This table confirms that, as shown in the bottom left panes of Figs. 4 and 5, the median \hat{p} Type 1 error rate is closer to 0.05 than Cliff's d for both two-group and four-group simulations. It also appears that \hat{p} performs marginally (but not significantly) better than Cliff's d in terms of its power difference against $StdMD$, particularly for two-group experiment simulations.

5 Meta-Analysis Simulations

Although our aim was simply to compare meta-analysis using non-parametric effect sizes with meta-analysis using $StdMD$, we found several practical problems identifying an appropriate meta-analysis method for $StdMD$. In our introduction to this paper, we commented that for large sample sizes and normal data, the standard meta-analysis method was well-understood; however, there are difficulties in applying the standard meta-analysis when sample sizes are small. In particular:

- Luo et al. (2022) found large variations in standardized mean difference ($StdMD$) estimates using different calculation methods on the same experiments, particularly with small sample sizes. It appeared that researchers were unclear about which of the different calculations they should use.
- Lin (2018) found that applying the small sample size adjustment to $StdMD$, could lead to a larger bias in meta-analysis results than aggregating the uncorrected $StdMD$ values. He was unable to specify under which conditions the $StdMD$ estimate should be corrected for small sample bias.
- Kitchenham and Madeyski (2020) reported inconsistencies in the published formulas for the variance of $StdMD$.

In the event of large samples and many independent experiments, disagreements with respect to formulas may not have a major impact on the aggregated values. However, this cannot be guaranteed in the context of small sample sizes and few experiments. We discuss this issue in Section 5.1, and explain our choice of standardized mean difference effect size. We then report our meta-analysis simulations in Section 5.2.

Table 10 Comparison of \hat{p} and Cliff's d power levels

Exp type	Num observations	\hat{p} Power > Cliff's d Power	\hat{p} Power = Cliff's d power	\hat{p} Power < Cliff's d power
Two group	90	88	2	0
Four group	120	106	14	0

Table 11 Comparison of \hat{p} and Cliff's d

Criterion	Exp type	Num observations	Cliff's d	\hat{p}
Median Type 1 Error Rate	Two-Group	30	0.0405	0.0539
Median Type 1 Error Rate	Four-Group	40	0.042	0.0493
Power Difference > 0	Two-Group	90	37	56
Power Difference > 0	Four-Group	120	58	62

5.1 A Meta-Analysis Example

In this section, we present an example that makes the difference between taking the unweighted average and using a meta-analytic weighted approach clearer. Suppose that we have the data reported in Table 12, which is simulated data representing a family of five experiments. The data for each experiment were based on two groups with five observations per group. The control group was sampled from a normal distribution with a mean 0 and variance 1, and the treatment group was sampled from a normal distribution with a mean 0.8 and variance 1. No additional variance was added to simulate random differences between experiments, so, for formal meta-analysis of parametric effect sizes, we used the R language *metafor* package with a fixed effects model (Viechtbauer 2010). For Cliff's d and \hat{p} , we used both methods Kromey investigated, i.e., the unweighted average and formal meta-analysis.

For parametric meta-analysis, we considered eight different possible meta-analysis methods depending on whether or not *StdMD* was adjusted for small sample size, whether the small-sample size formula for the variance of *StdMD* was used or the approximate Normal variance, whether or not the meta-analysis process used weighted or unweighted means, and whether the effect size was calculated after analysing the data as a single large experiment:

1. The *MDUnweighted* method. This is equivalent to meta-analysing the mean difference rather than the *StdMD*. The overall mean difference is the unweighted mean of the mean difference of each experiment, and the overall variance is the mean of the variance of each experiment. The overall *StdMD* is calculated as the overall average divided by the square root of the overall variance.
2. The *StdMDUnweighted* and the *StdMDAdjUnweighted* methods. For these methods, we took the average of *StdMD* and *StdMDAdj* values and calculated the variance directly using the average effect size in the normal variance approximation formula (see the Supplementary Material (Kitchenham and Madeyski 2023))
3. The *ApproxVarWeight* method. This involves aggregating *StdMD* and *StdMDAdj* using the relevant approximate normal variance⁷.
4. The *ExactVarWeight* method. This involves aggregating *StdMD* and *StdMDAdj* using the relevant exact variance.
5. The *HedgesSmallSample* method, which Hedges and Olkin recommend for aggregating results from small sample size experiments (see Hedges and Olkin 1985, Chapter 6, Section F.1).

Formulas for the exact and approximate variance of *StdMD* and *StdMDAdj* are reported in the Supplementary Material (Kitchenham and Madeyski 2023).

⁷ Hedges and Olkin propose another form of approximate variance for *StdMDAdj* (see (Hedges and Olkin 1985, Equation (8), Chapter 5)). However, the alternative formula is closely related to the approximate normal formula, so we do not consider this variance formula in this paper.

