Comparative analysis of meta-analysis methods: when to use which?

Oscar Dieste	Enrique Fernández	Ramón García Martínez	Natalia Juristo
Universidad Politécnica de Madrid	Universidad Nacional de La Plata	Universidad Nacional de Lanús	Universidad Politécnica de Madrid
odieste@fi.upm.es	enriquefernandez@educ.ar	rgarcia@unla.edu.ar	natalia@fi.upm.es

Abstract*— *Background: *Several meta-analysis methods can be used to quantitatively combine the results of a group of experiments, including the weighted mean difference, statistical vote counting, the parametric response ratio and the non-parametric response ratio. The software engineering community has focused on the weighted mean difference method. However, other meta-analysis methods have distinct strengths, such as being able to be used when variances are not reported. There are as yet no guidelines to indicate which method is best for use in each case* ***Aim:*** *Compile a set of rules that SE researchers can use to ascertain which aggregration method is best for use in the synthesis phase of a systematic review.* ***Method:*** *Monte Carlo simulation varying the number of experiments in the meta-analyses, the number of subjects that they include, their variance and effect size. We empirically calculated the reliability and statistical power in each case* ***Results:*** *WMD is generally reliable if the variance is low, whereas its power depends on the effect size and number of subjects per meta-analysis; the reliability of RR is generally unaffected by changes in variance, but it does require more subjects than WMD to be powerful; NPRR is the most reliable method, but it is not very powerful; SVC behaves well when the effect size is moderate, but is less reliable with other effect sizes. Detailed tables of results are annexed.* ***Conclusions:*** *Before undertaking statistical aggregation in software engineering, it is worthwhile checking whether there is any appreciable difference in the reliability and power of the methods. If there is, software engineers should select the method that optimizes both parameters.*

Keywords-component: Meta-analysis, reliability, statistical power, effect size, weighted mean difference (WMD), response ratio (RR), vote counting.

I. Introduction

The results of several experimental studies can be aggregated (or combined) through quantitative synthesis [1] (also known as research synthesis [2] or meta-analysis [3]). Aggregated results are more reliable (and potentially more generally applicable) than individual experiment results. Quantitative synthesis is a common practice in disciplines with a strong experimental tradition, such as medicine, psychology or physics.

The synthesis method used in experimentally mature disciplines is the weighted mean difference (WMD) [4]. For the WMD method to be reliable, the set of studies to be aggregated has to meet two strict constraints: (i) there must be a minimum number of subjects per treatment comparison (WMD's performance has been investigated for over 10 experimental subjects [4]), and (ii) certain statistical parameters (means, variances or standard deviations and number of experimental subjects) must be reported.

WMD's constraints considerably restrict its applicability in the current experimental state of SE. In SE there tend to be few subjects per experiment [5], less than 10 in many cases, and some studies do not provide the statistical parameters required for meta-analysis [5] (since they are not strictly necessary to describe the statistical data analysis of the results of a single experiment as, for example, when ANOVA is used).

Although WMD is the most widespread meta-analysis technique, it is not the only method for quantitative synthesis [6]. Hedges [4] proposes statistical vote counting (SVC) as a less restrictive alternative to WMD, whereas Gurevic & Hedges [7] propose the parametric response ratio (PRR) and non-parametric response ratio (NPRR) as alternative methods to WMD. These methods enable quantitative synthesis under conditions where WMD is not applicable, such as when variances or even treatment means are not reported.

PRR and NPRR have been widely used in other experimental disciplines like ecology [8] and education [9]. In SE, SVC was discussed in Pickard et al. [10], and Conradi appears to have used vote counting in [11], but the technique applied in that paper is a non-statistical version of vote counting, which is much less precise than the statistical form. To the best of our knowledge, PRR and NPRR have not yet been used to combine SE experiments.

As several quantitative synthesis methods are available to SE experimental researchers, they need to make a decision on which to apply under particular aggregation circumstances: Which method is preferable under equal conditions?

If some of a set of experiments for aggregation can be combined using WMD (or PRR) but others cannot, is it better to combine some studies using WMD (or PRR) or all of them using NPRR (or SVC)?

Other disciplines have investigated the relative performance of the synthesis methods for their experimental conditions. Lajeunesse [12] observed for ecology that the type I and II error rates of WMD and SVC were unacceptable in combinations of few experiments (≈5), whereas PRR and NPRR rates were within high but reasonable bounds. Friedrich and colleagues [13] demonstrated for medicine that WMD and PRR methods were equivalent under a range of conditions (small, medium and large effect sizes, several variances, etc.).

