Supplementary Material for the paper "Recommendations for Analysing and Meta-Analysing Small Sample Size Software Engineering Experiments"

Barbara Kitchenham and Lech Madeyski

Abstract—This document provides supplementary material to support the paper entitled "Recommendations for Analysing and Meta-Analysing Small Sample Size Software Engineering Experiments". The supplementary material is available online from https://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiRAMASSSEsupplement.pdf
If the paper is accepted for publication, this Supplementary Material will be made publicly available on Zenodo to ensure more reliable access and better support reproducible research.

Index Terms—evidence-based software engineering, meta-analysis, effect size, nonparametric, crossover design, reproducible research in software engineering

1 Introduction

In this document, we report statistical methods and approaches that support the results reported in [1].

In Section 2, we discuss the origin of nonparametric effect sizes and provide a detailed discussion of the two nonparametric effect sizes we evaluate in [1]: i.e., the probability of superiority $\hat{p}$ and Cliff's d. We explain how these effect sizes and their variances are estimated. We also explain how the problems that occur if the effect size variance is zero[1] are addressed.

In Section 3 we discuss how to calculate $\hat{p}$ and Cliff's d and their variances for different statistical designs. In particular, we discuss randomized block designs and three variants of repeated measures designs, i.e., the pre/post-test control design, the AB/BA crossover, and the four-group crossover design. An important issue is that the method for analyzing randomized blocks is the basis both for the analysis of all designs more complex than a simple between-groups design and for the method used to aggregate nonparametric effect sizes from different experiments.

In Section 4, we discuss the functional forms of the four distributions we use in our simulations: the normal (or more formally, the Gaussian) distribution, the log-normal distribution, the gamma distribution, and the Laplace distribution. We discuss the relationship between the parameters of each distribution and the mean and variance of samples from each distribution. We define the values of the parameters needed to produce values of (0.2, 0.5, 0.8) for the parametric standardized mean difference effect size of the generated data sets. We also report the expected effect sizes of Cliff's d and $\hat{p}$ for various values of the distribution parameters. The normal distribution is the standard distribution against which all our simulation results are assessed. The other distributions have various different non-normal properties that allow us to assess the value of the nonparametric effect sizes when we cannot assume normality.

In Section 5 we present the tables of results for our simulations of four-group randomized block experiments and for our meta-analysis of families of experiments. The procedures used to generate the tables are discussed in our related main text [1].

An initially unanticipated problem associated with evaluating the use of Cliff's d and $\hat{p}$ because it was not obvious how to provide a fair parametric analysis as a basis for comparisons with the nonparametric effect size analyses. The problem arose because there are many different methods for calculating the variance of a standardized mean difference effect size and aggregating results from a set of comparable experiments (such as a family of experiments [2]). To illustrate the extent of the problem, we identified a range of methods for aggregating $StdMD$, which in turn relied on identifying the methods used to estimate the variance of the different $StdMD$ estimates.

In Section 6, we discuss options available to estimate the variance of different estimates of $StdMD$. The formulas depend on the experimental design, whether or not the standard mean difference was adjusted for a small sample size, and whether or not the formula is the exact variance or the approximate normal variance. Readers should note that to emphasize the difference between the two parametric effect sizes in the main text [1], we refer to the standardized mean difference as $StdMD$, and the small sample size adjusted standardized mean difference as $StdMDAdj$. However, to

- *B. Kitchenham is with the School of Computing and Mathematics, Keele University, Keele, Staffordshire, ST5 5BG, UK.*
- *Lech Madeyski is with the Wroclaw University of Science and Technology, Wyb.Wyspianskiego 27, 50370 Wroclaw, Poland. E-mail: Lech.Madeyski@pwr.edu.pl*

1. Zero variances happen if observations in one treatment group are all larger than the observations in the other treatment group.

make the equations presented in Section 6 easier to read, we use the symbol d to refer to $StdMD$ and the symbol $\hat{\delta}$ to refer to $StdMDAdj$.

In Section 7, we discuss the various ways in which the standardized mean difference estimates can be aggregated and the related overall variance calculated.

Finally, we discuss various issues related to simulation studies, including :

- Unexpected failures when generating and analysing very small samples.
- The method adopted for tests of significance when using one-sided and two-sided tests.
- How the simulation functions we developed were used to construct the tables in the main text [1].

. Researchers wishing to use the simulation functions should ensure that they have installed the latest version of `reproducer` [3]. Any old version of reproducer should be uninstalled using the `R` command: `remove.packages("reproducer")`. The latest version of `reproducer` can then be installed and used.

2 Robust Effect Sizes and Nonparametric Analysis Methods

In this section, we provide a detailed discussion of Cliff's d and the probability of superiority $\hat{p}$.

2.1 Underlying Principles

Both Cliff's d and $\hat{p}$ can be defined in terms of three probabilities:

1) p_1, which is the probability that a random observation from a participant in group $G1$ is greater than a random observation from a participant in group $G2$.
2) p_2, which is the probability that a random observation from a participant in group $G1$ is equal to a random observation from a participant in group $G2$
3) p_3, which is the probability that a random observation from a participant in group $G1$ is less than a random observation from a participant in group $G2$.

Since these three probabilities comprise all possibilities for the relationship between observations in the two groups:

$$p_1 + p_2 + p_3 = 1 \tag{1}$$

Robust effect sizes have an implied direction as well as magnitude, and the direction depends on whether we are testing whether observations from G1 are greater than observations from G2 or vice versa. Thus, when aggregating such effect sizes, it is important that they are based on effect sizes that all make the same comparison. To avoid overloading numerical indexes, we will assume that values from G1 correspond to the *alternative* software engineering technique and are referred to as the a-values, and values from G2 correspond to the *control* or baseline software engineering technique and are referred to as c-values. If we subscript an effect size with $_{ac}$, we identify an effect size that specifies the extent to which the alternative technique outperforms the control, while a subscript of $_{ca}$ means we are looking at the extent to which the control technique outperforms the alternative technique[2].

To introduce the methods for calculating nonparametric effect sizes based on p_i and test their significance, we use the data shown in Table 1. These are hypothetical data from two independent groups where group G1 corresponds to the treatment (i.e., alternative) condition, and group G2 corresponds to the control condition. The example is a subset of data from a real experiment that we adapted to have one pair of duplicate observations. The response variable data are in column *G1 Data* and *G2 Data*. The *G1 Rank* and *G2 Rank* variables are the rank values for the combined G1 and G2 data set. The purpose of this example is to demonstrate how to calculate $\hat{p}$ and Cliff's d, and to explain the relationship between the effect sizes. We do not recommend such small samples for real experiments!

TABLE 1
Example Data

Group	*G1 Data*	*G1 Rank*	Group	*G2 Data*	*G2 Rank*
G1	0.24	12	G2	-0.02	4
G1	0.06	9	G2	-0.24	3
G1	0.03	6.5	G2	0.03	6.5
G1	-0.33	1	G2	0.15	11
G1	-0.26	2	G2	0.09	10
G1	0	5	G2	0.04	8

2.2 The Probability of Superiority

McGraw and Wong [4] proposed an effect size based on p_1, which they called the common language effect size. Subsequent researchers criticized this effect size because it was based on the assumption that there were no tied values [5]. Hence, the probability of superiority $\hat{p}_{ac}$ is now defined as:

$$\hat{p}_{ac} = p_1 + \frac{p_2}{2} \tag{2}$$

In addition:

$$\hat{p}_{ca} = p_3 + \frac{p_2}{2} = 1 - \hat{p}_{ac} \tag{3}$$

and

$$\hat{p}_{ca} + \hat{p}_{ac} = 1 \tag{4}$$

These equations provide a means of coping with duplicate values and confirm that $\hat{p}_{ca}$ and $\hat{p}_{ac}$ are probabilities that vary from 0 to 1. If $\hat{p}_{ca} \approx \hat{p}_{ac} \approx 0.5$, we conclude that there is no difference between the observations in the two groups. We consider later how this conclusion can be formally tested.

To calculate the p_i values, we need to count the total number of times each value in G1 was less than, equal to, or greater than a value in G2. For example, the value -0.33 in G1 is less than all six *Data* values in group G2 (see Table 1).

The counting process can be understood using a *superiority* matrix such as that shown in the top section of Table 2[3], where the top row shows the participants' values in G2 and the first column shows the participants' values in G1. The internal superiority matrix displays a "1" if the G1

2. We omit the subscript if there is no ambiguity in our equation.
3. Wilcox's `R` functions for calculating both $\hat{p}$ and Cliff's d and their variances are based on constructing the superiority matrix [6].

TABLE 2
Superiority Matrix

G2:	−0.02	−0.24	0.03	0.15	0.09	0.04	$Positive_j$	$Equal_j$	$Negative_j$	Sum_j	$Mean_j$
G1:											
0.24	1	1	1	1	1	1	6	0	0	6	1
0.06	1	1	1	−1	−1	1	4	0	2	2	0.3336
0.03	1	1	0	−1	−1	−1	2	1	3	−1	−0.1667
−0.33	−1	−1	−1	−1	−1	−1	0	0	6	−6	−1
−0.26	−1	−1	−1	−1	−1	−1	0	0	6	−6	−1
0	1	1	−1	−1	−1	−1	2	0	4	−2	−0.3333
$Positive_i$	4	4	2	1	1	2					
$Equal_i$	0	0	1	0	0	0					
$Negative_i$	2	2	3	5	5	4					
Sum_i	2	2	−1	−4	−4	−2					
$Mean_i$	0.3333	0.3333	−0.1667	−0.6667	−0.6667	−0.3333					

participant value is greater than a G2 participant value, "0" if the two values are equal, and "−1" if the G1 value is less than the G2 value. The values in this matrix are referred to as d_{ij} where i refers to the row and j refers to the column. For each column and each row of the superiority matrix, we show the number of positive, negative, and equal values, and the Sum and Mean values. The sample estimates of p_i values are obtained by dividing the total number of positive (for p_1), equal (for p_2), and negative values (for p_3) by the total number of elements in the matrix. Thus:

$$p_1 = \frac{14}{36} = 0.3889 \tag{5}$$

$$p_2 = \frac{1}{36} = 0.02778 \tag{6}$$

$$p_3 = \frac{21}{36} = 0.5833 \tag{7}$$

Then we have:

$$\hat{p}_{ac} = 0.3889 + 0.02778/2 = 0.4028$$

$$\hat{p}_{ca} = 0.5833 + 0.02778/2 = 0.5972$$

and

$$\hat{p}_{ac} + \hat{p}_{ca} = 0.4028 + 0.5972 = 1$$

For the example data, $\hat{p}_{ac} < 0.5$, the calculated effect size suggests that the control method has outperformed the alternative method. We discuss later in this section how to assess whether or not the performance improvement is statistically significant.

Although $\hat{p}$ can be easily calculated from p_i, it can also be calculated by ranking the data. This is because the ranking information measures the number of values that a random value exceeds. For example, across all the *Data* values in Table 1, the value -0.33 is the smallest and, as shown in the table, is given a rank 1 of 12, while the value 0.24 is the largest and is given a rank of 12. If there are duplicated values, they are given the appropriate *midrank*. For example, we have two 0.03 values that should be assigned ranks 6 and 7, so both numbers are assigned the mid-rank 6.5, and the rank values 6 and 7 are considered to be used.

The value of $\hat{p}$ can then be calculated from the difference between the average ranks:

$$\hat{p}_{ac} = \frac{\hat{R}_1 - \hat{R}_2}{N} + 0.5 \tag{8}$$

if ranks are assigned across the values in both groups, $\hat{R}_1$ is the average rank of the observations in G1, and $\hat{R}_2$ is the average rank of the observations in G2, and N is the total number of observations. For our example data, $\hat{R}_1 = 5.9167$ and $\hat{R}_2 = 7.0833$, so:

$$\hat{p}_{ac} = \frac{5.9167 - 7.0833}{12} + 0.5 = -0.09722 + 0.5 = 0.4028 \tag{9}$$

It is important to note that under the null hypothesis $\hat{p} = 0.5$, so it is usual to base tests of the significance of $\hat{p}$ on $\hat{p} - 0.5$, which Rahlfs et al. [7] refer to as the Mann-Whitney centred statistic or the Average Risk Difference.

As we noted in a previous paper [8], the commonly used Mann-Whitney-Wilcoxon rank test for two independent groups and the Kruskal-Wallis rank test for differences among more than two groups are not robust unless based on exact permutation statistics. Treating ranks as if they are random variables for the purpose of constructing significance tests is inadvisable because, if the null hypothesis is false, the variance of the rank averages in each group will be significantly different. Brunner and Munzel [9] proposed a method that explicitly allows for the heterogeneity caused by differences between the variances for each group. It is based on Welch's test [10] and provides a significance test that $\hat{p}$ is significantly different from 0.5 and assumes that the variance of $\hat{R}_1$ and the variance of $\hat{R}_2$ are *not* equal. Wilcox provides an implementation of the Brunner and Munzel method (i.e., the `bmp` function) that calculates the standard error and confidence interval upper and lower bounds for $\hat{p}$. The formulas to construct the variance of $\hat{p}$ and its confidence interval can be found in Section 2.4.

The first row of Table 3 shows the estimate of $\hat{p}$ together with its standard error and its 95% upper and lower confidence interval bounds (CIB). Since the confidence interval of $\hat{p}$ spans 0.5, we *cannot* reject the null hypothesis of no difference between the groups.

TABLE 3
Analysis of Example Data

Metric	*Estimate*	*Standard Error*	*Upper 95% CIB*	*Lower 95% CIB*
$\hat{p}$	0.4028	0.1836	0.8259	0
Cliff's d	-0.1944	0.3730	0.4652	-0.7152

2.3 Cliff's d

Given the definitions of p_1, p_2, and p_3, Cliff's d is defined as:

$$d_{ac} = p_1 - p_3 = \hat{p}_{ac} - \hat{p}_{ca} \tag{10}$$

or:

$$d_{ca} = p_3 - p_1 = -d_{ac} \tag{11}$$

Also, given Equation (4):

$$d_{ac} = 2\hat{p}_{ac} - 1 = 1 - 2\hat{p}_{ca} \tag{12}$$

Since $0 \leq \hat{p}_{ac} \leq 1$, we have $-1 \leq d \leq 1$. If $d \approx 0$, then we conclude that there is no difference between the two groups. It should be noted that in the literature, d is seldom sub-scripted; however, it has a direction that needs to be specified to make sure that the results are not misinterpreted.

Moreover, if the elements of the superiority matrix are defined to be d_{ij}, Cliff's d can be estimated directly as:

$$d_{ac} = \frac{\sum_i \sum_j (d_{ij})}{n_1 n_2} \tag{13}$$

With the values in Table 2, $d = (14 - 21)/36 = -0.1944444$. Also, from Equation (10), d_{ac} is estimated to be $\hat{p}_{ac} - \hat{p}_{ca} = 0.4027778 - 0.5972222 = -0.1944444$. The negative value for d_{ac} indicates that, in this case, the control technique has outperformed the alternative technique. In addition, d can be calculated from the mean of the row means (in Table 2) or the mean of column means. The formula needed to construct the variance of Cliff's d can be found in Section 2.4 together with the non-standard formula Cliff recommends for constructing confidence intervals around d.

The second row of Table 3 shows the estimate of Cliff's d for the data in Table 1. Since the confidence interval spans zero, we *cannot* reject the null hypothesis that there is no difference between the groups.

In this section, we have explained that Cliff's d is functionally related to $\hat{p}$. However, the variances of the two effect sizes do not use the same method of constructing confidence intervals, so it is possible for the significance tests for each effect size to deliver different results.

2.4 Nonparametric Effect Size Variances

This section specifies the formulas for constructing the variances of nonparametric effect sizes and the methods used to construct confidence intervals.

2.4.1 The Variance of $\hat{p}$

Wilcox [6] reports how to calculate the variance of $\hat{p}$ starting from ranks of the pooled data and the mean rank for each group, i.e.:

$$\overline{R}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} R_{ij} \tag{14}$$

where j is the group identifier i.e., 1 or 2 for a simple between-groups experiment, n_j is the number of participants in the group j and $N = \sum_{i=1}^{2} n_i$.

The ranks for each group, ignoring the other groups, also need to be calculated and are referred to as V_{ij}, where i refers to the rank of an individual observation in the group j. Then, for two-group experiments, the variance of $\hat{p}$ is

$$varp = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right) \tag{15}$$

where

$$s_j^2 = \frac{S_j^2}{(N - n_j)^2} \tag{16}$$

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(R_{ij} - V_{ij} - \overline{R}_j + \frac{n_j + 1}{2} \right)^2 \tag{17}$$

The test statistic is:

$$W = \frac{\overline{R}_2 - \overline{R}_1}{\sqrt{varp}} \tag{18}$$

and the degrees of freedom for the $t-$test are $\hat{v} = \frac{U_1}{U_2}$, where

$$U_1 = \left(\frac{S_1^2}{n_2} + \frac{S_2^2}{n_1} \right)^2 \tag{19}$$

and

$$U_2 = \frac{1}{n_1 - 1} \left(\frac{S_1^2}{n_2} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{S_2^2}{n_1} \right)^2 \tag{20}$$

The approximate $(1 - \alpha)$ interval for $\hat{p}$ is $\hat{p} \pm t \times \sqrt{varp}$, where t is the $1 - \alpha/2$ quartile of a Student's t distribution with degrees of freedom $\hat{v}$.

2.4.2 The Variance of Cliff's d

The variance of d is calculated using the superiority row means and column means:

$$\overline{d}_{.j} = \frac{1}{n_1} \left(\sum_{i=1}^{n_1} d_{ij} \right) \tag{21}$$

and

$$\overline{d}_{i.} = \frac{1}{n_2} \left(\sum_{j=1}^{n_2} d_{ij} \right) \tag{22}$$

Then we need to calculate three variance components, referred to as s_1^2, s_2^2, and $\tilde{\sigma}^2$.

$$s_1^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\overline{d}_{.j} - d)^2 \tag{23}$$

$$s_2^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\overline{d}_{i.} - d)^2 \tag{24}$$

$$\tilde{\sigma}^2 = \frac{1}{n_1 n_2} \sum_i \sum_j (d_{ij} - d)^2 \tag{25}$$

Then, $\hat{\sigma}^2$ (i.e., the variance of d) is

$$\hat{\sigma}^2 = \frac{(n_2 - 1)s_1^2 + (n_1 - 1)s_2^2 + \tilde{\sigma}^2}{n_1 n_2} \tag{26}$$

Instead of the more usual confidence limits, Cliff recommends using the following equation for the lower bound of the confidence interval:

$$CI_{lb} = \frac{d - d^3 + z\hat{\sigma}\sqrt{(1 - d^2)^2 + z^2\hat{\sigma}^2}}{1 - d^2 + z^2\hat{\sigma}^2} \tag{27}$$

where z is the *lower* quantile of the normal distribution for the appropriate significance level. The following equation specifies the upper bound:

$$CI_{ub} = \frac{d - d^3 - z\hat{\sigma}\sqrt{(1-d^2)^2 + z^2\hat{\sigma}^2}}{1 - d^2 + z^2\hat{\sigma}^2} \quad (28)$$

2.4.3 A Limitation of Cliff's d and the Probability of Superiority $\hat{p}$

As mentioned by Neuhäuser et al. [11], there is a problem estimating the variance of Cliff's d and $\hat{p}$ for small sample sizes. When all the results from the treatment group are greater than all the results from the control group, all the values in the superiority matrix will be 1 and $d = 1$ and $\hat{p} = 1$. While such an outcome might be called a *perfect* result by an experimenter, it is a problem for statistical analysis because the estimated variance of Cliff's d will be 0. In addition, although the variance of $\hat{p} = 1$ is calculated slightly differently, the estimated variance of $\hat{p}$ is always zero if $\hat{p} = 0$ or $\hat{p} = 1$.

This is a problem because in Section 3 we demonstrate how to apply nonparametric effect sizes to statistical designs more complex than simple between-groups studies, but our method depends on having a valid variance for the nonparametric effect size in each subset of the data. Our method of addressing this problem is explained below and is illustrated with a real software engineering data set.

An example of the issue is shown in the software data set shown in Table 4. This example arose in a pilot study of distributed and face-to-face software architect evaluation meetings [12], which used an AB/BA crossover design (see [13], Section 4.2.2). Looking at the time period difference (see the column labelled TP Diff), a significant difference between the results for crossover groups A and group B with a positive effect size would confirm that when teams met in distributed meetings, the results of their evaluation were better than when they met in face-to-face meetings.