From Table 4, we know that the expected values of the effect size are $\hat{p} \approx 0.714$, Cliff's $d \approx 0.428$, and $StdMD = 0.8$. Table 12 reports the effect size statistics obtained from each experiment. It is clear that the effect size estimates are very varied, and only Experiment 4 exhibits values close to the expected values.

Table 13 reports the results of aggregating the effect sizes and their variances and shows that:

- As expected, a formal meta-analysis of Cliff's d and \hat{p} delivers extremely inflated effect sizes while aggregation based on the unweighted means of Cliff's d and \hat{p} gives estimates that are close to the expected values. So, we recommend the meta-analysis method based on unweighted means for the non-parametric effect sizes, and use it for all our meta-analysis simulations. We provide algorithms to perform unweighted meta-analysis for the non-parametric effect in our reproducer R package (Madeyski et al. 2023): `metaanalyse.PHat` and `metaanalyse.Cliffd`.
- All the estimates $StdMD$ and $StdMDAdj$ underestimated the true effect size. The best estimate underestimated by 8%, the worst by 29%.
- Aggregating $StdMD$ and $StdMDAdj$ based on formal meta-analysis was the least accurate of the parametric methods, whether based on their exact or approximate variance.

The systematic underestimation of the parametric effect sizes occurs because, over the set of five experiments, the average $StdMD$ underestimates. However, the small sample size adjustment always *reduces* the $StdMD$ values, and smaller $StdMD$ values have smaller variances⁸ and are given greater weight in the meta-analysis. So, when a set of experiments tends to underestimate, using the small sample size adjustment makes the underestimation worse. Clearly, if a set of $StdMD$ values tends to overestimate, then using the small sample size adjustment will tend to reduce the bias.

In our opinion, the selection of an appropriate parametric meta-analysis method based on the standardized mean difference effect size for small samples is a problem in its own right, and that problem is outside the scope of this study. Therefore, we decided to assess Cliff's d and \hat{p} against the *MDUnweighted* method. This is similar to the Individual Participant Data (IPD) stratified approach described by Santos et al. (2020), where each experiment is analysed as a separate entity. It has the following advantages:

- Our approach is consistent with the method Santos et al. recommend for families of experiments.
- It delays the estimation of the overall effect size and its variance until the best estimates of the overall mean and variance of the data are based on relatively large sample sizes. This means that the small sample size adjustment is not necessary, and the normal approximation to the variance of $StdMD$ can be used.
- It is a similar basic approach to that we use for meta-analysis of the non-parametric effect sizes, so it seems to be a fair comparison method, although it is only likely to be useful in practice for families of experiments.
- We can use parametric analysis methods for both two-group (i.e., the usual R t -test) and four-group designs (i.e., Wilcoxon's `lincon` method with trimming set to zero (Wilcox 2012)), that do not require the assumption of variance heterogeneity. This makes comparisons between the parametric and non-parametric effect sizes fairer.

⁸ Because the variance formula uses the estimate of the effect size.

Table 12 Meta-analysis example data based on a simulation of five two-group experiments each with normal data with variance=1, mean group difference=0.8, group size=5

Exp	Mean Diff	Var	Std MD	df	tval	Sig- nif	Cliff's d	Cliff's d var	$\hat{\rho}$	$\hat{\rho}$ var	$\hat{\rho}$ df	StdMD Adj	StdMDAdj var.exact	StdMDAdj var.approx	StdMD var.exact	StdMD var.approx
1	1.86	1.27	1.65	7.98	2.61	Yes	0.84	0.04	0.92	0.01	6.63	1.49	0.63	0.44	0.77	0.54
2	0.29	1.51	0.24	6.84	0.38	No	0.20	0.18	0.60	0.04	6.63	0.21	0.45	0.32	0.57	0.40
3	-0.39	1.28	-0.35	7.05	-0.55	No	-0.04	0.21	0.48	0.05	5.08	-0.31	0.45	0.32	0.57	0.41
4	0.60	0.68	0.72	6.68	1.14	No	0.44	0.15	0.72	0.04	5.61	0.64	0.49	0.34	0.63	0.43
5	1.44	1.04	1.41	7.00	2.23	Yes	0.76	0.06	0.88	0.01	8.00	1.26	0.61	0.40	0.77	0.51

Table 13 Meta-analysis example results (given the simulation parameters, we expected Cliff's $d \approx 0.429$, $\hat{p} \approx 0.714$, and StdMD= 0.8)

Effect Size (ES)	Method	Mean	Significant	ES Var
Cliff's d	MA (FE)	0.640	TRUE	0.017
Cliff's d	Average	0.440	TRUE	0.026
\hat{p}	MA (FE)	0.827	TRUE	0.004
\hat{p}	Average	0.720	TRUE	0.006
StdMD	Average MD	0.706	TRUE	0.088
StdMD	StdMDUnwghtd	0.736	TRUE	0.088
StdMD	MA (Exact Var)	0.642	TRUE	0.130
StdMD	MA (Approx Var)	0.654	TRUE	0.090
StdMDAdj	MA (Hedges)	0.663	TRUE	0.097
StdMDAdj	StdMDAdjUnwghtd	0.657	TRUE	0.083
StdMDAdj	MA (Exact Var)	0.569	TRUE	0.103
StdMDAdj	MA (Approx Var)	0.579	TRUE	0.071

5.2 Meta-Analysis of Small Sample Size Experiments from Different Distributions

In this section, we evaluate the non-parametric and parametric effect sizes for the meta-analysis of families of experiments using an approach very similar to the method we used for single experiment simulations.