However, such findings cannot be directly imported to SE, as the conditions under which the methods were assessed were particular to the disciplines researched, and both have essentially different experimental settings than SE. In particular, experiments in SE have quite small sample sizes [5], and experience has shown that the number of experiments per meta-analysis is also small in many cases (e.g., [6]).

We have performed a comparative analysis of meta-analysis methods for SE experiments. We study the reliability and statistical power of the different quantitative synthesis methods through a simulation process using similar conditions to those current in SE experiments today.

This paper is structured as follows. In Section 2, we review existing quantitative synthesis methods. In Section 3, we detail the limitations of the meta-analysis methods and describe earlier work on analyzing performance in other experimental disciplines. Section 4 describes the research methodology that we follow. In Section 5 we describe the simulation processes. Section 6 presents the results of simulating the application of the synthesis methods under different conditions. In Section 7, we present guidelines to determine which method to use. Section 8 outlines the limits of our work. And, finally, Section 9 discusses our conclusions.

II. Quantitative Synthesis Methods

The quantitative synthesis of experiments involves aggregating the result of a set of studies by analyzing the performance of a pair of treatments (circumstances or interventions [14]) with the aim of giving a synthetic quantitative estimate of all the available studies [15]. As quantitative synthesis observes studies previously developed and analyzed by their authors, these synthesis studies also go by the name of meta-analysis, a term coined by Glass [3] in the field of psychology.

If all the experiments considered in a meta-analysis were equally precise and used exactly the same response variables, it would suffice just to average the results of each study to arrive at a final conclusion. In practice, though, not all studies are equally reliable. Therefore, a greater weight must be assigned to the studies from which more reliable information can be gained. The results are combined using a weighted mean [4], where a value derived from the variances or sample sizes of the experiments is used as a weight. Regarding the non-uniformity of the response variables, meta-analytical methods express their results using an *effect size* index, which is a non-scalar estimator of the relationship between treatments [4] and is applicable to any measure of difference between the results of two groups.

The weighted mean difference is the most commonly used method for the quantitative synthesis of continuous outcomes [15] (like results in SE, measuring productivity, effectiveness, efficiency, etc.). There are several other methods for discrete outcomes, such as odds ratio [15] or relative risk [15]. We do not address these methods here as discrete outcomes are not so frequently used in SE experiments. The weighted mean difference is widely used in medicine and psychology. However, there are other alternative methods for calculating the effect size, like the parametric response ratio [7], statistical vote counting [4] and non-parametric response ratio [7]. The four methods are briefly described in the following.

Weighted mean difference (WMD) uses the *individual effect* estimator that represents the rate of improvement of one treatment over another in each experiment. The individual effect size is measured as a quotient of the differences between the means and the combined standard deviation.

One of the drawbacks of the effect size estimated using WMD is that it is not straightforward to interpret. That is, it is not immediately clear how much better one treatment is than the other. Generally, a result equal to 0 is assumed to mean a null effect (the treatments behave equally), a result equal to 0.2 means a small effect size (one of the treatments is slightly better than the other), a result equal to 0.5 means a medium effect size (one of the treatments is clearly better than the other) and a result equal to 0.8 means a large effect size (one of the treatments is very much better than the other) [4]. As the effect size estimated using WMD is symmetrical with respect to the treatments, a positive value indicates that first treatment is preferable to the second, whereas a negative value means the opposite.

After estimating the effect size for each study, we can estimate the *global effect* that represents the overall rate of improvement of one treatment over another. This is calculated as the weighted mean of the effect estimators of the individual studies, where each study is weighted based on its inverse variance [4].

Parametric response ratio (PRR) involves estimating an effect index or ratio between two treatments by calculating the quotient of the two means. This quotient estimates the proportion of improvement between the two treatments [7].

The method is applied similarly to WMD. First, we have to estimate the ratio for each experiment and then, based on these ratios, we take a weighted mean of the individual ratios to estimate the global ratio, where each study is weighted based on its inverse variance [7].

Non-Parametric response ratio (NPRR) is similar to the parametric version, the only difference being the way in which the studies are weighted. The NPRR estimates variance based on the number of experimental subjects [16].