TABLE 4
Data Set with maximum Cliff's d

CO Group	Team ID	Treatment used first	TP1 score	TP2 score	TP Diff	Rank
A	G2	F2F	47	87	40	7
A	G5	F2F	88	96	8	5
A	G6	F2F	59	73	14	6
B	G1	Dist	67	27	-40	1
B	G3	Dist	85	70	-15	2
B	G4	Dist	79	74	-5	4
B	G7	Dist	65	59	-6	3

In this situation, Wilcox's algorithm calculates the upper and lower confidence for Cliff's d based on the binomial distribution and uses the relationship between d and $\hat{p}$ to calculate the upper and lower bounds for $\hat{p}$. Unfortunately, the binomial distribution method leads to large confidence intervals for small sample sizes, which can cause inconsistent results. For example, using the example shown in Table 4, Wilcox's `cid` algorithm reported that Cliff's=1, with a standard error of zero and confidence interval [-0.054,1]. Because the confidence interval spans zero, the binomial test suggests that the null hypothesis cannot be rejected. However, suppose the TP Diff value for team G4 had been 8 rather than -5, this would correspond to the minimum possible overlap between values, which in terms of rank would mean that teams G4 and G5 would both be allocated a rank of 4.5. From the point of view of the experimenter, this would be a *nearly perfect* result. In this case, Wilcox's algorithm reports Cliff's d=0.92, with a standard error of 0.0168 and a confidence interval [0.313,0.993]. Because the interval does not span 0, this would imply that the null hypothesis can be rejected. Thus, the levels of significance in the two cases are contradictory because when there is no overlap between the groups, the algorithm implies that the result is not significant, but when there is some overlap, and the value of Cliff's d is lower, the algorithm implies that the difference is significant. In addition, the length of the confidence interval for the nearly perfect case is $0.993 - 0.313 = 0.68$, while the length of the confidence interval for the perfect case is 1.054. Although for perfect experimental outcomes, we only get contradictions with respect to significance for sample sizes of 8 observations or less, the length of the confidence interval remains much larger than the equivalent confidence interval for nearly perfect experiments.

In Section 3, we demonstrate how to apply nonparametric analysis methods to statistical designs more complex than simple between-groups studies, but our method depends on having a valid variance for d in each subset of the data. In order to cope with the problem of perfect experiments, we decided to use the variance of the equivalent nearly perfect experimental outcome. So in the case of the perfect outcome discussed above, we would assume that $d = 1$, with confidence limits [0.313,1] (i.e., we extend the upper bound of the confidence interval to include the value of d, but the lower bound of the confidence interval remains greater than 0) and we assume that the standard error d is approximately 0.0168. This means that our estimate of the standard error of d is conservative for perfect experiments, but it is not zero, and our decisions regarding the significance of perfect and nearly perfect experiments are consistent.

3 Calculating Nonparametric Effect Sizes for Different Experimental Designs

A nonparametric analysis is usually applied to k-group between groups randomized experiments, where the only difference between the k groups of experimental units (which, in the case of SE experiments, would often be human participants) is the experimental technique that participants in each group used. The experimental design discussed in Section 2 has $k = 2$. However, if nonparametric effect sizes are to be useful, they must be applicable to more complex experimental designs too.

Vegas et al. [14] reviewed experimental papers from six top-ranked software engineering sources and identified 82 papers that reported 124 SE experiments with human participants in the years 2012 to 2014 inclusive. 38 experiments were classified as independent measures (i.e., between-groups experiments), 68 were crossover designs, 16 were defined as repeated measures[4] and two were matched pairs.

4. Formally, crossovers are a specific type of repeated measures design.

This study revealed the importance of the crossover design in software engineering and, thus, the need for methods of how to apply nonparametric effect sizes to this form of design.

In this section, we explain how to analyze data from randomized blocks (which are a form of a between-group experiment) and repeated measures designs, including crossover designs. We are particularly interested in repeated measures designs because one of the motivations for this paper was to investigate the use of nonparametric effect sizes in the context of families of experiments, and families of experiments often use repeated measures designs. The importance of randomized block experiments is that they are the building blocks we need to develop procedures for handling four-group duplicated crossover design experiments.

3.1 Analysing Randomized Blocks Experiments

Randomized block experiments are experiments in which a blocking factor is used either to extend the generality of the experiment or to reduce extraneous variation. For example, if we want to extend the generality of a code reading experiment, we might use two different programs and two different code reading techniques and assign each participant to one of the four different conditions (i.e., program 1 with technique 1, program 1 with technique 2, program 2 with technique 1, program 2 with technique 2). If we are concerned with reducing extraneous variation, we might assign participants to two blocks on the basis of skill and then randomize the assignment of participants in each block to each of the two different techniques.

One way to analyze such a design using a nonparametric approach is to use a within-block analysis. This means finding the value of an NP (i.e., nonparametric) effect size for the two treatment groups in each block and taking the average of the two values, i.e.,

$$\overline{NPES} = \frac{NPES_1 + NPES_2}{2} \tag{29}$$

$NPES_1$ is the estimate of a specific nonparametric effect size for block 1, and $NPES_2$ is the estimate of the nonparametric effect size for block 2. Based on the following two standard statistical results for independent variables x and y and a constant c:

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

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

we can use the variance of the NPES for each treatment in each block to estimate the pooled variance of $\overline{NPES}$:

$$var(\overline{NPES}) = \frac{var(NPES_1) + var(NPES_2)}{4} \tag{30}$$

If the NP effect size is Cliff's d, we calculate the variance of each group using Equation (26). If the NP effect size is $\hat{p}$, then calculate the variance of each group using Equation (15). The pooled variance can then be used to construct the confidence intervals on $\overline{NPES}$. For Cliff's d, we propose using the equations for confidence limits reported in Equation (27) and Equation (28). For $\hat{p}$, we propose using the sum of the degrees of freedom from each block to construct the usual $t-$distribution-based confidence intervals.

We have provided a function `Calc4GroupNPStats` in our `reproducer` R package [3] to analyze such experiments, assuming two treatment conditions and two blocking conditions.

3.2 Repeated Measures Experiments

In a review of the meta-analysis methods used in families of experiments [15], we found that the majority of families used repeated measures designs for individual experiments in a family. In particular, they used AB/BA crossover designs, 4-group AB/BA crossover designs, and one pre-test/post-test control design. Although many authors used nonparametric tests for some or all of their individual experiments, they did not discuss the implications of their basic design.

In this section, we show that if *difference* measures are used, both the AB/BA crossover design and the pre-test/post-test control design can be analyzed as simple between-group experiments, while the 4-group AB/BA design can be analyzed as a randomized block experiment. This means they can be analysed using the non-parametric methods that we propose.

3.2.1 Pre-test/Post-test Control Design

For a pre-test/post-test control design, participants are randomly allocated to two groups. In the first phase of the experiment, referred to as TP1, participants perform one or more software engineering tasks using the standard/control technique and the same software engineering materials, M1 (e.g., a computer program or a software design document), and a response value is measured (e.g., the time to complete the tasks, the number of correct answers to a comprehension questionnaire, or the number of defects detected). After a training period, in a second experimental session, referred to as $TP2$, participants in group G1 use the new technique to perform equivalent tasks on the second set of software engineering materials (M2) and participants in group G2 use the same technique they used in the first session with materials M2, and the same response variable is measured. Table 5 reports the expected response outcome values for a participant y_j in group $G1$ and participant y_k in group $G2$, for session TP1 and TP2 . We make the standard linear assumptions that the response values for each participant will:

- Reflect the specific skill of each participant, which is modelled as the term μ_j for participant j.
- Include an adjustment related to the specific session, which is considered to be zero for the first session and to be a constant P in the second session and is the same for participants in each group.
- Include a term that specifies the effect of using technique A or technique B. Participants in group G1 use technique A in session 1 and technique B in session 2. Participants in group G2 use technique A in both sessions.
- Include an adjustment related to the specific software engineering materials, i.e., M1 in session 1 and M2 in session 2.

If we consider the difference (D) between the expected values for each participant (i.e., the outcome value in session 2 minus the outcome value in session 1), for G1 we have:

$$D_{1j} = \tau_B - \tau_A + P + M2 - M1 \quad (31)$$

while for G2 we have:

$$D_{2k} = \tau_A - \tau_A + P + M2 - M1 = P + M2 - M1 \quad (32)$$

Therefore, if we conduct a nonparametric analysis of the *difference* values in the two groups, we can test whether the term $\tau_b - \tau_a$ is different from zero. If the confidence interval includes zero, we cannot reject the hypothesis that there is no difference between the techniques. Otherwise, if the lower confidence interval bound is greater than zero, technique B outperforms technique A, and if the upper confidence interval bound is less than zero, technique A outperforms technique B. This is simply a nonparametric between-group analysis.

3.2.2 AB/BA Crossover Design

The basic design for an AB/BA crossover is similar to that of the pre-test/post-test control, but instead of using the same technique as group 1 participants in session 1, participants in group G2 use technique B and swap to technique A in session 2 (see Table 6). In addition, a term λ_A may be added to the participant effect in G1 to model any interaction effect due to using method A before using method B, and a term λ_B may be added to the participant effect in group G2 to model the equivalent effect when method B is used first. In practice, we assume that the interaction terms are zero. Senn points out that the design is inappropriate if the interaction terms are nonzero [16]. The expected values for this design are shown in Table 6. Statistical tests of significance (both parametric and nonparametric) are based on analyzing the difference of the differences, which corresponds to testing whether or not $\tau_B - \tau_A = 0$

Both the pre-test/post-test control design and the AB/BA crossover design are more powerful than designs without repeated measures because the variability due to different participants is removed. However, an important difference between the AB/BA crossover design and the pre-test/post-test control design is that the difference for AB/BA crossover design is $\tau_B - \tau_A - (\tau_A - \tau_B) = 2(\tau_B - \tau_A)$. Thus, the crossover is more powerful than the pre-test post-test design. However, the pre-test post-test design avoids the issue of spurious interactions between method and order.

For non-parametric analysis of crossover designs, Senn recommended using a Mann-Whitney between-groups rank test on the difference data [16]. However, Cliff's d or $\hat{p}$ are a more robust basis for statistical tests [6]. Again, constructing the confidence intervals for the NP effect sizes allows us to assess whether there is a significant difference between the techniques and, if the confidence interval does not include zero, whether technique A outperforms technique B or vice versa.

3.2.3 Four-Group Crossover Design

As discussed in [15], the 4-group crossover is currently popular among SE researchers. The design is basically a duplicated AB/BA crossover, with the difference between the two individual AB/BA crossovers being the order in which the software engineering materials are used. We show the design in Table 7 along with the expected difference values (assuming no interaction effects).

The difference values show that Groups 1 and 2 constitute one AB/BA crossover, and Groups 3 and 4 constitute another. In each pair, the only difference between the expected values in each group is that one includes the term $\tau_A - \tau_B$ and the other includes the term $\tau_B - \tau_A$. Thus, the difference values in Groups 1 and 2 represent one block in a randomized block design, and the difference values in Groups 3 and 4 can represent the other block in a two-way randomized block design.

Then, if we calculate the specific NP effect size for G1 and G2, and the equivalent effect size for G3 and G4, and their respective variances, we can use Equation (29) to calculate the mean value of the NP effect size for the experiment and Equation (30) to calculate its variance.

3.2.4 The Meaning of Repeated Measures Effect Sizes

When analyzing repeated measures data sets, it is important to recall that there are two different types of standardized effect size (see [17] and [13]):

1) The standardized mean difference effect size of the *average personal improvement* which is calculated from difference values.
2) The standardized mean difference effect size of the *difference between treatments*, either based on the first time period data only or based on the average effect for both time periods).

We have concentrated on the analysis of difference values in order to confirm that the nonparametric effect sizes can be used for these more complex designs. However, it is important to recognize that when effect sizes are aggregated, you must always aggregate the same type of effect size. Furthermore, assuming that there is a positive correlation between repeated measures, you should expect the magnitude of the nonparametric effect sizes based on a personal improvement to be larger than the magnitude of the nonparametric effect sizes based on the difference between treatments. This is the same as the effect for parametric effect sizes and arises for the same reason. When there is a positive correlation between repeated measures, taking the difference removes some of the variability due to differences between participants. This makes systematic differences due to differences between treatments easier to detect.

4 The Distributions Used in the Simulation Studies

Our simulation studies reported in [1] were based on four different distributions: the normal distribution, log-normal distribution, the gamma distribution, and the Laplace distribution. We discuss the distributions in more detail in the following sections

Examples of the four distributions discussed in this section are shown in Figure 1. For the normal, log-normal and Laplace distributions, we used $\mu = 0$ and a spread parameter of 1, for the gamma distribution, we used a shape parameter equal to 3 and a rate parameter equal to 1. We show sample sizes of 40 and 1000 for each distribution.

TABLE 5
Expected Outcome for Participants in a Pre-test/Post-test Control Design

Sequence Group	Participant ID	Session TP1	Session TP2	Difference
$G1$	j	$y_{1,1,j} = \mu_j + \tau_A + M1$ (technique A)	$y_{2,2,j} = \mu_j + P + \tau_B + M2$ (technique B)	$P + \tau_B - \tau_A + M2 - M1$
$G2$	k	$y_{2,1,k} = \mu_k + \tau_A + M1$ (technique A)	$y_{1,2,k} = \mu_k + \tau_A + P + M2$ (technique A)	$P + M2 - M1$

TABLE 6
Expected Outcome for Participants in an AB/BA Crossover Design

Sequence Group	Participant ID	Session TP1	Session TP2	Difference (Assuming $\lambda_A = \lambda_B = 0$)
$G1$	j	$y_{1,1,j} = \mu_j + \tau_A + M1$ (technique A)	$y_{2,2,j} = \mu_j + P + \tau_B + M2 + \lambda_A$ (technique B)	$P + \tau_B - \tau_A + M2 - M1$
$G2$	k	$y_{2,1,k} = \mu_k + \tau_B + M1$ (technique B)	$y_{1,2,k} = \mu_k + \tau_A + P + M2 + \lambda_B$ (technique A)	$P + \tau_A - \tau_B + M2 - M1$

TABLE 7
Expected Outcome for Participants in an 4-Group Crossover Design

Sequence Group	Participant ID	Session TP1	Session TP2	Difference
$G1$	j	$y_{1,1,j} = \mu_j + \tau_A + M1$ (technique A)	$y_{2,2,j} = \mu_j + P + \tau_B + M2$ (technique B)	$P + \tau_B - \tau_A + M2 - M1$
$G2$	k	$y_{2,1,k} = \mu_k + \tau_B + M1$ (technique B)	$y_{1,2,k} = \mu_k + \tau_A + P + M2$ (technique A)	$P + \tau_A - \tau_B + M2 - M1$
$G3$	l	$y_{1,1,l} = \mu_l + \tau_A + M2$ (technique A)	$y_{2,2,l} = \mu_l + P + \tau_B + M1$ (technique B)	$P + \tau_B - \tau_A + M1 - M2$
$G4$	m	$y_{2,1,m} = \mu_m + \tau_B + M2$ (technique B)	$y_{1,2,m} = \mu_m + \tau_A + P + M1$ (technique A)	$P + \tau_A - \tau_B + M1 - M2$

We report the sample statistics in Table 8. Skewness is a measure of lack of symmetry. It is formally defined to be the third moment about the mean. Kurtosis is formally defined as the fourth moment about the mean. It was originally defined in terms of the peakedness and extent of outliers. Nowadays, it is defined in terms only of the preponderance of outliers.

Although we used similar parameter values for normal, log-normal and Laplace distibutions, they exhibited very different mean and variance estimates on the raw data scale. This motivated us to to chose parameters for the non-normal distribution simulations in [1] that delivered mean values and variance that were, for large samples, equal to the expected mean and variance of the normal simulations.

It is clear from these results that small sample frequency plots and parameter estimates can differ widely from those of large samples.

4.1 The Normal Distribution

Most standard statistical analyses and tests (e.g., $t-$tests and analysis of variance and co-variance) assume the underlying distribution of a data sample is normal. The normal distribution (also called the Gaussian) is defined by the mean μ and variance σ^2. The mean and variance of a random sample from a normal population are unbiased estimates of μ and σ^2. The standardized normal distribution has $\mu = 0$ and $\sigma^2 = 1$.

The functional form of the normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{33}$$

where e is Euler's constant and π is the constant of the same name. The standardised normal distribution has $\mu = 0$ and $\sigma^2 = 1$ and has the functional form:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \tag{34}$$

Important properties of the normal distribution are:

- If we have a sample from a normal distribution, the mean and variance of the sample are estimates of the parameters μ and σ^2.
- The parameters μ and σ^2 are independent, i.e., changing one of the parameter values does not cause a change in the other parameter.
- Any sample of data from a normal distribution can be standardized by subtracting the population mean from each value and dividing by the population standard deviation (σ). In practice, we subtract the

TABLE 8
Sample Statics for Simulated Data from Four Distributions

Run ID	Mean	Median	Variance	Skewness	Kurtosis	Outliers
Normal40	-0.128	0.028	0.957	0.356	2.394	0
Normal1000	-0.033	-0.020	0.963	0.093	2.949	6
LogNormal40	1.477	1.061	2.304	-2.316	9.027	3
LogNormal1000	1.606	1.077	3.355	-3.621	23.985	78
Gamma40	3.535	3.367	3.001	-0.931	3.550	1
Gamma1000	3.063	2.692	3.350	-1.153	4.734	27
Laplace40	-0.745	-0.206	2.242	1.312	4.410	3
Laplace1000	0.010	-0.026	2.142	-0.107	4.486	62

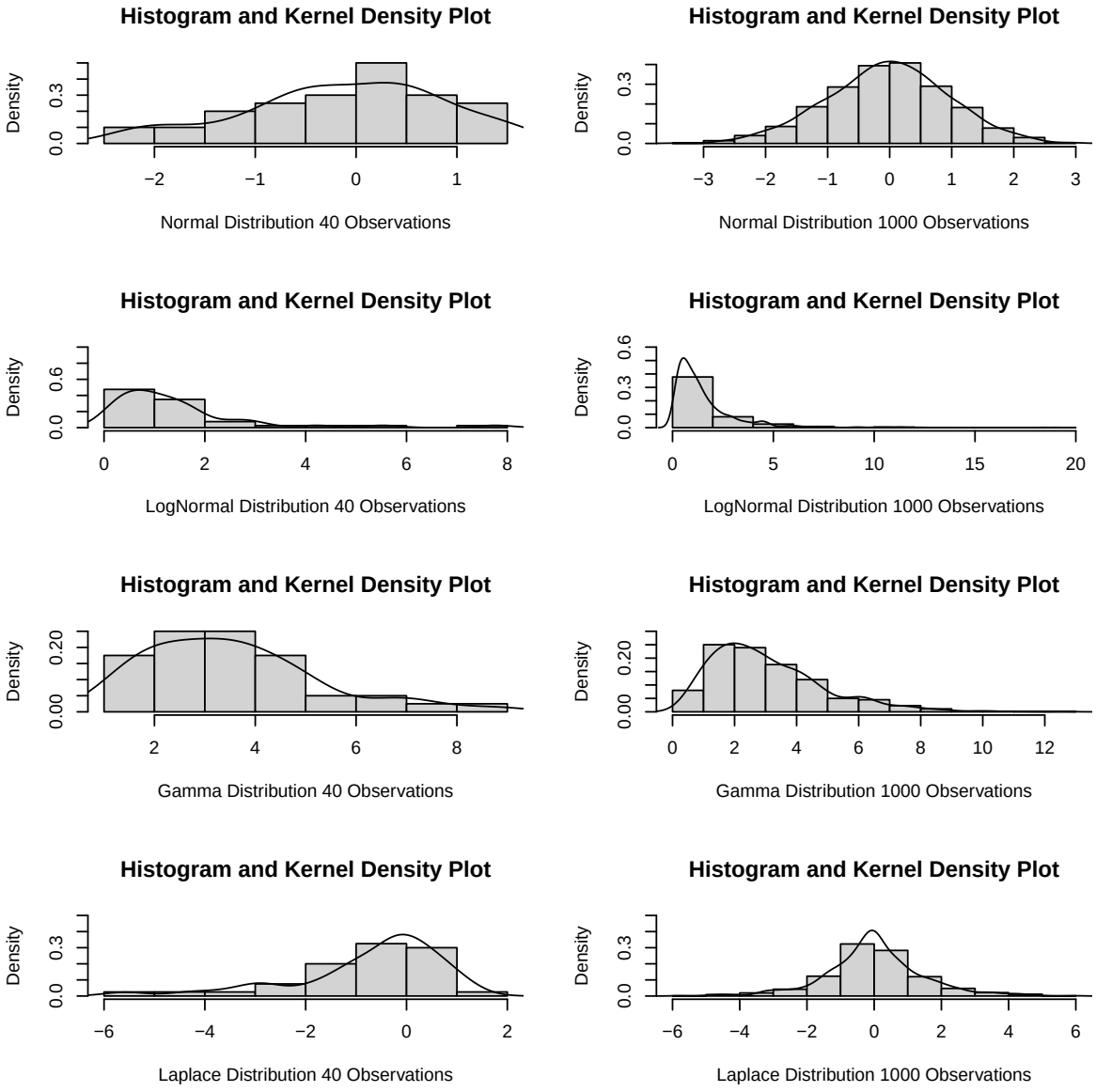

Fig. 1. Histograms and Kernel Density Plots for Simulated Data from Four Distributions

sample mean and divide it by the sample standard deviation.