For each meta-analysis, we simulated 10000 families, each comprising five experiments, where each family shared the same properties for its individual experiments in terms of design type (two-group or four-group), observations per group (5, 10, 15 or 20), and mean difference (zero, small, medium or large) for each of the four distribution as defined in Table 3. For Normal and Laplace data samples, we simulated data samples with and without variance heterogeneity. For four-group data samples, we added a fixed block effect for all Normal and Laplace samples, while for log-normal and Gamma distributions, we produced data samples with and without the block effect. Thus, for the individual experiments in a family, we used the same range of simulation conditions as we did for the single group evaluations, except that we reduced the range of sample size because, for meta-analysis, we were only interested in small sample sizes. We used the same evaluation criteria for our meta-analysis simulations as for our individual experiment simulation.

To make the simulations more realistic, we introduced heterogeneity between families. For normal, Laplace and log-normal distributions, when the mean difference was greater than zero, we introduced additional heterogeneity between experiments from different families by adding a small random amount to the control mean for each family. For the gamma distribution, we added the random value to the rate parameter. The random value for a specific family was obtained by generating a random normal variable from a distribution with mean 0 and standard deviation 0.5.

Like for the single experiment simulations, we produced four main results tables: two tables reporting the results for power, small sample bias and individual experiment estimate error for each type of experiment, and two tables reporting the Type 1 error rates for each type of experiment. The results tables can be found in Section 5 of the Supplementary

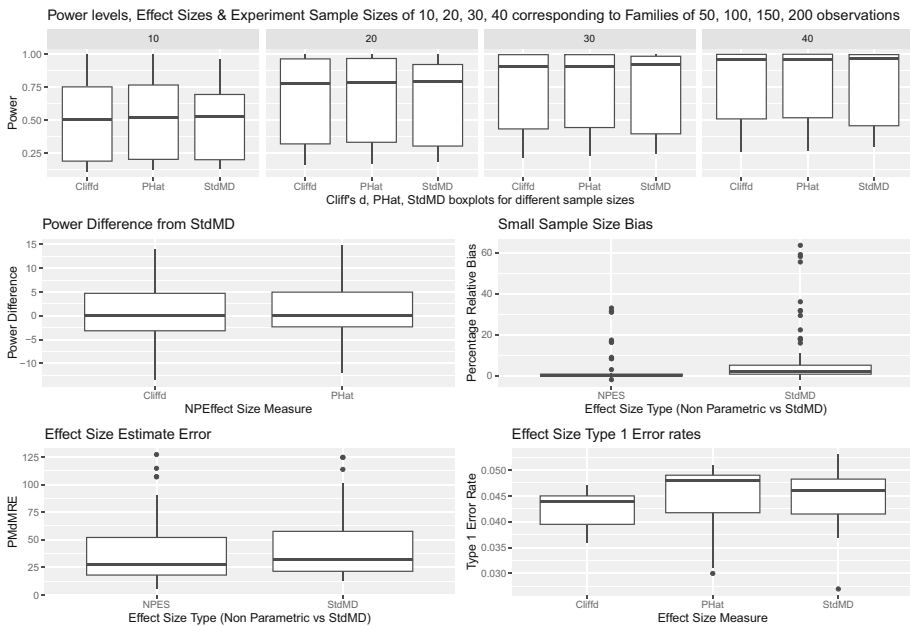


Fig. 6 Two-group meta-analysis simulation results summary (boxplots defined by the x-axis and obtained from all four data distributions)

Material (Kitchenham and Madeyski 2023), and we summarize the results in two multi-pane figures.

The results for families of experiments using two-group designs are shown in Fig. 6 and the results for families of experiments using four-group designs are shown in Fig. 7. As would be expected, the power levels shown in the top pane of both figures are much better for families of experiments than for single experiments. However, families of experiments with small sample sizes and small effect sizes still exhibit unacceptably low power.