Very little information is required to apply the statistical vote counting (SVC) method. All the information experimenters need to know is whether or not there is a difference between the treatment means (what is called "vote") and the number of experimental subjects used in each study (used as a weighting factor of the "vote") [4]. Based on these data, a maximum likelihood estimation process is enacted for the purpose of determining the effect size (generally selected from a list ranging from -0.5 to 0.5). For more details on how to apply SVC, see [4].

Although all meta-analysis methods have the same objective of getting a synthetic estimate of the effect size between treatments, the different methods are not alike.

WMD, PRR, SVC, NPRR differ on two points: (1) the use of sample population parameters as opposed to parameter estimates, and (2) the number of primary studies that they can handle. Means, variances and sample sizes are required by WMD and PRR. NPRR and SVC do not require all the parameters and can therefore aggregate studies that report less information. The downside is that the less demanding methods need to estimate the sample parameters, leading to a loss of precision in the process.

Although this might appear to be a purely technical problem, it actually puts researchers using meta-analysis methods into a dilemma. Suppose that we have a set of x studies reporting means, variances and sample sizes; another set of y studies reporting means and sample sizes and a third set of z studies reporting the effect direction and sample sizes (the other possibilities are not meta-analyzable). Which is the best option? Should we

- aggregate x studies using WMD/PRR, or
- aggregate $x+y$ studies using NPRR, or
- aggregate $x+y+z$ studies using SVC?

For large values of x, y and z, the reliability and statistical power of the meta-analysis methods are similar, meaning that the above question is of no practical importance. In SE, on the contrary, the values of x, y and z tend to be small. Therefore the discrepancy between the theoretical and real distributions of WMD, PRR, SVC and NPRR (the statistical estimators) is likely to be sizeable in SE, and, if used without due care, they can lead to mistaken findings.

III. Related Work

There are two works that have addressed the application conditions of the meta-analysis methods. Lajeunesse [12] aims to determine which of the four methods is more advisable for ecology. He uses a Monte Carlo simulation process to analyze the type I and type II errors for each of the methods when applied to an increasing number (5-30) of experiments in contexts where there is no effect.

Lajeunesse's work shows that the reliability of WMD and SVC is poor for five experiments, whereas both versions of RR behave quite acceptably (or much more favorably). This changes as of 15 experiments, when the WMD method starts to perform better than RR. As regards SVC, its reliability starts to grow significantly as of 15 experiments (but is lower than for RR). The statistical power of SVC is still low even using a much greater number of experiments. From this research we can infer that it is advisable to use RR (parametric or non-parametric) when the aggregation contains few experiments, whereas it is better to use WMD when there are over 15 experiments. Unfortunately, this research does not analyze in detail how other factors, such as the number of subjects per experiment or the variances of the tested treatments, affect the methods.

Friedrich and colleagues [13] aim to determine whether PRR might be an alternative to WMD for meta-analysis in medicine. They analyze WMD and PRR using a similar methodology to Lajeunesse's. The key difference is that Friedrich and colleagues look at how type I and type II statistical error behaves for each method when there are variations not only in the number of experiments but also in the number of experimental subjects and the variance. Their work shows that the levels of reliability and statistical power of PRR and WMD are similar. Therefore, they conclude that the use of PRR is recommendable in medicine. Unfortunately, this research does not study the behavior of the non-parametric methods (SVC and NPRR) since this type of methods are not generally used in medicine.

In summary, the two studies conducted to date to compare the performance of quantitative synthesis methods suggest that the precision of both the parametric and non-parametric versions of RR and WMD are comparable. Both of these studies were run in other disciplines and do not account for the needs of SE. On this ground, we have run a study tailored to the experimental context of SE, where:

- Non-parametric methods are considered due to reporting shortcomings.
- The number of subjects per experiment is considered due to the small size of the experiments.
- The variance of the tested treatments is considered due to the large disparity of variances across experiments, also caused by the small number of experimental subjects.

Our study will help SE experimenters to identify which aggregation method should be applied depending on the features of the studies to be combined.

IV. Research Methodology

Our research aims to study which meta-analysis methods are useful in SE and when to apply each method depending on the number of studies and the available information. We ran a simulation to exhaustively evaluate the behavior of WMD, PRR, SVC and NPRR.

We jointly analyze the number of experiments, the number of subjects per experiment, the effect sizes and the variance levels in the context of SE experiments today.