- The normal distribution is symmetric about its central value, so the mean and median of the distribution are equal. Skewness is a measure of lack of symmetry, so the theoretical skewness of the normal distribution is zero. Skewness is formally defined to be the third moment about the mean, and sample

skewness is calculated as:

$$\sum_{i}^{n} \frac{(x-\bar{x})^3}{n} \tag{35}$$

- The theoretical kurtosis of the normal distribution is 3. Kurtosis is formally defined as the fourth moment about the mean, and sample kurtosis is calculated as:

$$\sum_{i}^{n} \frac{(x-\bar{x})^4}{n} \tag{36}$$

For our simulation studies, we considered normal populations from two-group randomized experiments and four-group randomized block experiments. As is customary for statistical simulation studies, we simulated samples from standardized normal distributions with mean values of 0, 0.2, 0.5, and 0.8 to investigate the Type 1 error rate and the power of the effect sizes. These values are based on Cohen's observations of the standardized mean differences found in psychological studies, where 0.2 is considered a small effect size, 0.5 is a moderate effect size, and 0.8 is a large effect size.

For a standard normal distribution, it is simple to calculate the *theoretical* standardized mean difference effect size ($StdMD = \frac{\mu_d}{\sigma}$). In the case of a two-group randomized experiment which has one treatment group and one control group, the variance is not affected by the treatment if both groups have mean zero, the $StdMD = 0$. For any other difference between two groups, because $\sigma^2 = 1$, the theoretical $StdMD$ value equals that difference. In the cases of a four-group experiment, where each block comprises two groups, one representing the control condition and the other the treatment condition, and the variance is unaffected by the block or the treatment, the mean difference (μ_d is unaffected by the block because:

$$\mu_d = \frac{\mu_t - \mu_c + \mu_t + b - (\mu_c + b)}{2} \tag{37}$$

However, it is not impossible for the variance to be affected by the treatment. For example, a new software engineering method could exacerbate the difference between skilled and less skilled individuals, which would increase the variance among participants. Alternatively, it could increase the performance of less skilled individuals while not affecting the performance of skilled individuals, which would decrease the variance between participants. In either case, the variance of one group will be changed, as will the theoretical variance of the mean difference. Again we can calculate the theoretical $StdMD$, for example, for a two-group randomized block design:

$$\sigma^2 = \frac{(\sigma_t^2 + \sigma_c^2)}{2} \tag{38}$$

and the $StdMD$ will be calculated as:

$$StdMD = \frac{\mu_t - \mu_c}{\sigma} \tag{39}$$

We summarize the results of randomized experiments (i.e., two-group experiments) in Table 9, which show that increasing the variance of the treatment condition decreases the theoretical value of $StdMD$. We summarize the results of the randomized block (i.e., four-group experiments) in Table 10. We show that a blocking effect alone has no impact on the theoretical value of $StdMD$. Again, if the treatment condition leads to a change in the variance, the theoretical $StdMD$ will be decreased.

TABLE 9
Theoretical Effect Sizes for a Normal Distribution Two-Group Experiment

μ_c	σ_c	μ_t	σ_t	μ_d	σ^2	$StdMD$
0	1	0	1	0	1	0
0	1	0.20	1	0.2	1	0.2
0	1	0.5	1	0.5	1	0.5
0	1	0.8	1	0.8	1	0.8
0	1	0	1.5	0	1.620	00
0	1	0.2	1.5	0.2	1.620	0.157
0	1	0.5	1.5	0.5	1.620	0.392
0	1	0.8	1.5	0.8	1.620	0.628

TABLE 10
Theoretical Effect Sizes for Normal Distribution Four-Group Experiments

μ_c	σ_c	μ_t	σ_t	BE	μ_d	σ^2	$StdMD$
0	1	0.0	1.0	0.0	0.000	1.000	0.000
0	1	0.2	1.0	0.0	0.200	1.000	0.200
0	1	0.5	1.0	0.0	0.500	1.000	0.500
0	1	0.8	1.0	0.0	0.800	1.000	0.800
0	1	0.0	1.5	0.0	0.000	1.620	0.000
0	1	0.2	1.5	0.0	0.200	1.620	0.157
0	1	0.5	1.5	0.0	0.500	1.620	0.392
0	1	0.8	1.5	0.0	0.800	1.620	0.628
0	1	0.0	1.0	0.5	0.000	1.000	0.000
0	1	0.2	1.0	0.5	0.200	1.000	0.200
0	1	0.5	1.0	0.5	0.500	1.000	0.500
0	1	0.8	1.0	0.5	0.800	1.000	0.800
0	1	0.0	1.5	0.5	0.000	1.620	0.000
0	1	0.2	1.5	0.5	0.200	1.620	0.157
0	1	0.5	1.5	0.5	0.500	1.620	0.392
0	1	0.8	1.5	0.5	0.800	1.620	0.628

Thus, if we know the distribution a sample arises from, we can observe how close the estimates of parameters and effect sizes are to their theoretical values. This means that simulation studies based on a specific experimental design and a specific distribution can be used to investigate the accuracy of effect size estimates for different sample sizes. In addition, since we can simulate multiple samples from the same distribution and experimental design, we can also investigate:

- the power of statistical tests to detect non-zero effect sizes for different sample sizes,
- the rate of Type 1 errors for zero effect sizes which should estimate the α level of the test irrespective of sample size.

To compare the effectiveness of standardized effect size for the two-group randomized experiment with the effectiveness of Cliff'd d and $\hat{p}$, we need to know the theoretical value of the non-parametric effect sizes for specific distributions and experimental designs. However, there is no defined relationship between the parameters of a theoretical distribution and nonparametric effect sizes constructed from

random samples. Instead, we calculated large sample values of the effect sizes.

Since effect sizes get closer to their theoretical value as sample size increases, we used extremely large samples to obtain estimates of the large sample nonparametric effect sizes. To do this:

1) We generated a data set comprising two groups of independent normally distributed data with ten million observations per group.
2) We calculated the unstandardized mean difference effect size (MD), the combined variance (Var), the standardized mean difference ($StdMD$), Cliff's d, and $\hat{p}$ for the data set.
3) We did this for each combination of μ_c, σ_c, μ_t, σ_t shown in Table 9.

This process generated the results shown in Table 11. The values correspond to the conditions shown in Table 9. The estimates of $StdMD$ from the large sample are close to the theoretical $StdMD$ values; this means that the sample size is large enough to approximate the theoretical value to within three decimal places. So, we assume that the large sample size values of $\hat{p}$ and Cliff's d will also be close to the population values of the nonparametric effect sizes.

Like $StdMD$, $\hat{p}$ and Cliff's d are reduced if the variance of the treatment condition is increased. This makes sense because the nonparametric effect sizes are inversely proportional to the extent to which the control and treatment values overlap[5] and increasing the variance of the treatment group will increase the likelihood of the observations from the different groups overlapping.

TABLE 11
Large Sample Estimates of Effect Sizes for Normally Distributed Two-Group Randomised Experiments

Additional Variance	$\hat{p}$	Cliff's d	MD	Var	$StdMD$
No	0.500	0.000	0.001	1.00	0.001
No	0.556	0.113	0.200	1.000	0.200
No	0.638	0.276	0.500	1.000	0.500
No	0.714	0.428	0.800	1.000	0.800
Yes	0.500	-0.000	-0.001	1.625	-0.001
Yes	0.544	0.088	0.200	1.625	0.157
Yes	0.609	0.219	0.501	1.626	0.393
Yes	0.671	0.343	0.801	1.625	0.628

We used a similar process to simulate a four-group randomized block experiment based on four-group data sets. However, we also investigated the impact of introducing a non-zero blocking effect which was added to the mean values of the control and treatment means in one of the blocks. The results of this analysis are shown in Table 12. The values for all the effect sizes reported in Table 11 are extremely close to the equivalent values reported in Table 12, whether or not the simulation introduced a blocking effect. This confirms that blocking has no significant impact on effect sizes obtained using normal data.

5. That is, the more the observations overlap, the closer the effect sizes are to the null hypothesis condition.

TABLE 12
Large Sample Estimates of Effect Size for Normally Distributed Four-Group Randomised Experiments

Additional Variance	Block Effect	$\hat{p}$	Cliff's d	Md	Var	$StdMD$
No	No	0.500	-0.000	-0.000	1.000	-0.000
No	No	0.556	0.113	0.200	1.000	0.200
No	No	0.638	0.276	0.499	1.000	0.499
No	No	0.714	0.428	0.800	1.000	0.800
No	Yes	0.500	-0.000	-0.000	1.000	-0.000
No	Yes	0.556	0.112	0.200	1.000	0.200
No	Yes	0.638	0.276	0.500	1.000	0.500
No	Yes	0.714	0.429	0.800	1.000	0.800
Yes	No	0.500	-0.000	-0.001	1.625	-0.001
Yes	No	0.544	0.088	0.200	1.626	0.157
Yes	No	0.609	0.218	0.500	1.625	0.392
Yes	No	0.671	0.343	0.800	1.625	0.628
Yes	Yes	0.500	0.000	0.000	1.625	0.000
Yes	Yes	0.544	0.089	0.200	1.625	0.157
Yes	Yes	0.609	0.219	0.501	1.625	0.393
Yes	Yes	0.671	0.343	0.800	1.626	0.628

4.2 The Log-Normal Distribution

The log-normal distribution defines a function that is normally distributed after a logarithmic transformation. The log-normal distribution is often used in software cost estimation studies to normalize effort and size data and enable the relationship between effort and size to be represented by a linear equation.

The functional form of the log-normal distribution is:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}} \tag{40}$$

where μ and σ are the mean and variance of the related normal distribution, e is Euler's constant and π is the constant of the same name, and ln is the logarithm with base e.

The `R` system generates log-normal data by specifying the mean and standard deviation of the log-transformed data. If we generate log-normal data with mean=μ and variance=σ^2, the mean and variance of the *raw* data will be:

$$mean(X) = e^{\mu+\frac{\sigma^2}{2}} \tag{41}$$

$$variance(X) = (e^{\sigma^2} - 1) \times e^{2\mu+\sigma^2} \tag{42}$$

Equations (41) and (42) show the most important property of the log-normal distribution, that is a functional relationship between the mean and the variance of data samples from that distribution.

This means that if we increase the value of μ, we will increase both the mean of the raw data and the variance of the raw data. Similarly, increasing σ^2, also increases both the mean and variance of the raw data. Thus, any significant mean difference effect size will cause variance heterogeneity, and a fixed block effect will increase the variance of the data in the affected block. In addition, if we apply the same values we used for the normal simulations to simulate the log-normal data, we do not obtain the same effect size on the raw data scale.

In order to make our simulation results more comparable for different distributions, we set the parameters for the log-normal simulations to values that would generate effect

sizes of (0, 0.2, 0.5, and 0.8) on the *raw* data scale. This is shown in rows 1 to 4 of Table 13, which shows the theoretical effect of comparing two groups, with parameters μ_c and σ_c being the mean and standard deviation of the normal distribution used to generate the control group data and μ_t and σ_t being the mean and standard deviation of normal distribution used to generate the treatment group data. The column μ_d reports the theoretical mean difference that should be found on the raw data scale, the column σ^2 reports the average variance of the control and treatment data, $StdMD$ reports the standardized mean difference effect size. The table shows that if we use $\mu_t = 0.2665$, the theoretical standardized effect size should be 0.2[6].

Rows 1 to 4 in Table 13 show that increasing the mean difference between the treatment and control means substantially increases the variance on the raw data scale. Rows 5 to 8 in the table show that if we use the same values of μ_t but increase the value of σ_t from 1 to 1.5, we increase both the mean difference and variance on the raw data scale, which results in an increase in the standardized mean difference effect size. In fact, even if $\mu_c = \mu_t = 0$, the expected value standardized mean difference will be approximately 0.219. This result confirms that analyzing log-normal data using standard parametric methods without applying the appropriate transformation will deliver invalid results.

TABLE 13
Theoretical Parametric Effect Sizes for Two-Group Log-Normal Data

μ_c	σ_c	μ_t	σ_t	μ_d	σ^2	$StdMD$
0	1	0.00000	1.00	0.000	4.67	0.000
0	1	0.26600	1.00	0.502	6.31	0.200
0	1	0.72375	1.00	1.750	12.30	0.500
0	1	1.43633	1.00	5.280	43.60	0.800
0	1	0.00000	1.50	1.430	42.60	0.219
0	1	0.26600	1.50	2.370	70.90	0.282
0	1	0.72375	1.50	4.700	174.00	0.357
0	1	1.43633	1.50	11.300	714.00	0.423

Table 14 shows the theoretical effect of comparing four groups in a randomized block design. For each block, we use the same values for μ_c and σ_c and μ_t and σ_t. The difference between the blocks is modelled by the term BE, which is added to the mean values in the second block. The columns MD, Var, and $StdMD$ report, respectively, the theoretical mean difference, the variance, and the standardized mean difference on the raw data scale. Rows 1 to 8 in the table report the outcomes when the block effect is set to zero, while rows 9 to 16 show the theoretical effect sizes when there is a block effect. Rows 1 to 8 are exactly the same as Table 13. This confirms that if there is no block effect, there should be no significant difference between the results of analyzing the 2-group data and the results of analyzing the 4-group data.

Rows 9 to 16 in Table 14 show the effect of introducing a positive blocking effect. Rows 9 to 12 use the same variance for the control and treatment conditions, while rows 13 to 14 increase the variance for the treatment condition. The results suggest that block effects have a substantial impact on μ and σ^2 and a relatively small impact on $StdMD$. These results are *not* consistent with the results we would expect if the data were analyzed after a normal transformation.

6. This value and the other values used to achieve standardized effect sizes of 0.5 and 0.8 on the raw data were found by manual iteration (i.e., trial and error).

TABLE 14
Theoretical Parametric Effect Sizes for Four-Group Log-Normal Data

μ_c	σ_c	μ_t	σ_t	BE	μ_d	σ^2	$StdMD$
0	1	0.00000	1.0	0.0	0.000	4.67	0.000
0	1	0.26600	1.0	0.0	0.502	6.31	0.200
0	1	0.72375	1.0	0.0	1.750	12.30	0.500
0	1	1.43633	1.0	0.0	5.280	43.60	0.800
0	1	0.00000	1.5	0.0	1.430	42.60	0.219
0	1	0.26600	1.5	0.0	2.370	70.90	0.282
0	1	0.72375	1.5	0.0	4.700	174.00	0.357
0	1	1.43633	1.5	0.0	11.300	714.00	0.423
0	1	0.00000	1.0	0.5	0.000	8.68	0.000
0	1	0.26600	1.0	0.5	0.665	11.70	0.194
0	1	0.72375	1.0	0.5	2.320	22.80	0.486
0	1	1.43633	1.0	0.5	7.000	81.10	0.777
0	1	0.00000	1.5	0.5	1.900	79.20	0.213
0	1	0.26600	1.5	0.5	3.140	132.00	0.273
0	1	0.72375	1.5	0.5	6.230	323.00	0.347
0	1	1.43633	1.5	0.5	15.000	1330.00	0.411

We simulated the large sample estimates of $\hat{p}$, Cliff's d and $StdMD$ using the same process described in Section 4.1, using log-normal data with μ_t values of (0, 0.266, 0.72375, 1.43633). This process generated the results shown in Table 15. The values correspond to the conditions shown in Table 13. The large sample $StdMD$ estimates are close to the theoretical $StdMD$ values, however, $\hat{p}$ and Cliff's d behave differently. Table 13 rows 5 to 8 show:

- In the case when there is additional variation but there is no difference between the values of μ_c and μ_t, the $StdMD$ value is significantly different from 0. In fact, it is larger than the value we would usually call a small effect size. In contrast, $\hat{p}$ and Cliff's d have values very close to their null values of 0.5 and 0, respectively.
- In the case where there is additional variation and $\mu_c \neq \mu_t$, $StdMD$ values increase at a much slower rate than when there is no increase in the variance, for example, for the largest treatment difference, the value of $StdMD$ is equivalent to a medium effect. The values of $\hat{p}$ and Cliff's d are lower than they were when there was no additional variance, but the effect size increase remains consistent with interpretations of small, medium, and large effects.

This is an example of strong disagreement between the parametric effect size and the nonparametric effect sizes. Furthermore, since applying an appropriate normalizing transformation to the raw data would reproduce the values that generated the data, it confirms that analyzing the raw data with parametric methods would give a misleading assessment of the efficacy of a method or technique, while the nonparametric effect sizes would not.

The large sample effect sizes for the four-group randomized blocks experimental design are shown in Table 16. It is noticeable that the nonparametric effect sizes are not affected by the blocking factor.

TABLE 15
Large Sample Effect Sizes for Two-Group Log-Normal Experiments

Additional Variance	$\hat{p}$	Cliff's d	MD	Var	$StdMD$
No	0.500	-0.000	-0.002	4.673	-0.001
No	0.574	0.148	0.499	6.303	0.199
No	0.695	0.391	1.752	12.273	0.500
No	0.845	0.690	5.279	43.384	0.801
Yes	0.500	-0.000	1.430	41.784	0.221
Yes	0.559	0.117	2.369	69.973	0.283
Yes	0.656	0.312	4.708	174.973	0.356
Yes	0.787	0.574	11.286	706.874	0.424

TABLE 16
Large Sample Estimates of Effect Size for Log-Normally Distributed Four-Group Randomised Experiments

Additional Variance	Block Effect	$\hat{p}$	Cliff's d	MD	Var	$StdMD$
No	No	0.500	0.000	0.001	4.669	0.000
No	No	0.575	0.149	0.502	6.307	0.200
No	No	0.696	0.391	1.750	12.249	0.500
No	No	0.845	0.690	5.286	43.622	0.800
No	Yes	0.500	0.000	0.000	8.690	0.000
No	Yes	0.575	0.149	0.667	11.755	0.195
No	Yes	0.696	0.391	2.316	22.786	0.485
No	Yes	0.845	0.690	6.997	80.842	0.778
Yes	No	0.500	0.000	1.434	41.630	0.222
Yes	No	0.559	0.117	2.370	70.486	0.282
Yes	No	0.656	0.312	4.704	174.34	0.356
Yes	No	0.787	0.574	11.308	729.93	0.419
Yes	Yes	0.500	-0.000	1.895	78.71	0.214
Yes	Yes	0.559	0.117	3.129	128.27	0.276
Yes	Yes	0.656	0.312	6.231	327.77	0.344
Yes	Yes	0.787	0.574	14.974	1388.1	0.402

4.3 The Gamma Distribution

The gamma distribution has two parameters referred to as the rate (β) and shape (α) parameters. The gamma distribution is defined as:

$$x = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{43}$$

In software engineering, the shape might be used to represent the amount of work to be done, and the rate parameter would then represent the effort per unit of work. This is the basic idea behind the Putnam model [18]. The mean of a gamma distribution is:

$$mean = \frac{\alpha}{\beta} \tag{44}$$

where $\beta > 0$ and $\alpha > 0$. So, for example, if your overall workload (α) is 200 function points and your work rate (β) is 2 function points per day, simple arithmetic would indicate the workload would take 100 days to complete, corresponding to the mean of the gamma distribution with rate=2 and shape=200. The variance of a gamma distribution is:

$$variance = \frac{\alpha}{\beta^2} \tag{45}$$

with an equivalent formula being:

$$variance = \frac{mean}{\beta} \tag{46}$$

Thus, like the log-normal distribution, changes to the rate or the shape of the gamma distribution not only change the mean of a simulated data set but also change the variance.

In the context of software engineering experiments, the gamma distribution is relevant to experiments that investigate whether a technique increases the efficiency or productivity of a specific task. This would correspond to investigating whether a technique increases in the rate parameter of a gamma distribution, which in turn would *decrease* the mean of the raw data sample. In contrast to the way we modelled block effects for normal data, we assume that differences in the complexity of tasks or software development artefacts used in experiments investigating efficiency will change the shape parameter (i.e., the measure of the amount of work to be done) not the rate parameter (i.e., a measure of the efficiency with which the work is performed).

For other distributions, we assumed that the treatment was expected to increase the mean of the raw data, so we calculated the mean difference as:

$$MD = MD_T - MD_C \tag{47}$$

For positive effect sizes, this lead to positive values for $StdMD$ and Cliff's d values and values of $\hat{p} > 0.5$. In contrast, for the gamma distribution, if the treatment increases the rate parameter, the outcome of using Equation (47) is to reduce the MD_T value compared with the MD_C value. Thus, for a positive increase in the rate values, using Equation (47) leads to negative values of $StdMD$ and Cliff's d, and values of $\hat{p} < 0.5$.

For all our gamma distribution simulations, we set our control conditions for both two-group and four-group experiments to $rate = 1$ and $shape = 3$. We found, by manual iteration, that for zero and negative standardized mean differences of 0, -0.2, -0.5, and -0.8, the appropriate values of the mean difference are (0, 0.1225, 0.3415, 0.6224). However, for zero and positive standardized mean difference effects of 0, 0.2, 0.5, and 0.8, appropriate values of the mean difference are (0, -0.1095, -0.2545, -0.3838).

For the gamma distribution, we did not consider the impact of variance heterogeneity other than the impact of introducing a block effect in four-group experiments. The relationship between the shape and rate parameter makes the concept of heterogeneity difficult to understand and model. Furthermore, it is unnecessary to consider this issue (other than via changes to the block parameter) to investigate the accuracy and power of nonparametric effect sizes for individual experiments. For the meta-analysis, we applied a small random change to the *rate* parameter in order to model the differences between individual experiments in a family.