The lefthand pane in the middle row in both figures shows that there is not much to choose between the non-parametric effect sizes in terms of power compared with *StdMD*; both are better than *StdMD* in about half of the conditions and worse under the other conditions. Like the single experiment results, the data tables confirm that *StdMD* power is better than the non-parametric effect sizes for the Normal and Gamma samples and worse for the Laplace and log-normal samples. These results are reported in more detail in Section 5.3

Small sample size bias is shown on the righthand pane of the middle row of each figure; again, the non-parametric effect sizes are less biased than *StdMD*, although the effects are not as dramatic as they are for individual experiments, particularly in the case of four-group experiment meta-analysis. The median small sample bias for the two-group experiment meta-analysis was 0.195 for the NP effect sizes and 2.13 for *StdMD*. The median small sample effect size for the four-group experiment meta-analysis was 0.015 for the NP effect sizes and 0.775 for *StdMD*. The direct comparison of small sample bias for each entry in the two-group and four-group data tables is reported in Table 14.

The individual estimate error values measured by *PMdMRE* are shown in each figure's lefthand pane of the bottom row. The median *PMdMRE* values for the two-group experi-

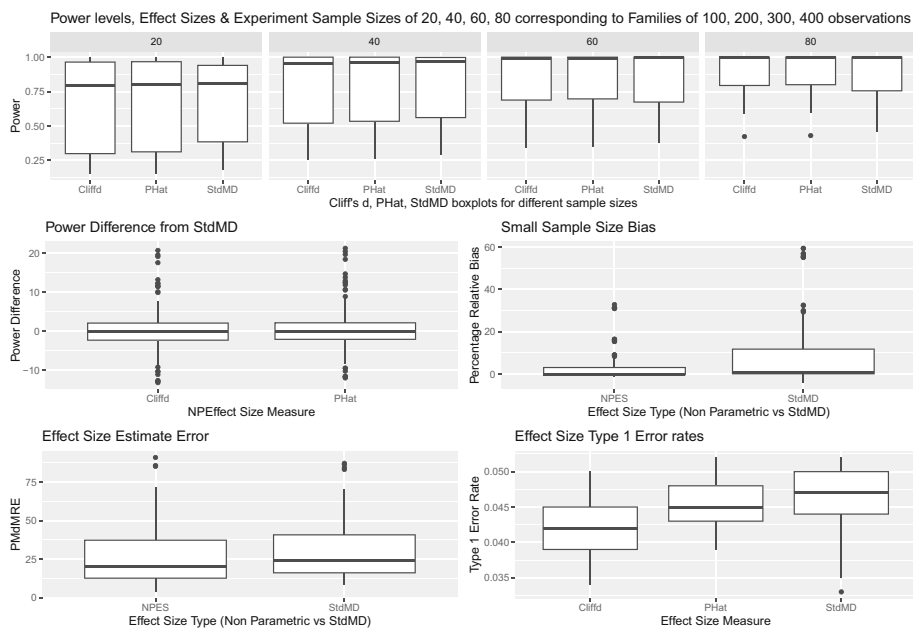


Fig. 7 Four-group meta-analysis simulation results summary (boxplots defined by the x-axis and obtained from all four data distributions)

ment meta-analysis were 28 for the non-parametric effect sizes and 32 for *StdMD*. For the four-group experiment meta-analysis, the median *PMdMRE* values were 20.5 for the non-parametric effect sizes and 24.4 for *StdMD*. The direct comparisons of related *PMdMRE* values from the same simulations are shown in Table 15. The results suggest the NP effect sizes are more accurate than the *StdMD* effect sizes, but the differences are not statistically significant.

The Type 1 error rates are shown in the lefthand pane of the bottom row of each figure. Cliff's *d* exhibits the most biased error rate in both figures, while \hat{p} has the most accurate Type 1 error rates for two-group experiment families, and *StdMD* has the most accurate Type 1 error rates for the four-group experiment families.

The results of the meta-analysis simulations illustrate the value of meta-analysis to increase power and reduce estimate bias. They also confirm the value of the non-parametric effect sizes in reducing small sample bias.

Table 14 Comparison of small sample bias of NP effect sizes and *StdMD* for meta-analysis simulations

Exp type	Num observations	Bias $NP < StdMD$	Bias $NP > StdMD$	Bias $NP = StdMD$
Two Group	72	69	2	1
Four Group	96	75	21	0

Table 15 Comparison of $PMdMRE$ values for NP effect sizes and $StdMD$ for meta-analysis simulations

Exp type	Num observations	$PMdMRE$ $NP < StdMD$	$PMdMRE$ $NP > StdMD$
Two Group	72	57	15
Four Group	96	78	18

5.3 Comparing Non-parametric Effect Sizes Cliff's d and \hat{p}

Table 16 confirms that, for both the two-group and the four-group experiment meta-analysis simulations, \hat{p} power levels were equal or better than Cliff's d power levels for all simulation conditions. Table 17 confirms that \hat{p} exhibited less Type 1 error rate bias for meta-analysis of both two-group families and four-group families. There was no significant difference between Cliff's d and \hat{p} in terms of their power difference effectiveness compared to $StdMD$ power.

6 Discussion

This paper has proposed the use of the non-parametric effects sizes \hat{p} and Cliff's d as effect sizes that are suitable both for summarizing the results of randomized experiments, and for subsequent meta-analysis of independent randomized experiments addressing the same research hypothesis. A novelty of our research is that we confirmed that \hat{p} could be used in meta-analysis like Cliff's d .