As in [12] and [13], we use the Monte Carlo technique for the simulation process. SE researchers have used the Monte Carlo method to simulate fault detection in formal protocol testing methods [17], code fault detection by reviewers with different profiles [18] and associated models [19], software fault inspection sampling [20], etc.

We have observed the typical values of the experiments covered by earlier systematic reviews in SE [21, 22, 23, 24] to define the population values to be used to output the sample values for meta-analysis simulation:

- For the number of subjects per experiment, we consider the range of 4 to 20 subjects per experimental group. It is hard to consider an experiment with fewer than four subjects per group. In SE there are many examples of experiments with from 4 to 20 subjects per group. The number of experiments is varied two by two in each case, as differences of one subject or experiment are unlikely to produce significant results.
- The number of experiments to be aggregated in each meta-analysis will range from 2 to 10, as these are typical values of the systematic reviews in SE.
- The population effect sizes (δ) are those defined in [13] (small (0.2), medium (0.5) and large (0.8)) plus the very

large effect size (1.2), as about 30% of the experiments published in SE have an effect size greater than 1 [25].

Regarding the simulation process:

- The population mean of the secondary treatment (μ^C) is set at 100 for the purposes of calculation, and, as in [13], standard deviation (σ) is set at the following percentages of the mean of the respective treatment: 10% (low variance), 40% (medium variance) and 70% (high variance).
- The population mean of the primary treatment will be estimated as defined in [13], that is, $\mu^E = 100 + \delta * \sigma$, and the population ratio used to validate the results generated by parametric and non-parametric RR will be estimated by RR = μ^E / μ^c.
- We apply the inverse of the response variable output used in [12] and [13] (type I and II error) (1-α) for reliability and (1-β) for statistical power. The inverse variables appear to better suit our purpose as they ascertain how often a meta-analysis method is wrong by determining whether (type I error) or not (type II error) there is a significant difference between two treatments.
- Following the recommendations in [12] on how to combine the values of the variables (effect size, number of experiments and number of subjects per experiment), we run 1000 simulations and then calculate the values of the response variables.

V. Simulation

Tables V to X in Annex A show the results of the simulation. The reliability-related tables indicate the percentage of times that the estimated confidence interval (α = 0.05) contained the population effect size value, whereas the statistical power-related tables indicate the percentage of times that the above confidence interval did not contain the value 0 for the WMD and SVC methods and the value 1 for the parametric and non-parametric RR methods. To make the tables clearer, we highlighted the cells where the estimated percentages were above the minimum set value (1 – α) * 100 = 95% for reliability and (1 – β) * 100 = 80% (which is the commonly recommended value [26]) for statistical power. The shaded cells reveal a clear and consistent pattern for both reliability and power. Let us discuss the two separately.

A. *Reliability*

For clarity's sake, Table I simplifies the detailed results reported in Tables V, VII and IX in three ways. First, we grouped the parameter values (*effect size, number of experiments* and *number of subjects per experiment*) depending on whether or not they passed the reliability criteria set beforehand (reliability ≥95%). We replaced the percentages by "R" or "UR" to indicate that the method behaved reliably or unreliably, respectively. Secondly, we found that the key factor affecting method reliability is the total number of subjects, not the number of experiments. Then we grouped the *number of experiments* and *number of subjects by treatment* together in a column called "number of subjects/experiment". Finally, as variance did not appear to affect the reliability of the methods, it was omitted.

TABLE I. Comparison of reliability

Effect	Subjects/ experiment	Total subjects	WMD	PRR	SVC	NPRR
Small	< 8	Irrelevant	R	UR	UR	R
	≥8	Irrelevant	R	R	UR	R
Medium	< 8	Irrelevant	R	UR	UR	R
	≥8	< 80	R	R	UR	R
	≥8	≥ 80	R	R	R	R
Large & very large	< 8	Irrelevant	UR	UR	UR	R
	≥8	Irrelevant	UR	R	UR	R

As Table I shows, NPRR is reliable throughout the entire test and does not appear to be affected either by the number of experiments or the number of subjects. On the other hand, WMD proves to be reliable in contexts where the effect size is small and medium, although it performs worse than expected where effect sizes are large and very large. NPRR does not appear to be affected by the number of subjects per experiment or the total number of subjects. PRR is insensitive to effect size, whereas it is more sensitive to the number of experimental subjects. When the number of subjects is greater than or equal to 8, PRR is reliable, irrespective of any other parameter. Therefore, PRR is quite reliable in many situations. Finally, SVC proves to be reliable only in settings where the effect size is medium and the number of subjects per experiment and the total number of subjects are really high (≥8 and ≥80, respectively).