The theoretical standardized effect size, variance, and standardized effect size for a two-group randomized experimental design are shown in Table 17. The first four rows show the expected raw data parameter values when the rate parameter is zero or positive. The last four rows show the expected raw data parameter values when the rate parameter is zero or negative.

Table 18 shows the theoretical parameter values and effect sizes for the gamma distribution based on a four-group randomized block design. For negative and positive effect sizes, we report the impact of both a zero block effect

TABLE 17
Theoretical Effect Sizes for a Two-Group Randomised Experiments based Gamma Data

$Rate_c$	$Shape_c$	$Rate_t$	$Shape_t$	μ_d	σ^2	$StdMD$
1	3	1.0000	3.0	0.000	3.00	0
1	3	1.1225	3.0	-0.327	2.69	-0.2
1	3	1.3415	3.0	-0.764	2.33	-0.5
1	3	1.6224	3.0	-1.150	2.07	-0.8
1	3	1.0000	3.0	0.000	3	0
1	3	0.8905	3.0	0.369	3.39	0.2
1	3	0.7455	3.0	1.020	4.20	0.5
1	3	0.6162	3.0	1.870	5.48	0.8

TABLE 18
Theoretical Effect Sizes for a Four-Group Randomised Block Experiments based on Gamma Data

$Rate_c$	Sh_c	$Rate_t$	Sh_t	BE	μ_d	σ^2	$StdMD$
1	3	1	3	0	0	3.000	0
1	3	1.123	3	0	-0.327	2.690	-0.200
1	3	1.341	3	0	-0.764	2.330	-0.500
1	3	1.622	3	0	-1.150	2.070	-0.800
1	3	1	3	0.5	0	3.250	0
1	3	1.123	3	0.5	-0.355	2.910	-0.208
1	3	1.341	3	0.5	-0.827	2.530	-0.520
1	3	1.622	3	0.5	-1.250	2.240	-0.833
1	3	1	3	0	0	3	0
1	3	0.8905	3	0	0.369	3.39	0.2
1	3	0.7455	3	0	1.02	4.20	0.5
1	3	0.6162	3	0	1.87	5.45	0.8
1	3	1	3	0.5	0	3	0
1	3	0.8905	3	0.5	0.4	3.67	0.208
1	3	0.7455	3	0.5	1.11	4.55	0.520
1	3	0.6162	3	0.5	2.02	5.90	0.833

and a non-zero block effect of 0.5 applied to the $shape$ parameter. The table confirms that the block effect increases both the mean difference and the variance of the raw data and leads to a small increase in the standardized mean difference that increases as the mean difference increases.

TABLE 19
Large Sample Size Effect Sizes for a Two-Group Randomised Experiments based on Gamma Data

$\hat{p}$	Cliff's d	MD	Var	$StdMD$
0.500	0.000	0.001	2.999	0.000
0.446	-0.108	-0.327	2.689	-0.199
0.365	-0.270	-0.764	2.333	-0.500
0.286	-0.429	-1.151	2.070	-0.800
0.500	0.000	0.001	3.000	0.000
0.554	0.108	0.369	3.390	0.200
0.635	0.269	1.023	4.200	0.499
0.715	0.429	1.870	5.449	0.801

The large sample size nonparametric effect sizes for two-group randomized experimental designs is shown in Table 19. For gamma data, the nonparametric effect sizes $\hat{p}$ and Cliff's d change in a manner consistent with the changes to the theoretical $StdMD$ values when the shape parameter is increased. I.e., when there is no difference between the rates in the control and treatment group but the shape parameter is increased, the values of $\hat{p}$, Cliff's d, and $StdMD$ are all consistent with a small increase in the magnitude of the effect size.

TABLE 20
Large Sample Size Effect Sizes for a Four-Group Randomised Block Experiments based on Gamma Data

Block Effect	$\hat{p}$	Cliff's d	MD	Var	$StdMD$
No	0.500	0.000	0.001	2.998	0.000
No	0.446	-0.108	-0.327	2.691	-0.200
No	0.365	-0.269	-0.763	2.333	-0.499
No	0.286	-0.428	-1.151	2.071	-0.800
Yes	0.500	0.000	0.000	3.250	0.000
Yes	0.444	-0.113	-0.355	2.914	-0.208
Yes	0.359	-0.281	-0.829	2.528	-0.521
Yes	0.277	-0.445	-1.247	2.242	-0.833
No	0.500	0.000	0.000	3.001	0.000
No	0.554	0.108	0.369	3.391	0.201
No	0.635	0.270	1.024	4.198	0.500
No	0.714	0.429	1.869	5.454	0.800
Yes	0.500	0.000	0.000	3.250	0.000
Yes	0.556	0.113	0.399	3.673	0.208
Yes	0.640	0.281	1.111	4.549	0.521
Yes	0.723	0.445	2.024	5.901	0.833

The large sample effect sizes for four-group randomized experiments are shown in Table 20.

4.4 Laplace Distribution Simulations

The Laplace distribution is a symmetrical distribution with a relatively large number of outliers. It has two parameters, a location parameter μ and a scale parameter b. It has not been used in software engineering research but does provide a useful contrast to normally distributed data with respect to its excessive number of outliers, but, in contrast to the skewed log-normal and gamma distributions, it is a symmetric distribution.

It is not generated by the R package, so we generated Laplace data by obtaining random variables from the uniform distribution in the interval (-1/2,1/2), then

$$X = \mu - b \times sgn(U) \times ln(1 - 2|U|) \tag{48}$$

where X is a random variable from the Laplace distribution, U is a random variable from a uniform distribution, $sgn(U)$ denotes a function that identifies the sign of U and ln denotes the natural logarithm.

Given parameters μ and b, the expected mean value of a Laplace data set is:

$$mean = \mu \tag{49}$$

and its variance is:

$$variance = 2 \times b^2 \tag{50}$$

Like the normal distribution, the mean and the variance of the Laplace distribution are functionally independent. We set the control condition to have $\mu = 0$ and $b = 1$ and found by manual iteration values for the treatment mean that would lead to a standardized mean differences effect size of (0, 0.2,0.5, 0,8). The two-group theoretical effect sizes are shown in Table 21, and the four-group theoretical effect sizes are shown in Table 22. These tables show that data sets from the Laplace distribution behave in a similar manner to normal data sets:

- The effect sizes for the two-group and four-group experimental designs are almost identical.
- Block effects that apply to mean values do not change effect sizes.
- Increasing the value of b increases the variance of the treatment group and decreases the standardized effect size.

TABLE 21
Theoretical Effect Sizes for a Two-Group Randomised Experiments based on Laplace Data

μ_c	b_c	μ_t	b_t	μ_d	σ^2	$StdMD$
0	1	0.000000	1.0	0.000	2.00	0.000
0	1	0.283000	1.0	0.283	2.00	0.200
0	1	0.707104	1.0	0.707	2.00	0.500
0	1	1.131374	1.0	1.131	2.00	0.800
0	1	0.000000	1.5	0.000	3.25	0.000
0	1	0.283000	1.5	0.283	3.25	0.157
0	1	0.707104	1.5	0.707	3.25	0.392
0	1	1.131374	1.5	1.131	3.25	0.628

TABLE 22
Theoretical Effect Sizes for a Four-Group Randomised Block Experiments based on Laplace Data

μ_c	b_c	μ_t	b_t	BE	μ_d	σ^2	$StdMD$
0	1	0.00000	1.0	0.0	0.000	2.00	0.000
0	1	0.28300	1.0	0.0	0.283	2.00	0.200
0	1	0.70710	1.0	0.0	0.707	2.00	0.500
0	1	1.13137	1.0	0.0	1.130	2.00	0.800
0	1	0.00000	1.5	0.0	0.000	3.25	0.000
0	1	0.28300	1.5	0.0	0.283	3.25	0.157
0	1	0.70710	1.5	0.0	0.707	3.25	0.392
0	1	1.13137	1.5	0.0	1.130	3.25	0.628
0	1	0.00000	1.0	0.5	0.000	2.00	0.000
0	1	0.28300	1.0	0.5	0.283	2.00	0.200
0	1	0.70710	1.0	0.5	0.707	2.00	0.500
0	1	1.13137	1.0	0.5	1.130	2.00	0.800
0	1	0.00000	1.5	0.5	0.000	3.25	0.000
0	1	0.28300	1.5	0.5	0.283	3.25	0.157
0	1	0.70710	1.5	0.5	0.707	3.25	0.392
0	1	1.13137	1.5	0.5	1.130	3.25	0.628

Large sample size nonparametric effect size values for two-group experiments are shown in Table 23 and for four-group experiments are shown in Table 24. The results are similar to the results for the Normal distribution. Blocking has a negligible impact on the effect sizes but they are reduced in the presence of heterogeneity.

5 Simulations of Four-Group Randomized Blocks Experiments

In this section, we report simulation studies results of four-group randomized block experiments and simulation studies of meta-analysis of two-group and four-group. The simulation details and graphical summaries of the results can be found in [1].

Table 25 includes the results for the simulations of four-group single experiments. The information in the Data Type column identifies the data distributions, which is N for Normal, Lap for Laplace, L for Lognormal and G for Gamma. The symbol "-H" is used to identify whether the simulations in a specific row were simulated with additional variance heterogeneity for one block. The Mean Diff column specifies that the difference between the control group mean and treatment group mean corresponded to a Small (S) difference (i.e., an intended 0.2 standardized mean difference on the raw data scale), Medium (M) difference (i.e., an intended 0.5 standardized mean difference on the raw data scale) or a Large (L) difference (i.e., an intended 0.8 standardized mean difference on the raw data scale). The column labelled Block Included indicates whether the group design included a fixed block effect. The Power Difference, Bias and MdMRE values were all multiplied by 100 to improve readability. Other tables include similarly labelled information.

TABLE 23
Large Sample Size Effect Sizes for Two-Group Randomised Experiments based on Laplace Data

Additional Variance	$\hat{p}$	Cliff's d	MD	Var	$StdMD$
No	0.500	0.001	0.002	2.001	0.001
No	0.570	0.140	0.283	2.002	0.200
No	0.666	0.333	0.707	2.000	0.500
No	0.747	0.495	1.131	1.998	0.800
Yes	0.500	-0.000	-0.001	3.251	-0.000
Yes	0.556	0.112	0.282	3.248	0.156
Yes	0.636	0.272	0.709	3.249	0.393
Yes	0.706	0.411	1.131	3.253	0.627

TABLE 24
Large Sample Size Effect Sizes for Four-Group Randomised Block Experiments based on Laplace Data

Additional Variance	Block Effect	$\hat{p}$	Cliff's d	MD	Var	$StdMD$
No	No	0.500	0.000	0.000	2.000	0.000
No	No	0.570	0.140	0.283	1.999	0.200
No	No	0.666	0.333	0.707	1.999	0.500
No	No	0.747	0.495	1.131	1.999	0.800
No	Yes	0.500	0.000	0.000	2.000	0.000
No	Yes	0.570	0.140	0.284	2.000	0.201
No	Yes	0.666	0.332	0.707	2.000	0.500
No	Yes	0.747	0.495	1.131	2.000	0.800
Yes	No	0.500	-0.000	-0.000	3.249	-0.000
Yes	No	0.556	0.112	0.283	3.251	0.157
Yes	No	0.636	0.271	0.707	3.250	0.392
Yes	No	0.706	0.411	1.131	3.249	0.627
Yes	Yes	0.500	0.000	0.001	3.250	0.000
Yes	Yes	0.556	0.112	0.284	3.250	0.157
Yes	Yes	0.635	0.271	0.707	3.249	0.392
Yes	Yes	0.706	0.411	1.130	3.251	0.627

TABLE 25: Four-Group Experiment Simulation Results

	Data Type	Block Included	Group Size	Mean Diff	NP Bias	StdMD Bias	NP PMdMRE	StdMD PMdMRE	PHat Observed	Cliffd Observed	StdMD Observed	PHat Power	Cliffd Power	StdMD Power	CliffPower Difference	PHatPower Difference
1	N	Yes	5	S	1.35	4.81	171.43	154.13	0.557	0.114	0.210	0.102	0.089	0.106	-1.670	-0.410
2	N	Yes	5	M	0.30	4.96	59.42	62.36	0.638	0.277	0.525	0.250	0.225	0.270	-4.470	-1.940
3	N	Yes	5	L	0.30	4.99	40.19	39.71	0.715	0.429	0.840	0.482	0.441	0.513	-7.240	-3.080
4	N	Yes	10	S	1.99	3.61	114.29	109.68	0.557	0.114	0.207	0.146	0.132	0.158	-2.650	-1.180
5	N	Yes	10	M	0.94	2.84	44.93	44.27	0.639	0.279	0.514	0.437	0.416	0.460	-4.340	-2.280
6	N	Yes	10	L	0.58	2.64	26.17	28.32	0.715	0.430	0.821	0.768	0.748	0.797	-4.910	-2.850
7	N	Yes	15	S	3.64	4.09	90.48	87.25	0.558	0.116	0.208	0.184	0.173	0.193	-2.020	-0.940
8	N	Yes	15	M	1.18	2.45	35.27	35.06	0.640	0.279	0.512	0.585	0.569	0.616	-4.750	-3.080
9	N	Yes	15	L	0.72	2.04	21.08	22.38	0.716	0.431	0.816	0.907	0.899	0.923	-2.440	-1.590
10	N	Yes	20	S	1.25	1.61	78.57	75.46	0.557	0.113	0.203	0.216	0.205	0.225	-2.060	-0.970
11	N	Yes	20	M	0.41	1.25	30.43	30.64	0.639	0.277	0.506	0.693	0.679	0.719	-3.990	-2.610
12	N	Yes	20	L	0.25	1.16	17.99	19.62	0.715	0.429	0.809	0.961	0.958	0.969	-1.120	-0.890
13	N	Yes	40	S	1.53	1.46	55.69	53.50	0.557	0.114	0.203	0.344	0.337	0.358	-2.140	-1.440
14	N	Yes	40	M	0.52	0.86	21.65	21.88	0.639	0.277	0.504	0.919	0.917	0.931	-1.430	-1.230
15	N	Yes	40	L	0.29	0.71	12.97	13.98	0.715	0.429	0.806	1.000	1.000	1.000	-0.040	-0.040
16	N-H	Yes	5	S	-1.43	2.36	218.18	200.64	0.543	0.087	0.161	0.086	0.075	0.088	-1.270	-0.160
17	N-H	Yes	5	M	-0.68	4.40	83.49	80.80	0.608	0.217	0.409	0.189	0.169	0.198	-2.940	-0.950
18	N-H	Yes	5	L	-0.16	4.75	52.05	51.14	0.671	0.341	0.658	0.333	0.304	0.361	-5.730	-2.790
19	N-H	Yes	10	S	-2.01	0.03	145.45	136.69	0.543	0.086	0.157	0.111	0.100	0.118	-1.840	-0.710
20	N-H	Yes	10	M	-0.54	1.55	55.96	55.44	0.608	0.217	0.398	0.302	0.282	0.325	-4.350	-2.280
21	N-H	Yes	10	L	-0.22	1.77	34.50	34.92	0.671	0.341	0.639	0.576	0.552	0.616	-6.450	-4.040
22	N-H	Yes	15	S	-1.42	-0.91	115.15	110.18	0.543	0.087	0.156	0.134	0.127	0.142	-1.580	-0.900
23	N-H	Yes	15	M	-0.29	0.69	44.95	44.32	0.609	0.217	0.395	0.409	0.393	0.436	-4.240	-2.660
24	N-H	Yes	15	L	-0.13	0.93	28.53	28.29	0.671	0.342	0.634	0.732	0.719	0.772	-5.320	-4.030
25	N-H	Yes	20	S	0.08	1.03	101.70	97.68	0.544	0.088	0.159	0.163	0.157	0.172	-1.540	-0.900
26	N-H	Yes	20	M	0.17	1.24	40.37	39.78	0.609	0.218	0.397	0.504	0.491	0.535	-4.400	-3.110
27	N-H	Yes	20	L	0.22	1.13	24.71	24.96	0.671	0.343	0.635	0.839	0.831	0.867	-3.590	-2.760
28	N-H	Yes	40	S	0.03	-0.13	71.59	69.21	0.544	0.088	0.157	0.241	0.235	0.255	-1.950	-1.370
29	N-H	Yes	40	M	0.05	0.35	28.15	27.69	0.609	0.218	0.393	0.761	0.755	0.789	-3.420	-2.820
30	N-H	Yes	40	L	0.11	0.31	17.03	17.58	0.671	0.342	0.630	0.981	0.981	0.989	-0.820	-0.750
31	Lap	Yes	5	S	0.87	11.53	128.57	157.25	0.571	0.141	0.223	0.119	0.104	0.107	-0.310	1.180
32	Lap	Yes	5	M	0.78	10.57	51.81	64.99	0.667	0.335	0.553	0.324	0.293	0.289	0.440	3.520
33	Lap	Yes	5	L	0.66	10.35	29.55	42.55	0.749	0.497	0.883	0.581	0.540	0.539	0.070	4.200
34	Lap	Yes	10	S	-0.85	4.16	92.86	109.75	0.569	0.139	0.208	0.179	0.162	0.155	0.700	2.410
35	Lap	Yes	10	M	-0.22	4.47	35.54	45.31	0.666	0.331	0.522	0.550	0.521	0.475	4.570	7.460
36	Lap	Yes	10	L	-0.08	4.57	23.08	30.00	0.747	0.494	0.837	0.865	0.851	0.796	5.440	6.840
37	Lap	Yes	15	S	-1.20	1.84	71.43	88.21	0.569	0.138	0.204	0.224	0.211	0.187	2.400	3.650
38	Lap	Yes	15	M	-0.28	2.50	29.05	35.84	0.666	0.331	0.512	0.711	0.695	0.613	8.220	9.760
39	Lap	Yes	15	L	-0.14	2.68	17.86	23.56	0.747	0.493	0.821	0.963	0.960	0.915	4.460	4.790
40	Lap	Yes	20	S	0.44	3.64	60.71	77.13	0.570	0.141	0.207	0.277	0.266	0.234	3.110	4.270
41	Lap	Yes	20	M	0.55	2.97	24.70	31.45	0.667	0.334	0.515	0.829	0.820	0.726	9.440	10.340
42	Lap	Yes	20	L	0.38	2.82	14.98	20.64	0.748	0.496	0.823	0.990	0.989	0.967	2.210	2.260
43	Lap	Yes	40	S	-0.09	1.71	44.20	54.31	0.570	0.140	0.203	0.450	0.443	0.357	8.610	9.380
44	Lap	Yes	40	M	0.17	1.29	17.73	22.58	0.666	0.333	0.506	0.980	0.979	0.938	4.140	4.200
45	Lap	Yes	40	L	0.17	1.20	10.93	14.94	0.747	0.495	0.810	1.000	1.000	1.000	0.020	0.020
46	Lap-H	Yes	5	S	-0.40	10.22	171.43	201.19	0.556	0.112	0.173	0.097	0.086	0.089	-0.310	0.820
47	Lap-H	Yes	5	M	0.30	11.18	62.36	82.52	0.636	0.272	0.436	0.240	0.217	0.213	0.370	2.730
48	Lap-H	Yes	5	L	0.06	11.26	41.75	53.44	0.706	0.412	0.699	0.442	0.408	0.398	1.050	4.460
49	Lap-H	Yes	10	S	-0.31	4.66	117.86	140.84	0.556	0.112	0.164	0.146	0.133	0.130	0.380	1.610
50	Lap-H	Yes	10	M	-0.18	5.19	47.60	58.44	0.635	0.271	0.412	0.411	0.389	0.350	3.870	6.100
51	Lap-H	Yes	10	L	0.00	5.17	28.64	37.82	0.706	0.412	0.660	0.712	0.690	0.622	6.890	8.990
52	Lap-H	Yes	15	S	0.13	2.97	91.27	111.62	0.556	0.112	0.162	0.175	0.163	0.146	1.670	2.880
53	Lap-H	Yes	15	M	-0.05	3.50	36.12	46.33	0.635	0.271	0.406	0.548	0.530	0.452	7.820	9.650
54	Lap-H	Yes	15	L	-0.19	3.49	22.98	30.00	0.706	0.411	0.650	0.866	0.856	0.776	8.020	9.010
55	Lap-H	Yes	20	S	-0.18	2.17	78.57	98.94	0.556	0.112	0.160	0.208	0.199	0.175	2.340	3.220
56	Lap-H	Yes	20	M	-0.26	2.50	31.92	40.39	0.635	0.270	0.402	0.661	0.645	0.541	10.410	11.920
57	Lap-H	Yes	20	L	-0.31	2.44	19.90	26.27	0.705	0.411	0.643	0.939	0.936	0.869	6.700	7.010
58	Lap-H	Yes	40	S	0.51	2.23	56.25	69.29	0.556	0.113	0.161	0.331	0.325	0.265	5.960	6.570
59	Lap-H	Yes	40	M	0.05	1.71	22.28	28.14	0.636	0.271	0.399	0.903	0.900	0.796	10.390	10.690
60	Lap-H	Yes	40	L	-0.11	1.43	14.08	18.29	0.706	0.412	0.637	0.999	0.999	0.990	0.930	0.950
61	L	No	5	S	-3.38	15.47	126.67	167.24	0.572	0.145	0.231	0.123	0.107	0.084	2.350	3.950
62	L	No	5	M	-1.08	22.75	42.86	59.58	0.694	0.388	0.614	0.411	0.378	0.288	9.030	12.310
63	L	No	5	L	-0.41	34.56	18.84	36.59	0.844	0.687	1.077	0.887	0.838	0.697	14.080	18.970
64	L	No	10	S	-1.39	12.57	86.67	114.48	0.574	0.148	0.225	0.196	0.180	0.157	2.300	3.860
65	L	No	10	M	-0.34	16.74	28.57	40.78	0.695	0.391	0.584	0.688	0.666	0.569	9.740	11.900
66	L	No	10	L	-0.01	24.16	11.59	27.58	0.845	0.690	0.993	0.995	0.993	0.951	4.280	4.400
67	L	No	15	S	-2.17	9.36	67.41	89.75	0.573	0.147	0.219	0.245	0.232	0.196	3.640	4.960
68	L	No	15	M	-0.80	13.04	23.58	31.95	0.694	0.389	0.565	0.844	0.833	0.724	10.880	11.960
69	L	No	15	L	-0.17	18.90	10.14	22.83	0.844	0.689	0.951	1.000	1.000	0.990	1.000	1.010
70	L	No	20	S	0.32	10.18	56.67	78.21	0.575	0.150	0.220	0.306	0.294	0.248	4.560	5.800
71	L	No	20	M	0.07	12.07	20.28	27.72	0.696	0.392	0.560	0.931	0.926	0.831	9.480	10.020
72	L	No	20	L	0.23	16.75	8.33	20.50	0.846	0.692	0.934	1.000	1.000	0.998	0.250	0.250
73	L	No	40	S	-0.33	6.23	41.25	53.28	0.575	0.150	0.212	0.494	0.486	0.389	9.720	10.470
74	L	No	40	M	-0.15	7.76	14.32	19.75	0.696	0.391	0.539	0.998	0.998	0.971	2.690	2.700
75	L	No	40	L	0.04	10.77	5.89	15.82	0.845	0.690	0.886	1.000	1.000	1.000	0.010	0.010
76	L	Yes	5	S	-3.38	19.04	126.67	172.90	0.572	0.145	0.231	0.123	0.107	0.084	2.350	3.950
77	L	Yes	5	M	-1.08	26.29	42.86	62.34	0.694	0.388	0.614	0.411	0.378	0.288	9.030	12.310
78	L	Yes	5	L	-0.41	38.55	18.84	39.52	0.844	0.687	1.077	0.887	0.838	0.697	14.080	18.970