Another novelty of this study is that we compared the effectiveness of \hat{p} and Cliff's d . There are some differences in power and Type 1 error rates between \hat{p} and Cliff's d due to the different methods recommended for constructing the effect size variance and for statistical tests. The main difference was that \hat{p} power was always better than, or equal to, Cliff's d power. Reference to the results tables confirms that Cliff's d power is equal to \hat{p} power only when power levels are very high (> 0.98); in all other cases, \hat{p} power is greater than Cliff's d power. Although \hat{p} overestimated Type 1 error rates for two-group experiments with five observations per group while Cliff's d was less biased, across all other simulation conditions, \hat{p} was less biased than Cliff's d . The results, therefore, suggest that overall \hat{p} is more effective than Cliff's d .

This proposal addresses the problem that SE researchers undertaking families of experiments may find if some experiments exhibit non-normality, but meta-analysis of a group of experiments testing the same hypothesis requires all the experiments to be summarized using the same effect size. For example, in Kitchenham et al. (2020a), we investigated 13 studies that applied meta-analysis to families of randomized experiments. Looking at the methods used to analyse individual experiments, we found three studies used only parametric test, whereas four used non-parametric tests only, four used either non-parametric tests or parametric tests depending on the normality of the experimental data, and two studies always

Table 16 Comparison of \hat{p} and Cliff's d Power levels for Meta-Analysis Simulations

Exp type	Num observations	\hat{p} Power > Cliff's d power	\hat{p} Power = Cliff's d power	\hat{p} Power < Cliff's d power
Two Group	72	58	14	0
Four Group	96	57	39	0

Table 17 Comparison of \hat{p} and Cliff's d for meta-analysis

Criterion	Exp type	Num observations	Cliff's d	\hat{p}
Median Type 1 Error Rate	Two-Group Families	24	0.044	0.048
Median Type 1 Error Rate	Four-Group Families	32	0.042	0.045
Power Difference >0	Two-Group	72	37	37
Power Difference >0	Four-Group	96	57	57

used both non-parametric and parametric statistical tests. However, in spite of most of the studies using non-parametric statistics tests, all 13 studies used parametric effect sizes for their meta-analyses. Our simulation results confirm that researchers can use \hat{p} both for summarizing the results of individual experiments and for meta-analyzing the results of families of experiments.

Our simulations have also confirmed that the power levels for both the parametric and non-parametric effect sizes are unacceptably low for single experiments with small sizes, particularly when effect sizes are small. This has been long recognised by SE researchers and had led to calls for using larger sample sizes (see Shepperd 2018; Jørgensen et al. 2016). Furthermore, for many years the Simula Laboratory in Norway pioneered the use of large sample size experiments using professional software engineers as participants (see, for example, Arisholm and Sjøberg 2004; Arisholm 2006; Arisholm et al. 2007). However, the mapping study by Santos et al. (2020) and our own research (Kitchenham et al. 2022) makes it clear that academic researchers have found that using families of experiments is easier to manage than undertaking single large-scale experiments. Fortunately, our simulations also show that high power levels can be achieved when experimental results are aggregated using meta-analysis. However, our simulations also confirm that for small effect sizes, aggregating the *StdMD* can lead to biased estimates of the overall effect size, while aggregating \hat{p} results in much less bias⁹.

Our results show that there is a potential for power loss using \hat{p} to summarize individual experiments if the data is known to normal or gamma. However, our simulations confirm that \hat{p} estimates for individual experiments are unbiased for all distributions. This means the \hat{p} estimates from individual experiments can be easily aggregated with other experiments testing the same hypothesis, without any of the adjustments required for the standardized mean difference effect sizes needed to adjust for its small sample bias.

We have demonstrated two other advantages that \hat{p} has over *StdMD*, in cases when samples are small and the distribution of the samples is unknown:

1. \hat{p} is robust to extreme non-normality, such as that arising for log-normal samples with variance heterogeneity.
2. The meta-analysis process is much simpler for \hat{p} than for *StdMD* because it does not require the data analyst to make a large number of arbitrary decisions about the method of constructing the effect sizes and their variance.

A criticism of using \hat{p} as an effect size is that it may be difficult to understand. However, it may also be difficult to understand the standardized mean difference or the point bi-serial correlation coefficient. Furthermore, in the context of academic experiments, which are based

⁹ Inspection of the meta-analysis result tables provided in the Supplementary Material confirms that both parametric and nonparametric effect sizes are most likely to be biased for gamma-distributed data.

on participants (often students) all undertaking the same task in a restricted timescale, we should not expect estimates of *StdMD* or *StdMDAdj* to be representative of the effects that would be found in an industrial context. It may be more useful to report \hat{p} , which gives a good indication of whether or not a technique improves the performance of individuals (even students) than a numerical estimate of the effect that is unlikely to be realistic.