B. *Statistical Power*

Table II shows a similar simplification as used in section V-A, "P" means that the method is statistically powerful, whereas "UP" means that it is not. As we found that statistical power is insensitive to the number of subjects per experiment, no distinction has to be made between the number of subjects per experiment and the total number of subjects to analyze power.

TABLE II. Comparison of the statistical power

Variance	Effect	Total subjects	WMD	PRR	SVC	NPRR
High	Small	< 160	UP	UP	UP	UP
		≥ 160	UP	UP	P	UP
	Medium	< 40	UP	UP	UP	UP
		≥ 40 & < 80	UP	UP	P	UP
		≥ 80 & < 120	P	UP	P	UP
		≥ 120 & < 160	P	P	P	UP
		≥ 160	P	P	P	P
	Large	< 20	UP	UP	UP	UP
		≥ 20 & < 48	UP	UP	P	UP
		≥ 48 & < 80	P	UP	P	UP
		≥ 80	P	P	P	P
	Very large	< 32	UP	UP	P	UP
		≥ 32 & < 48	P	P	P	UP
		≥ 48	P	P	P	P
Medium	Small	< 160	UP	UP	UP	UP
		≥ 160	UP	UP	P	UP
	Medium	< 48	UP	UP	UP	UP
		≥ 48 & < 112	UP	UP	P	UP
		≥ 112 & < 140	P	UP	P	UP
		≥ 140	P	P	P	UP
	Large	< 20	UP	UP	UP	UP
		≥ 20 & < 48	UP	UP	P	UP
		≥ 48 & < 100	P	P	P	UP
		≥ 100	P	P	P	P
	Very large	< 32	UP	UP	P	UP
		≥ 32 & < 80	P	P	P	UP
		≥ 80	P	P	P	P
Low	Small	< 160	UP	UP	UP	UP
		≥ 160	UP	UP	P	UP
	Medium	< 32	UP	UP	UP	UP
		≥ 32 & < 48	UP	UP	P	UP
		≥ 48 & < 112	UP	P	P	UP
		≥ 112	P	P	P	UP
	Large	< 16	UP	UP	UP	UP

Variance	Effect	Total subjects	WMD	PRR	SVC	NPRR
		≥ 16 & < 48	UP	UP	P	UP
		≥ 48 & < 64	UP	P	P	UP
		≥ 64	P	P	P	UP
	Very large	< 16	UP	UP	P	UP
		≥ 16 & < 32	UP	P	P	UP
		≥ 32	P	P	P	UP

As Table II shows, the power of all the methods is satisfactory as long as the total number of subjects is relatively large. The exact value depends on the method, variance and effect size. The greater the variance is, the more subjects are required to achieve a satisfactory statistical power, and vice versa. Effect size also has a considerable influence, and the number of subjects required drops steeply as the effect size increases. This could not be otherwise, as, statistical power is inversely proportional to the total α and directly proportional to the effect size and sample size. The results, therefore, obey all statistical inference requirements.

Note that NPRR, in particular, is not very powerful at all. This is all the more remarkable pitched against its high reliability. The requirements of this method are the most demanding of all the tested methods. For example, in a context of high variance and medium effects (not at all uncommon in SE), over 160 subjects per treatment would be required to assure a statistical power of 80% for NPRR. On the other hand, PRR reaches that power with a total of no more than 120 subjects per treatment. WMD and SVC are less demanding (80 and 40 subjects per treatment, respectively).

WMD and PRR perform quite similarly. WMD is less demanding than PRR in the case of medium/high variances (requiring between 1/2 and 1/3 of the number of subjects demanded by PRR), although things are the other way round in the case of low variances. Finally, as regards power, SVC is by far the best of the tested methods, requiring a good many fewer subjects under all circumstances.

C. Corroboration

The theoretical definitions suggest that the total number of subjects involved in the process is what affects the reliability and statistical power of the aggregation methods [27]. This groundwork points to a simulation process based on studies of equal size, which is what Friedrich [13] and Lajeunesse [12] did in their research. But such conditions are very hard to reproduce in the real world (real meta-analysis research involves studies of different sizes). For this reason, we decided to run a second simulation involving aggregation processes using experiments with disparate numbers of subjects. The aim of this simulation was not to generate new knowledge but to validate the evidence generated in the original process. The second simulation accounted for the population values (means, variances and effect sizes) defined earlier but combined studies of three different sizes: small, medium and large.