Continued on next page

TABLE 25 – continued from previous page

	Data Type	Block Included	Group Size	Mean Diff	NP Bias	StdMD Bias	NP PMdMRE	StdMD PMdMRE	PHat Observed	Cliffd Observed	StdMD Observed	PHat Power	Cliffd Power	StdMD Power	CliffPower Difference	PHatPower Difference
79	L	Yes	10	S	-1.39	16.05	86.67	118.24	0.574	0.148	0.225	0.196	0.180	0.157	2.300	3.860
80	L	Yes	10	M	-0.34	20.11	28.57	42.73	0.695	0.391	0.584	0.688	0.666	0.569	9.740	11.900
81	L	Yes	10	L	-0.01	27.84	11.59	29.83	0.845	0.690	0.993	0.995	0.993	0.951	4.280	4.400
82	L	Yes	15	S	-2.17	12.74	67.41	92.55	0.573	0.147	0.219	0.245	0.232	0.196	3.640	4.960
83	L	Yes	15	M	-0.80	16.29	23.58	33.59	0.694	0.389	0.565	0.844	0.833	0.724	10.880	11.960
84	L	Yes	15	L	-0.17	22.42	10.14	25.20	0.844	0.689	0.951	1.000	1.000	0.990	1.000	1.010
85	L	Yes	20	S	0.32	13.59	56.67	80.96	0.575	0.150	0.220	0.306	0.294	0.248	4.560	5.800
86	L	Yes	20	M	0.07	15.30	20.28	29.50	0.696	0.392	0.560	0.931	0.926	0.831	9.480	10.020
87	L	Yes	20	L	0.23	20.21	8.33	22.77	0.846	0.692	0.934	1.000	1.000	0.998	0.250	0.250
88	L	Yes	40	S	-0.33	9.51	41.25	55.22	0.575	0.150	0.212	0.494	0.486	0.389	9.720	10.470
89	L	Yes	40	M	-0.15	10.86	14.32	21.02	0.696	0.391	0.539	0.998	0.998	0.971	2.690	2.700
90	L	Yes	40	L	0.04	14.04	5.89	18.00	0.845	0.690	0.886	1.000	1.000	1.000	0.010	0.010
91	G	No	5	S	3.77	7.98	174.07	153.97	0.444	-0.112	-0.216	0.104	0.091	0.104	-1.300	0.040
92	G	No	5	M	1.19	5.88	62.96	60.13	0.363	-0.273	-0.529	0.247	0.220	0.257	-3.690	-1.070
93	G	No	5	L	0.90	5.96	40.19	37.09	0.284	-0.432	-0.848	0.483	0.445	0.519	-7.430	-3.610
94	G	No	10	S	-1.99	-0.67	118.52	108.02	0.447	-0.106	-0.199	0.136	0.121	0.144	-2.360	-0.860
95	G	No	10	M	-0.73	0.97	44.44	42.73	0.366	-0.268	-0.505	0.418	0.394	0.456	-6.130	-3.750
96	G	No	10	L	-0.20	1.71	26.17	26.02	0.286	-0.427	-0.814	0.760	0.741	0.815	-7.340	-5.420
97	G	No	15	S	-0.43	-0.48	93.42	86.63	0.446	-0.108	-0.199	0.169	0.158	0.178	-2.010	-0.930
98	G	No	15	M	-0.44	0.66	36.63	33.58	0.366	-0.269	-0.503	0.559	0.543	0.611	-6.790	-5.180
99	G	No	15	L	0.01	1.18	21.08	20.36	0.286	-0.428	-0.809	0.903	0.895	0.940	-4.470	-3.690
100	G	No	20	S	-1.14	0.11	81.48	76.09	0.447	-0.107	-0.200	0.199	0.192	0.223	-3.130	-2.420
101	G	No	20	M	-0.64	0.61	31.48	29.64	0.366	-0.268	-0.503	0.661	0.649	0.720	-7.090	-5.920
102	G	No	20	L	-0.08	0.87	18.22	17.94	0.286	-0.428	-0.807	0.961	0.959	0.979	-1.980	-1.830
103	G	No	40	S	0.30	0.34	57.99	53.40	0.446	-0.108	-0.201	0.319	0.313	0.350	-3.680	-3.110
104	G	No	40	M	-0.03	0.47	21.99	20.96	0.365	-0.270	-0.502	0.909	0.907	0.940	-3.350	-3.110
105	G	No	40	L	0.23	0.56	12.88	12.58	0.286	-0.429	-0.804	1.000	1.000	1.000	-0.020	-0.020
106	G	Yes	5	S	0.07	3.83	171.43	147.97	0.444	-0.112	-0.216	0.104	0.091	0.104	-1.300	0.040
107	G	Yes	5	M	-3.11	1.81	57.45	57.93	0.363	-0.273	-0.529	0.247	0.220	0.257	-3.690	-1.070
108	G	Yes	5	L	-3.17	1.76	37.22	35.68	0.284	-0.432	-0.848	0.483	0.445	0.519	-7.430	-3.610
109	G	Yes	10	S	-5.49	-4.49	114.29	103.87	0.447	-0.106	-0.199	0.136	0.121	0.144	-2.360	-0.860
110	G	Yes	10	M	-4.95	-2.92	43.26	40.82	0.366	-0.268	-0.505	0.418	0.394	0.456	-6.130	-3.750
111	G	Yes	10	L	-4.23	-2.32	25.56	24.95	0.286	-0.427	-0.814	0.760	0.741	0.815	-7.340	-5.420
112	G	Yes	15	S	-3.98	-4.30	90.48	82.80	0.446	-0.108	-0.199	0.169	0.158	0.178	-2.010	-0.930
113	G	Yes	15	M	-4.68	-3.21	35.38	32.14	0.366	-0.269	-0.503	0.559	0.543	0.611	-6.790	-5.180
114	G	Yes	15	L	-4.02	-2.83	20.28	19.76	0.286	-0.428	-0.809	0.903	0.895	0.940	-4.470	-3.690
115	G	Yes	20	S	-4.67	-3.74	78.57	73.35	0.447	-0.107	-0.200	0.199	0.192	0.223	-3.130	-2.420
116	G	Yes	20	M	-4.87	-3.26	29.96	28.64	0.366	-0.268	-0.503	0.661	0.649	0.720	-7.090	-5.920
117	G	Yes	20	L	-4.11	-3.13	17.60	17.52	0.286	-0.428	-0.807	0.961	0.959	0.979	-1.980	-1.830
118	G	Yes	40	S	-3.28	-3.52	55.69	51.27	0.446	-0.108	-0.201	0.319	0.313	0.350	-3.680	-3.110
119	G	Yes	40	M	-4.28	-3.40	21.32	20.15	0.365	-0.270	-0.502	0.909	0.907	0.940	-3.350	-3.110
120	G	Yes	40	L	-3.82	-3.42	12.42	12.33	0.286	-0.429	-0.804	1.000	1.000	1.000	-0.020	-0.020

Table 26 includes the Type 1 error rates for the four-group single experiment simulations.

TABLE 26
Four-Group Single Experiment Type 1 Error Rates

	Data Type	Group Size	Block Included	PHat Observed	Cliffd Observed	StdMD Observed	PHat Type1 Error Rate	Cliffd Type1 Error Rate	StdMD Type1 Error Rate
1	N	5	Yes	0.501	0.002	0.006	0.050	0.035	0.043
2	N	10	Yes	0.500	0.001	0.004	0.049	0.040	0.048
3	N	15	Yes	0.500	0.000	-0.001	0.054	0.046	0.051
4	N	20	Yes	0.500	0.001	0.001	0.049	0.044	0.048
5	N	40	Yes	0.501	0.002	0.003	0.048	0.044	0.048
6	N-H	5	Yes	0.499	-0.002	-0.004	0.052	0.037	0.044
7	N-H	10	Yes	0.500	-0.001	-0.002	0.048	0.037	0.047
8	N-H	15	Yes	0.502	0.004	0.006	0.048	0.040	0.047
9	N-H	20	Yes	0.501	0.002	0.005	0.050	0.044	0.051
10	N-H	40	Yes	0.500	0.000	0.001	0.051	0.047	0.050
11	Lap	5	Yes	0.499	-0.002	-0.002	0.049	0.033	0.041
12	Lap	10	Yes	0.499	-0.001	-0.002	0.050	0.041	0.047
13	Lap	15	Yes	0.499	-0.002	-0.003	0.050	0.044	0.054
14	Lap	20	Yes	0.499	-0.001	-0.003	0.052	0.045	0.051
15	Lap	40	Yes	0.500	0.001	-0.001	0.053	0.051	0.052
16	Lap-H	5	Yes	0.497	-0.006	-0.009	0.053	0.037	0.040
17	Lap-H	10	Yes	0.500	0.000	-0.000	0.048	0.038	0.045
18	Lap-H	15	Yes	0.499	-0.002	-0.003	0.049	0.041	0.044
19	Lap-H	20	Yes	0.500	-0.001	-0.000	0.052	0.046	0.050
20	Lap-H	40	Yes	0.500	0.001	0.001	0.048	0.045	0.050
21	L	5	No	0.499	-0.002	0.000	0.052	0.036	0.022
22	L	10	No	0.501	0.002	0.003	0.051	0.039	0.030
23	L	15	No	0.501	0.002	0.005	0.056	0.048	0.042
24	L	20	No	0.500	0.001	0.000	0.049	0.044	0.040
25	L	40	No	0.500	-0.001	0.001	0.050	0.047	0.043
26	L	5	Yes	0.499	-0.002	-0.002	0.051	0.035	0.021
27	L	10	Yes	0.501	0.003	0.001	0.049	0.038	0.033
28	L	15	Yes	0.499	-0.003	-0.002	0.049	0.041	0.035
29	L	20	Yes	0.500	-0.000	-0.001	0.046	0.041	0.036
30	L	40	Yes	0.500	-0.001	0.001	0.049	0.046	0.044
31	G	5	No	0.501	0.003	0.004	0.047	0.031	0.037
32	G	10	No	0.500	-0.001	-0.001	0.048	0.038	0.045
33	G	15	No	0.500	-0.001	-0.002	0.052	0.046	0.051
34	G	20	No	0.500	-0.000	-0.001	0.049	0.043	0.049
35	G	40	No	0.500	0.000	0.000	0.047	0.044	0.048
36	G	5	Yes	0.501	0.002	0.002	0.050	0.034	0.041
37	G	10	Yes	0.500	-0.001	0.000	0.048	0.037	0.046
38	G	15	Yes	0.500	0.001	0.001	0.053	0.047	0.053
39	G	20	Yes	0.500	0.001	0.002	0.049	0.043	0.049
40	G	40	Yes	0.501	0.001	0.002	0.047	0.044	0.048

Table 27 includes the results of simulating meta-analyses of families of five two-group experiments.

TABLE 27: Two-Group Meta-Analysis Simulation Results

	Data Type	Group Size	Mean Diff	NP Bias	StdMD Bias	NP PMdMRE	StdMD PMdMRE	PHat Observed	Cliffd Observed	StdMD Observed	PHat Power	Cliffd Power	StdMD Power	CliffdPower Difference	PHatPower Difference
1	N	5	S	1.09	3.58	107.14	97.92	0.557	0.113	0.207	0.148	0.138	0.165	-2.770	-1.670
2	N	5	M	0.36	2.54	39.13	39.71	0.639	0.277	0.513	0.470	0.450	0.524	-7.440	-5.470
3	N	5	L	0.22	2.28	25.23	25.62	0.714	0.429	0.818	0.816	0.803	0.862	-5.970	-4.680
4	N	10	S	-1.05	-0.27	71.43	68.66	0.555	0.111	0.199	0.232	0.221	0.254	-3.300	-2.160
5	N	10	M	-0.48	0.36	27.54	27.78	0.637	0.275	0.502	0.753	0.740	0.787	-4.710	-3.420
6	N	10	L	-0.30	0.52	16.82	17.76	0.713	0.427	0.804	0.982	0.980	0.989	-0.900	-0.650
7	N	15	S	0.71	0.96	57.94	55.49	0.556	0.113	0.202	0.312	0.300	0.334	-3.360	-2.110
8	N	15	M	0.18	0.69	22.38	22.67	0.638	0.276	0.503	0.900	0.895	0.919	-2.440	-1.980
9	N	15	L	0.18	0.62	13.40	14.56	0.714	0.429	0.805	0.998	0.998	0.999	-0.110	-0.100
10	N	20	S	0.72	0.72	49.11	47.29	0.556	0.113	0.201	0.385	0.376	0.406	-3.090	-2.180
11	N	20	M	0.19	0.50	19.20	19.28	0.638	0.277	0.503	0.959	0.957	0.970	-1.290	-1.070
12	N	20	L	0.13	0.45	11.68	12.45	0.714	0.429	0.804	1.000	1.000	1.000	0.000	0.010
13	N-H	5	S	2.93	4.55	127.27	124.74	0.545	0.091	0.164	0.119	0.110	0.128	-1.780	-0.920
14	N-H	5	M	0.80	3.20	52.29	50.39	0.610	0.220	0.405	0.337	0.319	0.380	-6.090	-4.300
15	N-H	5	L	0.59	2.70	32.16	32.03	0.672	0.344	0.645	0.628	0.609	0.687	-7.790	-5.920
16	N-H	10	S	-1.97	-0.95	90.91	87.77	0.543	0.086	0.156	0.169	0.160	0.184	-2.420	-1.570
17	N-H	10	M	-0.83	0.29	35.78	35.40	0.608	0.216	0.393	0.568	0.554	0.610	-5.570	-4.240
18	N-H	10	L	-0.38	0.44	21.64	22.46	0.670	0.341	0.631	0.892	0.885	0.922	-3.730	-3.050
19	N-H	15	S	0.68	0.88	74.75	70.28	0.544	0.089	0.158	0.227	0.218	0.241	-2.320	-1.420
20	N-H	15	M	0.25	0.78	29.46	28.33	0.609	0.219	0.395	0.727	0.718	0.770	-5.200	-4.360
21	N-H	15	L	0.28	0.59	18.13	18.04	0.671	0.343	0.632	0.976	0.975	0.985	-1.040	-0.930
22	N-H	20	S	0.49	0.32	63.64	60.35	0.544	0.088	0.158	0.270	0.263	0.294	-3.090	-2.410
23	N-H	20	M	0.20	0.47	25.23	24.43	0.609	0.218	0.394	0.835	0.831	0.862	-3.110	-2.690
24	N-H	20	L	0.21	0.35	15.20	15.65	0.671	0.343	0.630	0.995	0.995	0.997	-0.220	-0.190
25	Lap	5	S	-0.14	4.96	82.86	98.68	0.570	0.140	0.210	0.180	0.170	0.168	0.130	1.220
26	Lap	5	M	0.23	4.25	32.53	40.78	0.666	0.333	0.521	0.603	0.583	0.531	5.220	7.220
27	Lap	5	L	0.02	4.09	20.97	26.56	0.748	0.496	0.833	0.906	0.898	0.862	3.590	4.400
28	Lap	10	S	-0.37	1.88	54.29	68.32	0.570	0.139	0.204	0.304	0.290	0.259	3.120	4.530
29	Lap	10	M	-0.10	1.85	22.89	28.13	0.666	0.332	0.509	0.885	0.876	0.796	8.020	8.880
30	Lap	10	L	-0.38	1.85	13.71	18.43	0.747	0.494	0.815	0.997	0.997	0.987	0.920	0.950
31	Lap	15	S	0.97	2.43	46.03	56.13	0.571	0.141	0.205	0.426	0.415	0.345	7.060	8.090
32	Lap	15	M	0.44	1.70	18.34	23.18	0.667	0.333	0.509	0.970	0.968	0.918	5.010	5.190
33	Lap	15	L	-0.06	1.53	11.29	15.18	0.748	0.496	0.812	1.000	1.000	0.999	0.090	0.090
34	Lap	20	S	-0.83	0.74	40.00	48.91	0.569	0.139	0.201	0.502	0.493	0.408	8.530	9.440
35	Lap	20	M	-0.17	0.85	15.96	20.26	0.666	0.331	0.504	0.993	0.993	0.966	2.650	2.670
36	Lap	20	L	-0.37	0.89	9.88	13.32	0.747	0.494	0.807	1.000	1.000	1.000	0.030	0.030
37	Lap-H	5	S	0.27	5.50	107.14	124.68	0.556	0.112	0.166	0.143	0.132	0.131	0.110	1.140
38	Lap-H	5	M	-0.22	4.82	38.75	50.94	0.635	0.270	0.411	0.456	0.435	0.383	5.180	7.270
39	Lap-H	5	L	-0.08	4.49	26.21	32.84	0.706	0.412	0.656	0.772	0.759	0.696	6.370	7.610
40	Lap-H	10	S	0.73	2.14	71.43	86.76	0.556	0.113	0.160	0.228	0.216	0.194	2.120	3.310
41	Lap-H	10	M	-0.21	2.25	28.41	35.67	0.635	0.270	0.401	0.734	0.721	0.618	10.250	11.570
42	Lap-H	10	L	-0.17	2.13	17.48	23.14	0.706	0.411	0.641	0.971	0.968	0.922	4.610	4.880
43	Lap-H	15	S	1.32	2.24	57.94	71.31	0.557	0.113	0.161	0.310	0.301	0.249	5.130	6.110
44	Lap-H	15	M	0.28	1.83	23.58	29.35	0.636	0.272	0.399	0.878	0.873	0.773	9.980	10.500
45	Lap-H	15	L	0.03	1.58	14.35	18.91	0.706	0.412	0.638	0.997	0.997	0.983	1.390	1.400
46	Lap-H	20	S	-0.80	0.46	50.00	62.03	0.556	0.111	0.158	0.369	0.361	0.295	6.540	7.380
47	Lap-H	20	M	-0.47	0.94	20.30	25.48	0.635	0.270	0.396	0.946	0.944	0.863	8.070	8.320
48	Lap-H	20	L	-0.43	0.92	12.38	16.62	0.705	0.410	0.634	1.000	1.000	0.997	0.250	0.250
49	L	5	S	0.93	8.41	76.00	100.95	0.576	0.151	0.217	0.207	0.192	0.166	2.610	4.140
50	L	5	M	0.25	11.15	28.57	35.78	0.696	0.393	0.556	0.748	0.731	0.611	11.950	13.640
51	L	5	L	0.24	17.89	12.46	23.65	0.846	0.692	0.943	0.999	0.998	0.963	3.540	3.590
52	L	10	S	-1.27	0.83	52.00	69.95	0.574	0.148	0.202	0.342	0.329	0.260	6.870	8.220
53	L	10	M	-0.48	4.25	18.37	24.93	0.695	0.390	0.521	0.957	0.954	0.846	10.750	11.110
54	L	10	L	-0.10	9.20	7.83	17.72	0.845	0.689	0.874	1.000	1.000	0.998	0.200	0.200
55	L	15	S	-0.17	1.32	42.52	56.69	0.575	0.150	0.203	0.463	0.451	0.346	10.490	11.710
56	L	15	M	-0.07	2.50	14.97	20.45	0.696	0.392	0.512	0.995	0.995	0.942	5.220	5.270
57	L	15	L	0.07	5.80	6.28	15.23	0.845	0.690	0.846	1.000	1.000	1.000	0.050	0.050
58	L	20	S	-0.78	-1.72	36.67	48.36	0.574	0.149	0.197	0.557	0.549	0.409	14.010	14.790
59	L	20	M	-0.34	0.30	12.76	18.38	0.695	0.391	0.501	1.000	0.999	0.973	2.620	2.640
60	L	20	L	-0.03	3.49	5.36	13.27	0.845	0.690	0.828	1.000	1.000	1.000	0.010	0.010
61	G	5	S	31.38	63.54	114.81	113.80	0.429	-0.142	-0.327	0.199	0.186	0.289	-10.330	-9.030
62	G	5	M	16.52	36.01	45.19	48.60	0.343	-0.315	-0.680	0.569	0.553	0.688	-13.430	-11.890
63	G	5	L	8.31	22.30	27.44	29.33	0.267	-0.466	-0.978	0.859	0.849	0.929	-8.060	-7.020
64	G	10	S	31.51	58.08	81.48	85.06	0.429	-0.142	-0.316	0.327	0.315	0.428	-11.380	-10.190
65	G	10	M	16.59	31.74	34.81	38.81	0.343	-0.315	-0.659	0.818	0.808	0.888	-8.010	-6.990
66	G	10	L	8.40	18.16	20.93	23.50	0.267	-0.466	-0.945	0.982	0.980	0.993	-1.270	-1.110
67	G	15	S	32.97	59.03	69.55	74.53	0.428	-0.144	-0.318	0.436	0.426	0.542	-11.600	-10.580
68	G	15	M	17.34	31.56	30.70	36.04	0.342	-0.317	-0.658	0.916	0.913	0.955	-4.120	-3.860
69	G	15	L	8.94	17.58	18.45	21.69	0.266	-0.468	-0.941	0.998	0.997	0.999	-0.220	-0.190
70	G	20	S	30.93	55.42	60.19	65.32	0.429	-0.141	-0.311	0.504	0.495	0.599	-10.350	-9.450
71	G	20	M	16.19	29.30	27.41	32.05	0.343	-0.314	-0.647	0.958	0.957	0.981	-2.460	-2.310
72	G	20	L	8.29	15.90	16.74	19.65	0.267	-0.466	-0.927	0.999	0.999	1.000	-0.070	-0.060