6.1 Limitations and Constraints

A general limitation of simulation experiments is that they can only consider a limited number of conditions. Our simulation studies included both normal and non-normal distributions with different non-normal properties, two design types, a variety of small sample sizes and limited variance heterogeneity. We have not used mixed distributions, which have higher probabilities of outliers, nor have we used artificially truncated distributions, where data values are restricted to be within finite upper and lower bounds, although, in principle, more complex data sets should favour the use of non-parametric effect sizes (Neuhäuser et al. 2007). In addition, we have not considered unbalanced experiments where there are different numbers of participants in different treatment groups and blocks, nor have we considered the impact of families of less than 5 experiments.

A specific limitation of our study is that the meta-analysis method we used for *StdMD* was recommended for families of experiments, not a meta-analysis of sets of independent experiments (Santos et al. 2020). However, the method that we used for both parametric and non-parametric meta-analysis is not actually invalid for any set of independent experiments; it is just equivalent to always choosing a random effects model. This is likely to be less powerful for *StdMD* than a standard meta-analysis when a fixed effects analysis is appropriate or when it is possible to obtain a reliable estimate of the excess variance due to experiment heterogeneity.

Our proposals are restricted to:

- experimental designs that can be decomposed into independent two-group experiments, which include randomized between-groups experiments, randomized blocks experiments and within-participant before/after experiments, and cross-over experiments. We explicitly exclude randomized factorial experiments that investigate the interaction effects between different techniques. In this case, there is no well-defined summary effect size for such experiments.
- outcome measures of ratio or interval scale but not binary outcome measures or short ordinal scale measures. This is because binary measures and ordinal measures cannot be converted into useful rank statistics. Binary outcome measures all share one of two ranks, and ordinal scale measures share a limited number of ranks.

6.2 Future Work and Conclusions

For future research, we need to trial our proposals on a variety of different SE data sets to ensure that our method and analysis tools are appropriate for experiments with unequal treatment groups and block sizes and perform as we anticipate for cross-over designs.

We present our key findings in the following textbox, which confirm that there are good reasons for using the non-parametric effect size \hat{p} for both statistical analysis and meta-analysis of randomized experiments with small sample sizes.

Key findings:

1. For individual experiments, the non-parametric effect sizes (Cliff's d and \hat{p}) had negligible small sample bias for all the combinations of sample sizes and distributions that we simulated. In contrast, *StdMS* exhibited a substantial small sample bias across the range of distributions and sample size distributions.
2. The non-parametric effect sizes delivered power levels that were better than *StdMD* for lognormal and Laplace data but marginally worse for gamma and normal data.
3. For meta-analysis, the method (Kromrey et al. 2005) propose for meta-analysis with Cliff's d can also be applied to \hat{p} , and it can also be used for non-parametric analysis of any single experiment that can be decomposed into blocks of one or more two-group random experiments.
4. For meta-analysis, the non-parametric effect sizes exhibit less small sample bias than *StdMD*.
5. Across all but one simulation condition, \hat{p} type 1 error rates were less biased than Cliff's d type 1 error rates and across all conditions, \hat{p} power was as good or better than Cliff's d power.

Overall conclusion: Using \hat{p} as an effect size is a low-risk option for analysing and meta-analysing data from small sample-size experiments. Parametric methods are only preferable if you have prior knowledge of the distribution of the data.

We hope that researchers who have published meta-analyses based on parametric effect size for lack of any alternative will *re-do* their meta-analysis using \hat{p} and publish any changes to their previous conclusions.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10664-024-10504-1>.

Acknowledgements The authors thank Prof. Wilcox for providing the source of R functions for robust statistical analysis. The paper is the result of the research internship of Lech Madeyski at Keele University at the invitation of Prof. Kitchenham and, partly, the research internship of Lech Madeyski at BTH at the invitation of Prof. Mattsson and at Lancaster University at the invitation of Prof. Hall.

Author Contributions **Barbara Kitchenham:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Visualization. **Lech Madeyski:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization.

Data Availability The paper is backed by the reproduction package (*reproducer* (Madeyski et al. 2023)) written in R and available from CRAN, the official repository of R packages.