Table 28 includes the Type 1 error rates from the simulations of meta-analysis of families of five **two**-group experiments.

TABLE 28
Two-Group Meta-Analysis Type 1 Error Rates

	Data Type	Group Size	PHat Observed	Cliffd Observed	StdMD Observed	PHat Type1 Error Rate	Cliffd Type1 Error Rate	StdMD Type1 Error Rate
1	N	5	0.501	0.002	0.003	0.044	0.037	0.042
2	N	10	0.499	-0.002	-0.002	0.050	0.044	0.048
3	N	15	0.500	0.000	0.001	0.049	0.044	0.050
4	N	20	0.500	0.000	0.000	0.049	0.046	0.047
5	N-H	5	0.501	0.003	0.004	0.046	0.038	0.042
6	N-H	10	0.499	-0.002	-0.003	0.048	0.043	0.046
7	N-H	15	0.500	0.000	0.001	0.048	0.045	0.050
8	N-H	20	0.500	0.000	0.000	0.049	0.046	0.046
9	Lap	5	0.500	0.000	0.002	0.042	0.036	0.039
10	Lap	10	0.500	0.001	0.000	0.045	0.040	0.044
11	Lap	15	0.501	0.002	0.002	0.050	0.045	0.047
12	Lap	20	0.499	-0.001	-0.001	0.051	0.046	0.049
13	Lap-H	5	0.500	0.001	0.003	0.043	0.036	0.040
14	Lap-H	10	0.500	0.000	0.000	0.046	0.040	0.046
15	Lap-H	15	0.501	0.002	0.002	0.047	0.044	0.046
16	Lap-H	20	0.499	-0.001	-0.001	0.051	0.047	0.047
17	L	5	0.501	0.002	0.003	0.044	0.036	0.029
18	L	10	0.499	-0.001	-0.004	0.050	0.044	0.039
19	L	15	0.500	0.000	0.003	0.049	0.044	0.042
20	L	20	0.500	0.000	0.001	0.050	0.047	0.044
21	G	5	0.500	-0.000	-0.001	0.045	0.038	0.051
22	G	10	0.500	0.000	0.001	0.049	0.043	0.049
23	G	15	0.499	-0.002	-0.002	0.049	0.044	0.053
24	G	20	0.500	0.000	0.003	0.048	0.044	0.052

Table 29 includes the results of simulating meta-analyses of families of five **four**-group experiments.

TABLE 29: Four-Group Meta-Analysis Simulation Results

	Data Type	Block Included	Group Size	Mean Diff	NP Bias	StdMD Bias	NP PMdMRE	StdMD PMdMRE	PHat Observed	Cliffd Observed	StdMD Observed	PHat Power	Cliffd Power	StdMD Power	Cliffd Power Difference	PHat Power Difference
1	N	Yes	5	S	0.33	1.28	71.43	66.62	0.556	0.112	0.203	0.211	0.202	0.247	-4.460	-3.540
2	N	Yes	5	M	0.08	1.01	27.54	27.24	0.638	0.276	0.505	0.734	0.722	0.793	-7.100	-5.910
3	N	Yes	5	L	0.11	0.95	15.89	17.47	0.714	0.428	0.808	0.976	0.974	0.990	-1.600	-1.340
4	N	Yes	10	S	0.50	0.56	50.00	47.16	0.556	0.113	0.201	0.368	0.356	0.405	-4.850	-3.650
5	N	Yes	10	M	0.14	0.46	19.57	19.31	0.638	0.276	0.502	0.956	0.953	0.972	-1.940	-1.600
6	N	Yes	10	L	0.08	0.43	11.68	12.48	0.714	0.428	0.803	1.000	1.000	1.000	0.000	0.000
7	N	Yes	15	S	0.13	-0.06	40.48	38.65	0.556	0.112	0.200	0.496	0.487	0.529	-4.280	-3.330
8	N	Yes	15	M	-0.01	0.11	15.62	15.64	0.638	0.276	0.501	0.993	0.993	0.997	-0.380	-0.320
9	N	Yes	15	L	0.01	0.15	9.45	10.03	0.714	0.428	0.801	1.000	1.000	1.000	0.000	0.000
10	N	Yes	20	S	-0.03	-0.19	35.27	33.37	0.556	0.112	0.200	0.599	0.591	0.631	-3.930	-3.150
11	N	Yes	20	M	-0.03	0.03	13.95	13.65	0.638	0.276	0.500	1.000	1.000	1.000	-0.040	-0.040
12	N	Yes	20	L	-0.01	0.09	8.29	8.76	0.714	0.428	0.801	1.000	1.000	1.000	0.000	0.000
13	N-H	Yes	5	S	-0.69	0.07	90.91	86.31	0.544	0.087	0.157	0.156	0.149	0.183	-3.370	-2.700
14	N-H	Yes	5	M	-0.32	0.72	37.61	35.10	0.609	0.217	0.395	0.536	0.524	0.610	-8.570	-7.350
15	N-H	Yes	5	L	-0.13	0.72	21.64	22.34	0.671	0.342	0.633	0.878	0.873	0.924	-5.100	-4.560
16	N-H	Yes	10	S	0.50	0.24	65.91	60.59	0.544	0.088	0.157	0.265	0.257	0.288	-3.060	-2.280
17	N-H	Yes	10	M	0.25	0.48	25.69	24.66	0.609	0.219	0.394	0.822	0.815	0.866	-5.130	-4.380
18	N-H	Yes	10	L	0.24	0.37	15.79	15.58	0.671	0.343	0.630	0.994	0.994	0.997	-0.340	-0.290
19	N-H	Yes	15	S	-0.06	-0.25	51.52	48.93	0.544	0.088	0.157	0.351	0.344	0.382	-3.800	-3.050
20	N-H	Yes	15	M	0.03	0.13	20.49	19.77	0.609	0.218	0.393	0.937	0.935	0.959	-2.370	-2.130
21	N-H	Yes	15	L	0.10	0.07	12.54	12.62	0.671	0.342	0.628	1.000	1.000	1.000	-0.050	-0.050
22	N-H	Yes	20	S	-0.44	-0.85	45.45	43.30	0.544	0.088	0.156	0.430	0.423	0.461	-3.880	-3.140
23	N-H	Yes	20	M	-0.15	-0.12	18.12	17.54	0.609	0.218	0.392	0.981	0.980	0.988	-0.840	-0.760
24	N-H	Yes	20	L	0.00	-0.10	10.96	11.11	0.671	0.342	0.627	1.000	1.000	1.000	0.000	0.000
25	Lap	Yes	5	S	-1.35	1.15	60.00	67.98	0.569	0.138	0.202	0.281	0.269	0.245	2.400	3.560
26	Lap	Yes	5	M	-0.40	1.64	22.89	27.79	0.665	0.331	0.508	0.858	0.851	0.788	6.280	7.060
27	Lap	Yes	5	L	-0.09	1.78	14.17	18.43	0.747	0.494	0.814	0.996	0.995	0.989	0.670	0.700
28	Lap	Yes	10	S	-1.15	0.61	40.00	48.23	0.569	0.138	0.201	0.489	0.476	0.400	7.610	8.920
29	Lap	Yes	10	M	-0.30	0.82	15.66	20.06	0.665	0.331	0.504	0.991	0.990	0.970	1.950	2.030
30	Lap	Yes	10	L	-0.02	0.88	9.72	13.10	0.747	0.494	0.807	1.000	1.000	1.000	0.000	0.000
31	Lap	Yes	15	S	0.45	0.94	32.06	39.87	0.570	0.141	0.202	0.663	0.653	0.538	11.520	12.540
32	Lap	Yes	15	M	0.26	0.73	12.99	16.41	0.666	0.333	0.504	1.000	1.000	0.996	0.400	0.400
33	Lap	Yes	15	L	0.22	0.69	8.05	10.74	0.748	0.495	0.806	1.000	1.000	1.000	0.000	0.000
34	Lap	Yes	20	S	-0.18	0.43	28.21	34.29	0.570	0.140	0.201	0.767	0.759	0.637	12.250	12.970
35	Lap	Yes	20	M	0.06	0.43	11.45	14.24	0.666	0.332	0.502	1.000	1.000	1.000	0.040	0.040
36	Lap	Yes	20	L	0.12	0.45	7.09	9.40	0.747	0.495	0.804	1.000	1.000	1.000	0.000	0.000
37	Lap-H	Yes	5	S	-0.94	1.47	71.43	87.16	0.555	0.111	0.159	0.205	0.197	0.185	1.200	1.950
38	Lap-H	Yes	5	M	-0.61	2.03	29.15	35.56	0.635	0.269	0.400	0.694	0.684	0.609	7.480	8.460
39	Lap-H	Yes	5	L	-0.47	2.03	18.45	23.02	0.705	0.410	0.641	0.964	0.961	0.923	3.810	4.100
40	Lap-H	Yes	10	S	-1.35	-0.04	51.79	61.22	0.555	0.110	0.157	0.358	0.347	0.292	5.500	6.570

Continued on next page

TABLE 29 – continued from previous page

	Data Type	Block Included	Group Size	Mean Diff	NP Bias	StdMD Bias	NP PMdMRE	StdMD PMdMRE	PHat ObsPHat	Cliffd Observed	StdMD Observed	PHat Power Power	Cliffd Power Power	StdMD Power Power	Cliffd Power Difference	PHat Power Difference
41	Lap-H	Yes	10	M	0.76	0.75	20.60	25.15	0.635	0.269	0.395	0.939	0.935	0.862	7.310	7.700
42	Lap-H	Yes	10	L	-0.51	0.80	12.62	16.39	0.705	0.410	0.633	1.000	1.000	0.997	0.260	0.280
43	Lap-H	Yes	15	S	0.67	0.56	41.27	50.72	0.556	0.113	0.158	0.499	0.491	0.392	9.940	10.730
44	Lap-H	Yes	15	M	1.49	0.69	16.77	20.66	0.635	0.271	0.395	0.990	0.990	0.956	3.340	3.400
45	Lap-H	Yes	15	L	-0.17	0.57	10.25	13.43	0.706	0.411	0.632	1.000	1.000	1.000	0.010	0.010
46	Lap-H	Yes	20	S	0.06	0.18	35.71	43.68	0.556	0.112	0.157	0.595	0.587	0.471	11.650	12.430
47	Lap-H	Yes	20	M	1.36	0.47	14.61	17.78	0.635	0.271	0.394	0.999	0.999	0.987	1.150	1.170
48	Lap-H	Yes	20	L	-0.24	0.40	8.86	11.49	0.706	0.411	0.630	1.000	1.000	1.000	0.000	0.000
49	L	No	5	S	-1.09	1.55	52.00	68.06	0.574	0.148	0.203	0.312	0.299	0.249	4.950	6.220
50	L	No	5	M	-0.39	4.60	18.37	24.65	0.695	0.390	0.523	0.946	0.942	0.840	10.120	10.570
51	L	No	5	L	0.06	9.33	7.83	17.83	0.845	0.690	0.875	1.000	1.000	0.998	0.250	0.250
52	L	No	10	S	-0.63	-1.43	37.33	47.99	0.575	0.149	0.197	0.540	0.525	0.402	12.320	13.850
53	L	No	10	M	-0.24	0.42	12.76	17.95	0.696	0.391	0.502	1.000	0.999	0.975	2.430	2.440
54	L	No	10	L	-0.03	3.49	5.51	13.91	0.845	0.690	0.828	1.000	1.000	1.000	0.010	0.010
55	L	No	15	S	-0.50	-2.28	30.07	39.05	0.575	0.149	0.195	0.710	0.702	0.525	17.620	18.490
56	L	No	15	M	-0.19	-1.07	10.43	14.70	0.696	0.391	0.495	1.000	1.000	0.996	0.400	0.400
57	L	No	15	L	-0.01	1.21	4.41	11.67	0.845	0.690	0.810	1.000	1.000	1.000	0.010	0.010
58	L	No	20	S	-1.40	-4.04	26.00	33.89	0.574	0.148	0.192	0.811	0.806	0.614	19.200	19.740
59	L	No	20	M	-0.50	-2.49	9.18	13.33	0.695	0.390	0.488	1.000	1.000	0.998	0.170	0.170
60	L	No	20	L	-0.08	-0.44	3.84	10.28	0.845	0.689	0.796	1.000	1.000	1.000	0.000	0.000
61	L	Yes	5	S	-1.16	2.78	52.00	70.23	0.574	0.148	0.199	0.312	0.299	0.238	6.060	7.390
62	L	Yes	5	M	-0.41	5.69	18.37	25.64	0.695	0.390	0.514	0.946	0.942	0.827	11.470	11.890
63	L	Yes	5	L	0.02	10.67	7.83	18.76	0.845	0.690	0.860	1.000	1.000	0.997	0.330	0.330
64	L	Yes	10	S	-0.25	0.07	37.33	49.76	0.575	0.150	0.194	0.542	0.527	0.395	13.240	14.730
65	L	Yes	10	M	-0.11	1.43	12.76	18.70	0.696	0.392	0.493	1.000	0.999	0.971	2.830	2.840
66	L	Yes	10	L	0.00	4.62	5.51	14.59	0.845	0.690	0.813	1.000	1.000	1.000	0.000	0.000
67	L	Yes	15	S	-0.59	-1.89	30.07	39.97	0.575	0.149	0.190	0.708	0.700	0.503	19.630	20.490
68	L	Yes	15	M	-0.22	-0.53	10.43	15.43	0.696	0.391	0.483	1.000	1.000	0.993	0.660	0.660
69	L	Yes	15	L	-0.02	2.11	4.41	12.32	0.845	0.690	0.793	1.000	1.000	1.000	0.000	0.000
70	L	Yes	20	S	-1.38	-3.46	26.00	35.02	0.574	0.148	0.187	0.812	0.806	0.598	20.780	21.340
71	L	Yes	20	M	-0.48	-1.97	9.18	13.80	0.695	0.390	0.476	1.000	1.000	0.998	0.240	0.240
72	L	Yes	20	L	-0.08	0.33	3.84	11.04	0.845	0.689	0.780	1.000	1.000	1.000	0.000	0.000
73	G	No	5	S	30.97	59.35	85.19	84.32	0.429	-0.141	-0.319	0.309	0.297	0.430	-13.280	-12.060
74	G	No	5	M	16.35	32.57	36.30	39.16	0.343	-0.314	-0.663	0.790	0.781	0.885	-10.440	-9.530
75	G	No	5	L	9.01	18.89	21.50	24.05	0.267	-0.467	-0.951	0.978	0.976	0.993	-1.740	-1.530
76	G	No	10	S	31.33	56.95	62.96	66.10	0.429	-0.142	-0.314	0.495	0.485	0.613	-12.840	-11.820
77	G	No	10	M	16.36	29.85	28.15	32.64	0.343	-0.314	-0.649	0.956	0.953	0.981	-2.850	-2.480
78	G	No	10	L	8.86	16.16	17.29	20.12	0.267	-0.466	-0.929	1.000	1.000	1.000	-0.030	-0.030
79	G	No	15	S	31.17	55.22	53.91	57.32	0.429	-0.142	-0.310	0.616	0.607	0.719	-11.190	-10.310
80	G	No	15	M	16.53	29.21	24.77	29.57	0.343	-0.315	-0.646	0.990	0.989	0.997	-0.760	-0.660
81	G	No	15	L	9.14	15.76	15.89	18.66	0.266	-0.467	-0.926	1.000	1.000	1.000	0.000	0.000
82	G	No	20	S	31.01	55.10	47.69	52.72	0.429	-0.141	-0.310	0.712	0.707	0.795	-8.850	-8.320
83	G	No	20	M	16.27	28.75	23.52	28.18	0.343	-0.314	-0.644	0.998	0.998	0.999	-0.110	-0.110
84	G	No	20	L	8.86	15.19	14.95	18.00	0.267	-0.466	-0.921	1.000	1.000	1.000	0.000	0.000
85	G	Yes	5	S	32.88	59.47	85.71	83.16	0.426	-0.149	-0.332	0.334	0.320	0.450	-13.010	-11.650
86	G	Yes	5	M	15.95	32.46	33.33	38.56	0.337	-0.327	-0.689	0.817	0.808	0.901	-9.300	-8.420
87	G	Yes	5	L	8.32	18.73	20.18	23.77	0.258	-0.483	-0.989	0.981	0.980	0.995	-1.500	-1.360
88	G	Yes	10	S	31.38	56.82	60.71	64.48	0.426	-0.147	-0.326	0.516	0.506	0.633	-12.740	-11.670
89	G	Yes	10	M	15.47	30.02	26.95	32.36	0.337	-0.326	-0.676	0.963	0.961	0.984	-2.330	-2.120
90	G	Yes	10	L	7.99	16.22	16.59	19.89	0.259	-0.482	-0.968	1.000	1.000	1.000	0.000	0.000
91	G	Yes	15	S	31.20	55.83	52.38	56.69	0.427	-0.147	-0.324	0.643	0.635	0.739	-10.400	-9.580
92	G	Yes	15	M	15.46	29.31	24.19	29.51	0.337	-0.326	-0.672	0.991	0.991	0.997	-0.580	-0.570
93	G	Yes	15	L	7.98	15.56	15.20	18.57	0.259	-0.482	-0.963	1.000	1.000	1.000	0.000	0.000
94	G	Yes	20	S	31.38	55.45	47.77	52.15	0.426	-0.147	-0.323	0.733	0.728	0.816	-8.790	-8.280
95	G	Yes	20	M	15.39	28.86	22.52	27.89	0.337	-0.325	-0.670	0.998	0.998	0.999	-0.110	-0.090
96	G	Yes	20	L	7.95	15.19	14.24	17.97	0.259	-0.481	-0.960	1.000	1.000	1.000	0.000	0.000

Table 30 included the Type 1 error rates from simultions of families of five **four**-group experiments.

TABLE 30
Four-Group Meta-Analysis Type 1 Error Rates

	Data Type	Block Included	Group Size	PHat Observed	Cliffd Observed	StdMD Observed	PHat Type1 Error Rate	Cliffd Type1 Error Rate	StdMD Type1 Error Rate
1	N	Yes	5	0.500	-0.001	-0.000	0.040	0.036	0.045
2	N	Yes	10	0.500	0.000	0.000	0.046	0.041	0.048
3	N	Yes	15	0.500	-0.000	-0.000	0.045	0.042	0.046
4	N	Yes	20	0.500	-0.000	-0.001	0.051	0.049	0.051
5	N-H	Yes	5	0.500	-0.001	-0.001	0.043	0.038	0.046
6	N-H	Yes	10	0.500	0.000	0.000	0.047	0.043	0.046
7	N-H	Yes	15	0.500	-0.000	-0.001	0.044	0.042	0.046
8	N-H	Yes	20	0.500	-0.000	-0.001	0.050	0.048	0.050
9	Lap	Yes	5	0.499	-0.002	-0.002	0.040	0.035	0.041
10	Lap	Yes	10	0.499	-0.001	-0.001	0.043	0.039	0.044
11	Lap	Yes	15	0.500	0.001	0.000	0.049	0.045	0.047
12	Lap	Yes	20	0.500	-0.000	-0.001	0.045	0.042	0.048
13	Lap-H	Yes	5	0.500	-0.000	-0.002	0.039	0.034	0.039
14	Lap-H	Yes	10	0.499	-0.001	-0.001	0.046	0.041	0.044
15	Lap-H	Yes	15	0.500	0.001	0.000	0.050	0.046	0.048
16	Lap-H	Yes	20	0.500	-0.000	-0.001	0.045	0.042	0.050
17	L	No	5	0.499	-0.001	-0.001	0.040	0.036	0.035
18	L	No	10	0.500	0.000	-0.001	0.045	0.040	0.042
19	L	No	15	0.500	-0.000	0.001	0.045	0.042	0.042
20	L	No	20	0.500	-0.001	-0.001	0.052	0.050	0.047
21	L	Yes	5	0.499	-0.001	-0.001	0.040	0.036	0.033
22	L	Yes	10	0.500	0.000	0.000	0.045	0.040	0.041
23	L	Yes	15	0.500	0.000	0.000	0.045	0.041	0.045
24	L	Yes	20	0.499	-0.001	-0.001	0.052	0.050	0.048
25	G	No	5	0.500	-0.001	-0.001	0.041	0.036	0.049
26	G	No	10	0.500	-0.000	-0.001	0.047	0.042	0.050
27	G	No	15	0.500	-0.001	-0.001	0.048	0.046	0.052
28	G	No	20	0.500	-0.000	-0.000	0.048	0.044	0.051
29	G	Yes	5	0.499	-0.001	0.000	0.043	0.039	0.048
30	G	Yes	10	0.500	-0.000	0.000	0.048	0.044	0.051
31	G	Yes	15	0.501	0.001	0.001	0.048	0.045	0.052
32	G	Yes	20	0.500	-0.000	-0.001	0.047	0.045	0.050

6 Estimating the Variance of Standardized Mean Difference Effect Sizes

In this section, we discuss how we estimated the variance of the standardized mean difference values calculated in our simulation studies. For this reason, we consider only situations where experiments are completely balanced.

In the case of between-groups experiments with only two groups, the *population* standardized mean difference is defined to be

$$\delta = \frac{\mu_y - \mu_x}{\sigma} \tag{51}$$

where μ_y is the population mean one group, μ_x is the population mean of the other group and σ is the population standard deviation. For other statistical designs, the formula varies, but it is always based on dividing an estimate of the mean difference by the population standard deviation[7].

6.1 The Exact Variance of Estimates of δ

Given the form of Equation (51) we might assume that δ can be estimated from the sample statistics as:

$$d = \frac{m_1 - m_2}{s} \tag{52}$$

where m_1 is the mean value of group 1, m_2 is the mean value of group 2, s is the pooled within-group standard deviation. However, d is a *biased* estimator of δ. For small sample sizes, the best estimate of δ is obtained from the formula:

$$\hat{\delta} = J(df) \times d \tag{53}$$

where:

$$J(df) = \sqrt{\frac{2}{df}}\left(\frac{\Gamma(\frac{df}{2})}{\Gamma(\frac{df-1}{2})}\right) \tag{54}$$

and Γ is the gamma distribution and df is the number of degrees of freedom. In standard textbooks (e.g. , [19]), J(df) is replaced by an approximation:

$$c(df) \approx 1 - \frac{3}{4df - 1} \tag{55}$$

However, since we are dealing with very small sample sizes, we used $J(df)$ in all our simulations.

Assuming a balanced experiment with n independent observations in each group, given that d is related to $t-$statistic:

$$t = \frac{m_1 - m_2}{s\sqrt{\frac{2}{n}}} = d \times \sqrt{\frac{n}{2}} \tag{56}$$

and the variance of a t-variable is:

$$\sigma_t^2 = \left(\frac{df}{df-2}\right)(1+\phi^2) - \frac{\phi^2}{[J(df)]^2} \tag{57}$$

where $\phi = \delta\sqrt{n/2}$, The variance of d is, then, obtained by multiplying σ_t^2 by $\frac{2}{n}$:

$$s_d^2 = \frac{2df}{n(df-2)}\left(1 + \frac{n}{2}\delta^2\right) - \frac{\delta^2}{[J(df)]^2}$$

$$= \frac{df}{(df-2)}\left(\frac{2}{n} + \delta^2\right) - \frac{\delta^2}{[J(df)]^2}$$

which, since $J[df]d$ is an unbiased estimates δ, gives:

$$s_d^2 = \frac{df}{(df-2)}\left(\frac{2}{n} + [J(df)]^2 d^2\right) - d^2 \tag{58}$$

and the variance of $\hat{\delta}$ is obtained by multiplying s_d^2 by $[J(df)]^2$:

$$s_{\hat{\delta}}^2 = \frac{[J(df)]^2 df}{(df-2)}\left(\frac{2}{n} + \hat{\delta}^2\right) - \hat{\delta}^2 \tag{59}$$

For large values of df (where, $(df)/(df-2) \approx 1$) and $J(df) \approx 1$) and $\delta = 0$, both $d \approx 0$ and $\hat{\delta} \approx 0$, so if $n = 10$, $\sigma_{\hat{\delta}}^2 = \sigma_d^2 \approx \frac{2}{10} = 0.2$.

For balanced randomized block designs with two blocks and two treatment conditions, the equivalent equations are

$$s_d^2 = \frac{df}{(df-2)}\left(\frac{1}{n} + [J(df)]^2 d^2\right) - d^2 \tag{60}$$

and

$$s_{\hat{\delta}}^2 = \frac{[J(df)]^2 df}{(df-2)}\left(\frac{1}{n} + \hat{\delta}^2\right) - \hat{\delta}^2 \tag{61}$$

where it is critical to understand that n is the number of observations in each of the four conditions (i.e., each combination of block and treatment). Thus, if we have 5 observations in each of the 4 conditions, with large values of df and $\delta = 0$, $\sigma_{\hat{\delta}}^2 = \sigma_d^2 \approx \frac{1}{5} = 0.2$, i.e., exactly the same variance estimates as the two-group variances.

The more general form of the variances for balanced randomized designs with more blocks (but still assuming equal size groups and no repeated measures in experimental units) is based on the number of observations in each treatment condition, so if $N = N_A = N_B$ where N_A is a number of observations arising from treatment A and N_B is the number of observations arising from treatment B, the equations are:

$$s_d^2 = \frac{df}{(df-2)}\left(\frac{2}{N} + [J(df)]^2 d^2\right) - d^2 \tag{62}$$

and

$$s_{\hat{\delta}}^2 = \frac{[J(df)]^2 df}{(df-2)}\left(\frac{2}{N} + \hat{\delta}^2\right) - \hat{\delta}^2 \tag{63}$$

This means that if we have a family of k experiments, and N is the same for each experiment and we have estimates of $\hat{\delta}$ and d obtained from combining the results of the experiments, the equations are:

$$s_d^2 = \frac{df}{(df-2)}\left(\frac{2}{Nk} + [J(df)]^2 d^2\right) - d^2 \tag{64}$$

and

$$s_{\hat{\delta}}^2 = \frac{[J(df)]^2 df}{(df-2)}\left(\frac{2}{Nk} + \hat{\delta}^2\right) - \hat{\delta}^2 \tag{65}$$

7. In the main text, we used $StdMD$ and $StdMDAdj$, while in this document, we use $d = StdMD$, and $\hat{\delta} = StdMDAdj$ to simplify the equations.

6.2 The Approximate Normal Variance of Estimates of δ

For "large" samples[8], there is also a normal approximation, based on the large sample variance of a $t-$variable [21]:

$$\sigma_{\phi}^2 \approx \left(1 + \frac{\phi^2}{2df}\right) \tag{66}$$

Using the same arguments as before:

$$s_{\delta}^2 \approx \frac{2}{n} + \frac{\delta^2}{2df} \tag{67}$$

and

$$s_{d}^2 \approx \frac{2}{n} + \frac{[J(df)]^2 d^2}{2df} \tag{68}$$

and

$$s_{\delta}^2 \approx [J(df)]^2 \left(\frac{2}{n} + \frac{\hat{\delta}^2}{2df}\right) \tag{69}$$

For four-group randomized block experiments, the equations are:

$$s_{d}^2 \approx \frac{1}{n} + [J(df)]^2 \frac{d^2}{2df} \tag{70}$$

and

$$s_{\hat{\delta}}^2 \approx [J(df)]^2 \left(\frac{1}{n} + \frac{\hat{\delta}^2}{2df}\right) \tag{71}$$

where n is the number of observations in each block and treatment combination. It should be noted that, for large values of df, the two equations deliver similar results. We note that there are disagreements about the formula for the approximate normal variance in the literature [22], but for our simulations, we used the equations reported in this section.

For a family of k experiments all of the same size, the variance estimates are obtained by replacing n by nk.

6.3 The Degrees of Freedom

All the variance formulas reported above depend on the degrees of freedom of the experiment. With fully balanced experiments, no multiple measures on the same experimental unit, and assuming that the population variance for each treatment and block conditions are equal, the degree of freedom is equal to $N - m$ where m is the number of experimental conditions, and N is the total number of observations. Thus, the degrees of freedom for a two-group experiment would be $N - 2$ and for a four-group experiment, it would be $N - 4$.

However, the basic `R` $t-$test and the algorithms Wilcox developed to calculate the variance of $\hat{p}$ and variance of estimates of the mean difference in randomized blocks experiments follow Welch's approach ([23] and [24]) which does *not* assume variance equality. The impact of using Welch's method is that the values of degrees of freedom are decreased if the variances are not equal. This leads to the question of whether the degrees of freedom used in the standardized effect size should be based on the degrees of freedom from the test algorithms or the experimental design structure. We have not found any statistical literature that discusses this issue. In most of the standard statistical texts, the degrees of freedom is specified as $N - 2$; in addition, when the large sample approximate variance formula is used, the degrees of freedom is usually replaced by N. This is justified on the grounds that for large samples $df \approx N$.

However, we are concerned with small samples, where the effect of different choices for the degrees of freedom parameter will have a greater impact on the variance estimate [22]. This would make our simulation results slightly more conservative than studies that use the theoretical degrees of freedom but would affect the standardized effect size and $\hat{p}$ equally. Thus, we decided to use the Welch-based degrees of freedom in our simulations. Furthermore, our simulations investigating the Type 1 error rates suggest that the use of the Welch-based degrees of freedom is, in most cases, close to the theoretical value (i.e., 0.05), which suggests that our decision has not introduced a systematic bias. For families of experiments, we used the sum of the degrees of freedom from each experiment.

7 Aggregating Standardized Mean Differences

The standard meta-analysis of parametric effect sizes is based on using a weighted average of the effect size, where weights are based on the effect size variance. We used the `R` package `metafor` to aggregate d and $\hat{\delta}$ with using a fixed-effects model for our specific example and a randomized-effects model using PM method for our more extensive simulations (see [25] for a discussion of `metafor`).

However, there are several other methods of aggregating standardized effect sizes for families of experiments, and Lin [26] has reported bias using weighted methods for small sample sizes. So, in order to undertake a fair comparison between meta-analysis of nonparametric and parametric, we investigated four other means of aggregating results:

1) The *MDUnweighted* method which is based on calculating the average mean difference across the family of experiments.
2) The *StdMDUnweighted* method which is based on the average of the d values obtained from each experiment.
3) The *StdMDAdjUnweighted* method which is based on the average of the $\hat{\delta}$ values obtained from each experiment.
4) The *HedgesSmallSample* method which Hedges and Olkin [20] recommend for two-group experiments with 10 or fewer observations per group.

These methods are explained below.

7.1 The *MDUnweighted Method*

For each family, we calculated the average mean difference and the variance using the `R` $t-$test for two-group experiments and Wilcox's algorithm for testing linear combinations for four-group experiments. Then, we calculated the average mean difference across the k experiments using the formula:

$$MD_{ave} = \frac{\Sigma(MD_j)}{k} \tag{72}$$

8. We have not seen a specification of "large", but Hedges and Olkin [20] present an example of aggregating small sample-size two-group experiments which is based on experiments with group size $n \leq 10$.

where MD_j is the mean difference for experiment j. We calculated the estimate of the data variance as:

$$s_{ave} = \frac{\Sigma(s_j)}{k} \tag{73}$$

where s_j is the variance of experiment j. Then, we calculated the aggregated estimated of d as:

$$d = \frac{MD_{ave}}{s_{ave}} \tag{74}$$

We calculated the approximate normal variance of d as:

$$s_d^2 = \frac{2}{Nk} + \frac{[J(df)]^2 d^2}{2df} \tag{75}$$

where $df = \Sigma_{i=1}^{k}(df_j)$ where df_j is the Welch-based degrees of freedom from experiment j, k is the number of experiments in the family, J(df) is the exact small sample size adjustment, and N is the number of observations in the treatment and control condition in each experiment.

The $MDUnweighted$ method is an example of an Individual Participant Data (IDP) stratified method which is one of the methods recommended by [27]. We selected this method for our main meta-analysis simulations because:

- It allows us to perform an independent parametric analysis for each experiment without making any assumptions about variance homogeneity either within the individual experiment or across the set of experiments.
- It leaves the construction of the standardized effect size until we have enough degrees of freedom to feel confident that we can omit the small sample adjustment, and we can use the Normal approximation to calculate the effect size variance.

7.2 The *StdMDUnweighted* Method

For each family, we calculated the average mean difference and the variance using the R $t-$test for two-group experiments and Wilcox's algorithm for testing linear combinations for four-group experiments. Then, we calculated the standardized mean difference d_j for each experiment as:

$$d_j = \frac{MD_j}{s_j} \tag{76}$$

Then we used the unweighted mean of the d_j values as our overall estimate of d.

We calculated the approximate normal variance of d as:

$$s_d^2 = \frac{2}{Nk} + \frac{[J(df)]^2 d^2}{2df} \tag{77}$$

where $df = \Sigma_{i=1}^{k}(df_j)$ where df_j is the Welch-based degrees of freedom from experiment j, k is the number of experiments in the family, J(df) is the exact small sample size adjustment, and N is the number of observations in the treatment and control condition in each experiment.

7.3 The *StdMDAdjUnweighted* Method

The *StdMDAdjUnweighted* method is very similar to the *StdMDAdjUnweighted* method but aims to estimate $\hat{\delta}$. For each family, we calculated the average mean difference and the variance using the R $t-$test for two-group experiments and Wilcox's algorithm for testing linear combinations for four-group experiments. Then, we calculated the standardized mean difference $\hat{\delta}_j$ for each experiment as:

$$\hat{\delta}_j = J(df)\frac{MD_j}{s_j} \tag{78}$$

Then we used the unweighted mean of the $\hat{\delta}_j$ values as our overall estimate of $\hat{\delta}$.

We calculated the approximate normal variance of $\hat{\delta}$ as:

$$s_{\hat{\delta}}^2 = [J(df)]^2 \left(\frac{2}{Nk} + \frac{\hat{\delta}^2}{2df}\right) \tag{79}$$

where $df = \Sigma_{i=1}^{k}(df_j)$ where df_j is the Welch-based degrees of freedom from experiment j, k is the number of experiments in the family, J(df) is the exact small sample size adjustment, and N is the number of observations in the treatment and control condition in each experiment.

7.4 The *HedgesSmallSample* Method

The *HedgesSmallSample* method is a weighted aggregation method applied to the $\hat{\delta}_j$ values (see [20], pp 129-131). The weights are based on the variance of each $\hat{\delta}_j$ value, but the variance of each $\hat{\delta}_j$ is calculated using the unweighted mean of $\hat{\delta}_j$ as the estimate of δ in the exact variance equation for the appropriate statistical design. In the case of our simulations, all the experiments in a specific family were of the same size and the same design, so the only factor that influenced the values of the weights was differences between the Welch-based degrees of freedom among the members of the same family.

To summarize, the small sample methods require calculating the unweighted average $\bar{d}$:

$$\bar{d} = \frac{\Sigma\hat{\delta}_j}{k} \tag{80}$$

and recalculating the variance (v_j) of each experiment using $\bar{d}$ in Equation (59). Then, the estimate of the aggregated effect size is:

$$\hat{\delta} = \Sigma\frac{\delta_j}{v_i} \tag{81}$$

with variance:

$$s_{\hat{\delta}}^2 = \left[\Sigma\left(\frac{1}{v_j}\right)\right]^{-1} \tag{82}$$

We calculated the estimates of $\hat{\delta}$ and its variance can be calculated by using Equation (81) and Equation (82). However, they can also be calculated using the `R metafor` function `rma` with the parameter `method` set to *FE*.

8 Simulation Functions and Methods

In this section, we discuss a number of issues related to the reproducibility of our simulation results.

8.1 Simulation Problems

We found some problems when undertaking our simulations. We have already mentioned issues with zero nonparametric effect size variances (see Section 2.4.3) and the degrees of freedom that should be used when calculating the standardized effect size variance(see Section 6.3).

There were also two minor issues that might affect anyone trying to reproduce our results:

1) The `metafor` package `rma` function sometimes failed to converge for random effects meta-analysis.
2) Attempts to generate gamma distribution samples sometimes failed.

Both these issues appeared to be a side-effect of using very small sample sizes in four-group meta-analysis simulations. We provided a workaround for both by putting calls to `rma` and calls to our function for generating experiment data into `while` loops that detected failures and performed another data extraction with a new seed to obtain a new set of data while maintaining the total required number of repetitions.

8.2 Significance Tests

In the main text [1], we argued that we should use one-sided tests for simulation studies. The rationale for this was that when we ourselves have set the difference between the control and treatment condition in one direction, we should *not* consider a statistically significant effect size in the wrong direction as *correctly rejecting the null hypothesis*. Furthermore, such an occurrence (whether for parametric or nonparametric effect size) should *not* increase the count of statistically significant effect sizes used to estimate power.

However, a problem with using one-sided tests is that only the `R t.test` function correctly implemented one-sided tests. For analysis functions that did not support one-sided tests, we used a method based on confidence intervals to implement one-sided tests depending on the information provided by the statistical analysis function we were using. We explain the methods we used by reference to an example using the `R` language `t.test` functions.

Figure 2 shows a `R` listing of a two-sided test and a one-sided test based on the same generated data sets. In both cases, the $t - value$ is the same (as are the two-group mean values). Furthermore, since the t values are the same, the standard error of the mean difference MD, the pooled variance of the data (s^2), the standard error of the mean se and the standardized mean difference ($StdMD$) are unaffected by choice of the significance level. In fact, for this example, we have:

$$MD = 0.4487428 - (-0.0583762) = 0.507119$$

$$StdMD = t\sqrt{\frac{2}{n}} = 1.7737\sqrt{\frac{2}{20}} = 0.5608932$$

$$se = MD/t = 0.507119/1.7737 = 0.2859102$$

$$s^2 = se^2 \times \frac{n}{2} = 0.2859102^2 = 0.8174464$$

All these statistics can be obtained from the output from the `R` function `t.test`, if the output is saved into a new variable. For example, using the `R` statement `output=t.test(b, a, "greater")`, the standard error can be obtained by subsequently using the statement `output$stderr`.

```
set.seed(123)
a=stats::rnorm(20,0,1)-0.2
b=stats::rnorm(20,0.5,1)
#Two-sided test
#Confidence interval defaulted to 95%
stats::t.test(b,a)

     Welch Two Sample t-test

data:  b and a
t = 1.7737, df = 37.082, p-value =
0.08432
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 -0.07214401  1.08638208
sample estimates:
 mean of x  mean of y
 0.4487428 -0.0583762

#One-sided test
#Confidence interval 95% one-sided


stats::t.test(b,a,"greater")

     Welch Two Sample t-test

data:  b and a
t = 1.7737, df = 37.082, p-value =
0.04216
alternative hypothesis: true difference
in means is greater than 0
95 percent confidence interval:
 0.02479152        Inf
sample estimates:
 mean of x  mean of y
 0.4487428 -0.0583762
```

Fig. 2. Example of two-sided and one-sided tests

In this example, the $p - value$ for the two-sided test is $p = 0.08432$, which means that we cannot reject the assumption that the data sets come from the same population. In contrast, the $p - value$ for the one-sided test is $p = 0.04216$, indicating that the difference between the two-group means is significant at the 0.05 level.