Declarations

Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give

appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Acion L, Peterson JJ, Temple S, Arndt S (2006) Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics Med* 25:591–602
- Acuña ST, Gómez MN, Hannay JE, Juristo N, Pfahl D (2015) Are team personality and climate related to satisfaction and software quality? aggregating results from a twice replicated experiment. *Inf Softw Technol* 57(1):141–156
- Arcuri A, Briand L (2014) A hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering. *Softw Testing, Verification Reliability* 24(3):219–250. <https://doi.org/10.1002/stvr.1486>
- Arisholm E (2006) Empirical assessment of the impact of structural properties on the changeability of object-oriented software. *Inf Softw Technol* 48(11):1046–1055
- Arisholm E, Sjöberg DI (2004) Evaluating the effect of a delegated versus centralized control style on the maintainability of object-oriented software. *IEEE Trans Softw Eng* 30(8):521–534
- Arisholm E, Gallis H, Dyba T, Sjöberg DI (2007) Evaluating pair programming with respect to system complexity and programmer expertise. *IEEE Trans Softw Eng* 33(2):65–86
- Basili V, FShull, Lanubile E, (1999) Building knowledge through families of experiments. *IEEE Trans Softw Eng* 25(4):456–473. <https://doi.org/10.1109/32.799939>
- Borenstein M, Hedges LV, Higgins JPT, Rothstein HT (2009) *Introduction to Meta-Analysis*. John Wiley and Sons Ltd
- Brunner E, Munzel U (2000) The nonparametric Behrens-fisher problem: asymptotic theory and a small-sample approximation. *Biometrical J* 42:17–25. [https://doi.org/10.1016/S0378-3758\(02\)00269-0](https://doi.org/10.1016/S0378-3758(02)00269-0)
- Brunner E, Munzel U, Puri ML (2002) The multivariate nonparametric Behrens-fisher problem. *J Statistical Plan Inference* 108(1–2):37–53. [https://doi.org/10.1016/S0378-3758\(02\)00269-0](https://doi.org/10.1016/S0378-3758(02)00269-0)
- Ciolkowski M (2009) What do we know about perspective-based reading? an approach for quantitative aggregation in software engineering. In: *Proceedings of the 2009 3rd international symposium on empirical software engineering and measurement*, IEEE Computer Society, Washington, DC, USA, ESEM '09, pp 133–144. <https://doi.org/10.1109/ESEM.2009.5316026>
- Cliff N (1993) Dominance statistics: ordinal analyses to answer ordinal questions. *Psychological Bulletin* 114(3):494–509
- Cohen J (1992) A power primer. *Psychological Bulletin* 112(1):155–159
- Curtin F, Altman DG, Elbourne D (2002) Meta-analysis combining parallel and cross-over clinical trials. I: continuous outcomes. *Statistics Med* 21:2132–2144. <https://doi.org/10.1002/sim.1205>
- Derrick B, Broad A, Toher D, White P (2017) The impact of an extreme observation in a paired samples design. *Adv Methodol & Statistics/Methodološki Zvezki* 14(2)
- Faraone SV (2008) Interpreting estimates of treatment effects. *Pharmacy Therapeutics* 22(12):627–633
- García S, Fernández A, Luengo J, Herrera F (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. *Inf Sci* 180(10):2044–2064. <https://doi.org/10.1016/j.ins.2009.12.010>
- Hedges LV, Olkin I (1983) Nonparametric estimators of effect size in meta-analysis. Tech. Rep. Technical Report No. 193, Department of Statistics, Stanford University
- Hedges LV, Olkin I (1985) *Statistical methods for meta-analysis*. Academic Press, Orlando, Florida, USA
- Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ* 327(7414):557–560. <https://doi.org/10.1136/bmj.327.7414.557>
- Jørgensen M, Dybå T, Liestøl K, Sjöberg DI (2016) Incorrect results in software engineering experiments: How to improve research practices. *J Syst Softw* 116:133–145. <https://doi.org/10.1016/j.jss.2015.03.065>
- Jureczko M, Madeyski L (2015) Cross-project defect prediction with respect to code ownership model: an empirical study. *e-Informatica Softw Eng J* 9(1):21–35. <https://doi.org/10.5277/e-Inf150102>
- Kitchenham B, Madeyski L (2020) Inconsistencies with formulas for the standard error of the standardized mean difference of repeated measures experiments. *Statistics Med* 39:4101–4104