Another important difference between the two $t-$test analysis results is that the confidence interval for the mean difference for the two-sided $t.test$ has defined upper and lower limits, but the confidence interval for the one-sided $t.test$ only has a defined lower limit. In the two-sided test, the upper and lower intervals are determined from the 0.025 and the 0.0975 quantiles of the t distribution. Since the confidence limit includes zero, we cannot reject the hypothesis that the data sets come from the same distribution. However, the one-sided test is based on the 0.05 quantile

of the t-distribution. Since the lower bound is greater than zero, we can reject the hypothesis that the mean difference is zero at the 0.05 level.

There are several different ways to assess the significance of a $t-$test:

- Compare the t-value with its *critical value*.
- Check whether the $p-value$ is less than the required alpha level.
- Check whether the confidence interval excludes zero (or any other value consistent with the null hypothesis).

All these methods depend on whether a one-sided or two-sided test is required, and as the example shows, it is possible to have an experiment where a two-sided test would not reject the null hypothesis, but a one-sided test would reject the null hypothesis.

In our simulation studies of individual experiments, for all tests, except those using the `R` language `t.test` function, we used the confidence interval method to assess significance. For Cliff's d, we used a confidence interval based on the normal distribution as recommended in [28]. For $\hat{p}$, we used confidence intervals based on the t-distribution as recommended by Brunner and Munzel [9].

For two-sided tests, we constructed the $\alpha/2$ confidence intervals and checked whether the confidence interval included the null hypothesis value (i.e. 0 for parametric effect sizes and Cliff's d and 0.5 for $\hat{p}$).

For one-sided tests, we constructed the α confidence interval. Then, for positive effect sizes, we identified the effect size as significantly greater than zero if the *lower* confidence interval limit was greater than the null hypothesis value, which is 0 for Cliff's d and$StdMD$, and 0.5 for $\hat{p}$. For negative effect sizes, we identified the effect size as significantly less than zero if the *upper* confidence interval limit was less than the null hypothesis value.

For simulations of meta-analysis of families of experiments, we used the same process for confidence interval construction and significance testing, with the exception of testing the significance of $stdMD$ and $StdMDAdj$ obtained formal meta-analysis. In this situation, we used confidence intervals based on the normal distribution rather than confidence intervals based on the $t-$distribution. This method would be used in situations where details of the heterogeneity-adjusted degrees of freedom were not available. For small sample sizes, it leads to slightly wider confidence intervals than would be expected from confidence intervals based on the $t-$distribution.

8.3 Reproducing the Simulations

In this section, we explain how to reproduce our simulation results using functions available in our `R` language `reproducer` package.

To make sure you use the correct version of `reproducer` remove any existing version and re-install using the following commands:

```
utils::remove.packages("reproducer")
utils::install.packages("reproducer")
```

8.3.1 Analysis of Existing Software Datasets

In Section 2 of the main text, we reported the analysis of existing software engineering data sets. To obtain the first row in Table 1, use the commands:

```
File = reproducer::KitchenhamEtAl.
↪  CorrelationsAmongParticipants.Scanniello15EMSE
reproducer::crossoverResidualAnalysis(File,
↪  StudyID="S1", ExperimentNames=c("USB2"),
↪  Type=c("4G"), Metrics=c("Correctness", "Time",
↪  "Efficiency"))
```

This will deliver a table with 3 rows[9], and the row with Metrics="Time" is the first row in Table 1. To obtain information about the other data sets see [29].

8.3.2 The Laplace Distribution Data

The `R` language does not provide a function to generate Laplace-distributed data. To generate Laplace data, use the function `LaplaceDist`.

If you want to reproduce the Laplace distribution graph shown on the right-hand-side of Figure 1, use the command:

```
z = reproducer::LaplaceDist(1000, 0, 1)
hist(z, freq = F, xlab = "Laplace Distribution 1000
↪  Observations", main = "Histogram and Kernel
↪  Density Plot", ylim = c(0, 0.5))
```

Then, to obtain the distribution statistics, use the command:

```
reproducer::AnalyseResiduals(z,"Laplace1000")
```

8.3.3 Calculating Nonparametric Effect Sizes

The `R` package algorithms `PHat.test` and `Cliffd.test` will calculate the vaues of the non-parametrric effect sizes and their variances for data obtained from a two-group experiment. The algorithms also calculates the confidence interval for the effect sizes and performs one-siderd or two-sided statistical tests.

With data from two data vectors x and y, to test whether $y > x$. based on $\hat{p}$, use the function:

```
reproducer::PHat.test(x, y, alternative="greater")
```

To test whether $y > x$ using Cliff's d use the function:

```
reproducer::Cliffd.test(y, x, alternative="greater")
```

Warning: The change in the order of the x and y variables in the two algorithms is intentional. This is to remain consistent with the interface of the functions defined by Wilcox.

If you want to perform a one-sided test that $y < x$, change the value of the parameter `alternative` to `"less"`. For a two-sided test, either default the parameter $alternative$, or set it to `"two.sided"`.

The α-level of the test and confidence interval width depends on the parameter `alpha`, which defaults to 0.05. It returns the $100(1-\alpha)$, the default being the 95% confidence interval.

For a randomized blocks experiment with two treatments and two blocks, assuming we have four vectors with x1 corresponding to treatment 1 in block 1, y1 corresponding to treatment 2 in block 1, x2 corresponding to treatment 1 in block 2, and y2 corresponding to treatment 2 in block 2, the following function will test whether treatment 1 data values are less than treatment 2 data values:

```
reproducer::Calc4GroupNPStats(x1, y1, x2, y2,
↪  alpha=0.05, alternative="less")
```

9. Copy and paste from pdf is error-prone. If you cut and paste the commands, please make sure you remove any spurious spaces.

This function returns estimates of $\hat{p}$, Cliff's d a together with their variances and confidence intervals. It also returns the point bi-serial Kendall's tau.

8.3.4 Identifying Theoretical Parametric Effect Sizes

In Section 4, we presented the theoretical parametric effect sizes for normal, log-normal, gamma, and Laplace distributed data based on the parameter values we used in our simulation studies. To reproduce the σ^2 and $StdMD$ values reported in row 5 of Table 13 use the command:

```
reproducer::RandomizedDesignEffectSizes(m1=0,
↪  std1=1, m2=0, std2=1.5, type = "l")
```

To obtain the values found in other rows, change the values of the parameters `m2` and `std2` appropriately. Change the `type` parameter to obtain the results for other distributions. Set the parameter `type` to "n" for normal data, "g" for gamma data, and "lap" for Laplace data.

For four-group experiments, use the `RandomizedBlockDesignEffectSizes` command. For example, to reproduce the σ^2 and $StdMD$ entries in row 10 in Table 14, use the command:

```
reproducer::RandomizedBlockDesignEffectSizes(m1=0,
↪  std1=1, m2=0.266, std2=1, m3=0, std3=1,
↪  m4=0.266, std4=1, BE=0.5, type="l")
```

In all cases except the gamma distribution, variance instability is modelled by increasing the parameters `std2` and `std4` by 0.5, and non-zero block effects are modelled by setting the `BE` parameter to 0.5. In the case of the Gamma distribution, we model only the block effect, as an increase to the value of the `shape` parameter, because the gamma distribution does not have a parameter equivalent to a variance parameter. A non-zero `BE` parameter increases the treatment mean values (i.e., `m2` and `m4`).

8.3.5 Large Sample Size Nonparametric Effect Sizes

In Section 4, we presented the large sample size effect sizes for Normal, log-normal, gamma, and Laplace distributed data based on the parameter values we used in our simulation studies. This was mainly to show the large sample size nonparametric effect sizes for the different distributions. For example, for two-group gamma distributions, the following command will produce the large sample size nonparametric effect sizes corresponding to a mean difference of -0.2 on the raw data scale:

```
reproducer::
↪  calculateLargeSampleRandomizedDesignEffectSizes(
↪  meanC=1, sdC=3, diff=0.1223, N=10000000,
↪  type="g", StdAdj = 0)
```

The equivalent command for four-group experimental designs, including a non-zero block effect, is:

```
reproducer::
↪  calculateLargeSampleRandomizedBlockDesignEffectSizes(
↪  meanC=1, sdC=3, diff=0.1223, N=10000000,
↪  type="g", Blockmean=0.5, StdAdj=0)
```

Note. Even with ultra-large samples, there may be some very small disagreements of the order of 0.001, unless you use the same seed. However, for our purposes, this level of disagreement is irrelevant.

8.3.6 Analysis of Individual Experiments

To generate the results reported in rows 1-3 of Table 6 in the main text, use the command:

```
reproducer::calculate2GBias( mean=0,
↪  sd=1,diff=c(0.2,0.5,0.8),
↪  Expected.StdMD=c(0.2,0.5,0.8),
↪  Expected.PHat=c(0.556,0.638,0.714),
↪  N=5,reps=10000, type="n", seed=123, StdAdj =0)
```

`calculate2GBias` requires the user to identify the theoretical standardized mean difference effect size for each mean difference being simulated and the large sample size values of $\hat{p}$. These values are reported in [1] and Section 4, and the functions in `reproducer` that will provide the required values are described in Section 8.3.4 and Section 8.3.5.

In addition, to fully reproduce the results tables, the bias and MdMRE values must be multiplied by 100, and the non-parametric effect size power difference (PD) must be calculated as follows:

$$CliffdPD = (CliffdPower - StdESPower)100 \quad (83)$$

and

$$PHatPD = (PHatPower - StdESPower)100 \quad (84)$$

To obtain the results for other sample sizes, change the parameter `N` (which defines the group size) and the `seed` parameter.

To obtain the results for other distributions, you need to change the `type` parameter (which defines the distribution type), the `diff` parameter (which defines the values of the treatment parameter), and the `seed` parameter and the Expected.StdMD and Expected PHat parameters.

Warning: Undertaking 10000 simulations per condition takes a long time. On a Macbook with a 2 GHz Intel Core I7 Processor, the above command took 6 minutes 46 seconds to execute. The command constructed 3 table entries; however, the full table has 105 entries.

To generate the results of the four-group design individual experiment simulations shown in rows 1-3 of Table 25, use the command:

```
reproducer::calculate4GBias(mean=0,sd=1,
diff=c(0.2,0.5,0.8), Expected.StdMD=c(0.2,0.5,0.8),
Expected.PHat=c(0.556,0.638,0.714),
N=5, reps=10000, type="n", seed=17+123, StdAdj =0,
Blockmean=0.5)
```

To generate the other entries in the table, change the parameters, `N`, `type`, `diff`, `Expected.StdMD`, and `Expected.PHat` appropriately.

You should change the seed value for each simulation function call, although it will not make much difference to the results if you do not use the same seeds as we have used.

To calculate the Type 1 error rates for two-group experimental design simulations use the following command:

```
reproducer::calculate2GType1Error(mean=0,
↪  sd=1,N=5,reps=10000,type="n",seed=156,StdAdj =
↪  0)
```

This will generate the results reported in the first row of Table 7 in the main text. The values reported in other rows can be obtained by changing the values of the `type`, `seed`, and `N` parameters.

To calculate the Type 1 error rates for the four-group experimental design use the command:

```
reproducer::calculate4GType1Error(mean=0,sd=1,N=5,
reps=10000,type="n",seed=17+156,StdAdj = 0,
Blockmean=0.5)
```

This will reproduce the values in the first row of Table 26.

8.3.7 Meta-Analysis Example

To generate the example data shown in Table 12 of the main text [1], use the command:

```
reproducer::NP2GMetaAnalysisSimulation(mean=0, sd=1,
↪  diff=0.8, GroupSize=5, Exp=5, type="n",
↪  StdAdj=0, alpha=0.05, seed=457, StdExp=0,
↪  MAMethod="FE", returnES=TRUE)
```

To generate the results of the meta-analysis of the example data shown in Table 13 of the main text, use the command:

```
reproducer::NP2GMetaAnalysisSimulation(mean=0, sd=1,
↪  diff=0.8, GroupSize=5, Exp=5, type="n",
↪  StdAdj=0, alpha=0.05, seed=457, StdExp=0,
↪  MAMethod="FE", returnES=FALSE)
```

Note, however, that the information is not in the same format as the table.

We provide two functions to meta-analyse the non-parametric effect sizes. To try these out you can use the $\hat{p}$ and Cliff's d data shown in Table 12.

The following comands will meta-analyse the $\hat{p}$ values:

```
PHatMean <- c(0.92,0.60,0.48,0.72,0.88)
PHatMeanVar <- c(0.01,0.04,0.05,0.04,0.01)
PHatDF <- c(6.63,6.63,5.08,5.61,8)
reproducer::metaanalyse.PHat(PHat=PHatMean,
↪  PHatvar=PHatMeanVar, DFUnknown=FALSE, df=PHatDF)
```

The aggregate estimate of $\hat{p}$ and its variance reported by the function will correspond to those associated with the Average method in Table 13 of the main text.

```
Cliffd <- c(0.84,0.2,-0.04,0.44,0.76)
Cliffdvar <- c(0.04,0.18,0.21,0.15,0.06)
```

The following commands with meta-analyse Cliff's d values:

```
reproducer::metaanalyse.Cliffd (Cliffd=Cliffd,
↪  Cliffdvar=Cliffdvar, df=0,
↪  alternative="greater")
```

The aggregate estimate of Cliff's d and its variance reported by the function will correspond to those associated with the Average method in Table 13 of the main text.

8.3.8 Meta-Analysis Simulations

To generate lines 49-51 of the meta-analysis results table for two-group experiments reported in Table 27, use the command:

```
reproducer::calculateMABias(mean=0, sd=1, N=5,
↪  diff=c(0.266,0.72375,1.43633), Experiments=5,
↪  reps=10000, Expected.StdMD=c(0.2,0.5,0.8),
↪  Expected.PHat=c(0.575,0.696,0.845), type="l",
↪  FourG=F,seed= 13 + 1665, StdAdj = 0,
↪  Blockmean=0, StdExp=0.5)
```

Warning. The simulation of families of experiments takes longer than the simualtion of individual experiments. The above command generates 4 rows in Table 27, which comprises 72 rows, and took 1 hour 32 minutes and 10 seconds to execute, on a Macbook with a 2 GHz Intel Core I7 Processor.

The final four rows in the Type 1 error rates table for the two-group experiment meta-analysis simulations (see Table 28) are obtained using the command:

```
reproducer::calculateMAType1Error(mean=1, sd=3,
↪  N=c(5,10,15,20), reps=10000, type="g",
↪  Experiments=5, FourG=F, StdAdj=0, Blockmean=0.5,
↪  seed= 13+1013, StdExp=0.5)
```

To generate lines 46-48 of the meta-analysis results table for four-group experiments reported in Table 29, use the command:

```
reproducer::calculateMABias(mean=0, sd=1, N=20,
↪  diff=c(0.283,0.707104,1.131374), Experiments=5,
↪  reps=10000, Expected.StdMD=c(0.157,0.392,0.628),
↪  Expected.PHat=c(0.556,0.636,0.705), type="lap",
↪  FourG=T,seed = 1565, StdAdj = 0.5,
↪  Blockmean=0.5, StdExp=0.5)
```

The first four rows in Type 1 error rates table for the four-group design meta-analysis simulations (see Table 30) are obtained using the command:

```
reproducer::calculateMAType1Error(mean=0, sd=1,
↪  N=c(5,10,15,20), reps=10000, type="n",
↪  Experiments=5, FourG=T, StdAdj=0, Blockmean=0.5,
↪  seed=313, StdExp=0.5)
```

REFERENCES

[1] B. Kitchenham and L. Madeyski, "Recommendations for Analysing and Meta-Analysing Small Sample Size Experiments," 2023.

[2] V. Basili, F.Shull, and E. Lanubile, "Building knowledge through families of experiments," *IEEE Transactions on Software Engineering*, vol. 25, no. 4, pp. 456–473, 1999.

[3] L. Madeyski, B. Kitchenham, and T. Lewowski, *reproducer: Reproduce Statistical Analyses and Meta-Analyses*, 2023, R package. [Online]. Available: https://cran.r-project.org/web/packages/reproducer/reproducer.pdf

[4] K. McGraw and S. Wong, "A common language effect size statistic," *Psychological Bulletin*, vol. 111, pp. 361–265, 1992.

[5] A. Varga and H. D. Delany, "A Critique and Improvement of the Common Language Effect Size Statistics of McGraw and Wong," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 2, pp. 101–132, 2000.

[6] R. R. Wilcox, *Introduction to Robust Estimation & Hypothesis Testing*, 3rd ed. Elsevier, 2012.

[7] V. W. Rahlfs, H. Zimmermann, and K. R. Lees, "Effect size measures and their relationships in stroke studies," *Stroke*, vol. 45, pp. 627–633, 2013.

[8] B. Kitchenham, L. Madeyski, D. Budgen, J. Keung, P. Brereton, S. Charters, S. Gibbs, and A. Pohthong, "Robust Statistical Methods for Empirical Software Engineering," *Empirical Software Engineering*, vol. 22, no. 2, pp. 579–630, 2017. [Online]. Available: https://doi.org/10.1007/s10664-016-9437-5

[9] E. Brunner and U. Munzel, "The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation," *Biometrical Journal*, vol. 42, pp. 17 – 25, 2000.

[10] B. L. Welch, "The Significance of the Difference Between Two Means when the Population Variances are Unequal," *Biometrika*, vol. 29, no. 3-4, pp. 350–362, 1938.

[11] M. Neuhäuser, C. Lösch, and K.-H. Jöckel, "The Chen–Luo test in case of heteroscedasticity," *Computational Statistics & Data Analysis*, vol. 51, pp. 5055—5060, 2007.

[12] M. A. Babar, B. Kitchenham, L. Zhu, I. Gorton, and R. Jeffery, "An empirical study of groupware support for distributed software architecture evaluation process," *Journal of Systems and Software*, vol. 79, pp. 912–925, 2006.

[13] L. Madeyski and B. Kitchenham, "Effect Sizes and their Variance for AB/BA Crossover Design Studies," *Empirical Software Engineering*, vol. 23, no. 4, pp. 1982–2017, 2018. [Online]. Available: https://doi.org/10.1007/s10664-017-9574-5

[14] S. Vegas, C. Apa, and N. Juristo, "Crossover Designs in Software Engineering Experiments: Benefits and Perils," *IEEE Transactions on Software Engineering*, vol. 42, no. 2, pp. 120–135, 2016.

[15] B. Kitchenham, L. Madeyski, and P. Brereton, "Meta-analysis for families of experiments in software engineering: a systematic review and reproducibility and validity assessment," *Empirical Software Engineering*, vol. 25, no. 1, pp. 353–401, 2020. [Online]. Available: https://doi.org/10.1007/s10664-019-09747-0

[16] S. Senn, *Cross-over Trials in Clinical Research*, 2nd ed. John Wiley and Sons, Ltd., 2002.

[17] S. B. Morris and R. P. DeShon, "Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs," *Psychological Methods*, vol. 7, no. 1, pp. 105–125, 2002.

[18] B. Londeix, *Cost estimation for Software Development*. Addison-Wesley Publishers Ltd., 1987.

[19] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. T. Rothstein, *Introduction to Meta-Analysis*. John Wiley and Sons Ltd., 2009.

[20] L. V. Hedges and I. Olkin, *Statistical methods for meta-analysis*. Orlando, Florida, USA: Academic Press, 1985.
[21] N. L. Johnson and B. L. Welch, "Applications of the non-central t-distribution," *Biometrika*, vol. 31, no. 3-4, pp. 362–389, 1940.
[22] B. Kitchenham and L. Madeyski, "Inconsistencies with formulas for the standard error of the standardized mean difference of repeated measures experiments," *Statistics in Medicine*, vol. 39, pp. 4101—-4104, 2020.
[23] B. Welch, "The significance of the difference between two means when the population variances are unequal," *Biometrika*, vol. 29, no. 3/4, pp. 350–362, 1938.
[24] B. Welch, "The generalization of "student's" problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.
[25] W. Viechtbauer, "Conducting meta-analyses in R with the metafor package," *Journal of Statistical Software*, vol. 36, no. 3, pp. 1–48, 2010.
[26] L. Lin, "Bias caused by sampling error in meta-analysis with small sample sizes," *PLoS ONE*, vol. 13, no. 9, 2018.
[27] A. Santos, O. Gómez, and N. Juristo, "Analyzing Families of Experiments in SE: A Systematic Mapping Study," *IEEE Transactions on Software Engineering*, vol. 46, no. 5, pp. 566–583, 2020.
[28] N. Cliff, *Ordinal Methods for Behavioral Data Analysis*. New York: Psychology Press, 2014.
[29] B. Kitchenham, L. Madeyski, G. Scanniello, and C. Gravino, "The Importance of the Correlation in Crossover Experiments," *IEEE Transactions on Software Engineering*, vol. 48, no. 8, pp. 2802–2813, 2022. [Online]. Available: https://doi.org/10.1109/TSE.2021.3070480