- Kitchenham B, Madeyski L (2023) Supplementary Material for the paper "Recommendations for Analysing and Meta-Analysing Small Sample Size Experiments". <https://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiRAMASSSEsupplement.pdf>
- Kitchenham B, Madeyski L, Budgen D, Keung J, Brereton P, Charters S, Gibbs S, Pohthong A (2017) Robust statistical methods for empirical software engineering. *Empirical Softw Eng* 22(2):579–630. <https://doi.org/10.1007/s10664-016-9437-5>
- Kitchenham B, Madeyski L, Curtin F (2018) Corrections to effect size variances for continuous outcomes of cross-over clinical trials. *Statistics Med* 37(2):320–323. <http://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiCurtinSIM.pdf>
- Kitchenham B, Madeyski L, Brereton P (2019) Problems with statistical practice in human-centric software engineering experiments. In: Proceedings of the evaluation and assessment on software engineering, ACM, New York, USA, EASE '19, pp 134–143. <https://doi.org/10.1145/3319008.3319009>, <https://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiBreretonEASE19.pdf>
- Kitchenham B, Madeyski L, Brereton P (2020) Meta-analysis for families of experiments in software engineering: a systematic review and reproducibility and validity assessment. *Empirical Softw Eng* 25(1):353–401. <https://doi.org/10.1007/s10664-019-09747-0>
- Kitchenham B, Madeyski L, Scanniello G, Gravino C (2020b) Supplementary material to the paper "The Importance of the Correlation in Crossover Experiments". <https://doi.org/10.5281/zenodo.4475865>
- Kitchenham B, Madeyski L, Scanniello G, Gravino C (2022) The importance of the correlation in crossover experiments. *IEEE Trans Softw Eng* 48(8):2802–2813. <https://doi.org/10.1109/TSE.2021.3070480>
- Kraemer H, Andrews G (1982) A non-parametric technique for meta-analysis effect size calculation. *Psychological Bulletin* 91:404–412
- Kromrey JD, Hogarty KY, Ferron JM, Hines CV, Hess MR (2005) Robustness in meta-analysis: an empirical comparison of point and interval estimates of standardized mean differences and Cliff's delta. In: Proceedings of the joint statistical meetings, Minneapolis
- Lin L (2018) Bias caused by sampling error in meta-analysis with small sample sizes. *PLoS ONE* 13(9). <https://doi.org/10.1371/journal.pone.0204056>
- Long JD, Cliff N (1997) Confidence intervals for Kendall's tau. *British J Math Statistical Psychol* 50(1):31–41
- Luo Y, Funada S, Yoshida K, Noma H, Sahker E, Furukawa TA (2022) Large variation existed in standardized mean difference estimates using different calculation methods in clinical trials. *J Clinical Epidemiol* 149:89–97. <https://doi.org/10.1016/j.jclinepi.2022.05.023>
- Madeyski L, Jureczko M (2015) Which process metrics can significantly improve defect prediction models? An Empirical Study. *Softw Quality J* 23(3):393–422. <https://doi.org/10.1007/s11219-014-9241-7>
- Madeyski L, Kitchenham B (2017) Would wider adoption of reproducible research be beneficial for empirical software engineering research? *J Intell & Fuzzy Syst* 32:1509–1521. <https://doi.org/10.3233/JIFS-169146>
- Madeyski L, Kitchenham B (2018) Effect sizes and their variance for AB/BA crossover design studies. *Empirical Softw Eng* 23(4):1982–2017. <https://doi.org/10.1007/s10664-017-9574-5>
- Madeyski L, Orzeszyna W, Torkar R, Józala M (2014) Overcoming the equivalent mutant problem: a systematic literature review and a comparative experiment of second order mutation. *IEEE Trans Softw Eng* 40(1):23–42. <https://doi.org/10.1109/TSE.2013.44>
- Madeyski L, Kitchenham B, Lewowski T (2023) reproducer: Reproduce Statistical Analyses and Meta-Analyses. <https://cran.r-project.org/web/packages/reproducer/reproducer.pdf>, R package
- McGraw K, Wong S (1992) A common language effect size statistic. *Psychological Bulletin* 111:361–265
- Morales JM, Navarro E, Sánchez-Palma P, Alonso D (2016) A family of experiments to evaluate the understandability of TRiStar and i* for modeling teleo-reactive systems. *J Syst Softw* 114:82–100
- Neuhäuser M, Lösch C, Jöckel KH (2007) The Chen-Luo test in case of heteroscedasticity. *Comput Statistics & Data Anal* 51:5055–5060
- Rahlf s VW, Zimmermann H, Lees KR (2013) Effect size measures and their relationships in stroke studies. *Stroke* 45:627–633
- Ripley BD (2006) *Stochastic Simulation*. Wiley
- Santos A, Gómez O, Juristo N (2020) Analyzing families of experiments in SE: a systematic mapping study. *IEEE Trans Softw Eng* 46(5):566–583. <https://doi.org/10.1109/TSE.2018.2864633>
- Senn S (2002) *Cross-over Trials in Clinical Research*, 2nd edn. Wiley
- Shepherd M (2018) Replication studies considered harmful. In: Proceedings of the 40th international conference on software engineering: new ideas and emerging results, Association for Computing Machinery, New York, USA, ICSE-NIER '18, pp 73–76. <https://doi.org/10.1145/3183399.3183423>
- Varga A, Delany HD (2000) A critique and improvement of the common language effect size statistics of McGraw and Wong. *J Educ Behavioral Statistics* 25(2):101–132

- Vegas S, Apa C, Juristo N (2016) Crossover designs in software engineering experiments: benefits and perils. *IEEE Trans Softw Eng* 42(2):120–135. <https://doi.org/10.1109/TSE.2015.2467378>
- Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. *J Statistical Softw* 36(3):1–48. <https://doi.org/10.18637/jss.v036.i03>
- Welch B (1938) The significance of the difference between two means when the population variances are unequal. *Biometrika* 29(3/4):350–362
- Wilcox RR (2012) *Introduction to Robust Estimation & Hypothesis Testing*, 3rd edn. Elsevier

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.