The number of subjects to be assigned to each experiment was also decided according to the rules of the first simulation, assigning four subjects per treatment to the small experiments (this is the minimum size defined in the first simulation), 20 subjects per treatment to the large experiments (this is the maximum size defined in the first simulation) and 14 subjects per treatment to the medium-sized experiments (this being the nearest value to the average of the large and small experiments ((20 + 4)/2 = 12) used in the first simulation).

As we defined three experiment sizes, we decided to run combinations of three experiments, combining experiments of all sizes (except experiments of equal sizes because we had already evaluated this case), six experiments with twice the number of studies of each size as used in the first simulation and nine experiments with three times the number of studies of each size as used in the first simulation. In this new simulation, we corroborated the results about method reliability and statistical power (shown in Tables I and II). The tables with the results of this second simulation are available at http://www.grise.upm.es/sites/extras/5.

VI. Results

Our results are generally consistent with earlier studies [12, 13]. However, the diversity of contexts tested in our simulation highlights certain issues that have gone unnoticed until now and are important for establishing the relative reliability and power of the different meta-analysis methods.

First NPRR is not very powerful when the total number of subjects in the meta-analysis is small. Contrariwise, Lajeunesse [12] suggests that NPRR is virtually type II error free with as few as five experiments. This is true, but only under the conditions simulated by Lajeunesse: a very large effect size (around 1.0), a total number of experimental subjects per aggregation of around 75 and a medium variance (around 30%). In our simulation, the power using NPRR is also around 90% under Lajeunesse's conditions (see Table VI in Annex A). Now this high statistical power is an exception in contexts with small sample sizes (very common in SE).

Secondly, our study suggests that the outlook for WMD is not as gloomy as presented by Lajeunesse [12]. The type I and type II errors reported in [12] for WMD were as high as 71% and 88%, respectively, for five experiments including 75 subjects per aggregation, medium variances and very high effect sizes. Our simulation returns much better data for this technique. Our data are consistent with findings by Friedrich and colleagues [13], suggesting that WMD is reliable and powerful not only as of a total of 75 subjects, but also with quite a lot fewer subjects depending on the contextual conditions (effect sizes and variances).

Thirdly, the reliability of WMD in contexts with a high or very high effect size (d ≥ 1.2) is remarkably low. It is striking that as the number of experiments and subjects increase, there are losses rather than gains in WMD reliability of close to 50%. This appears to contradict earlier research by Hedges [28], partially reported in [4]. As regards the theory of large sample sizes, our results appear to defy common sense. However, this result is partly supported by Harwell's findings. Harwell [29] observed that the statistical power of WMD tends to be lower when it combines small studies even if the total number of subjects is high.

Fourthly, contrary to the claims in [12], SVC is only reliable in contexts with medium effect sizes (d = 0.5), albeit apparently for the same reason as explains the low reliability level for WMD: the possible heterogeneity of the primary studies. Although lower than in our results, the power of SVC shows a similar upward trend as reported by [12] as the number of experiments and experimental subjects increase.

Finally, our study fairly reliably replicates the results of [13] regarding PRR, confirming that its reliability and statistical power are similar to WMD's. This applies even in contexts with very high population effects, where, according to our findings, WMD is not as reliable. Friedrich and colleagues [13] did not analyze this point because the maximum effect size he tested was 0.8. Again there are no other results to corroborate our data [13]. Even so, PRR should logically be more robust than WMD in the presence of heterogeneity, as PRR uses a logarithmic transformation. This transformation is currently used as a means of controlling the heterogeneity of variances [7].

VII. Selecting A Meta-Analysis Method

Based on Tables I and II, we can say that:

- In contexts where effect sizes are low (d = 0.2), all the methods face the problem of powerlessness. This makes sense since it is hard to detect significant differences if they are small and the total number of experimental subjects is not very large;
- In contexts where effect sizes are medium (d = 0.5), the parametric methods (WMD and PRR) perform well, meaning they are reliable and powerful;
- In contexts where effect sizes are high (d = 0.8), WMD results are reversed, and it is powerful but not reliable, whereas the ratio-based methods (PRR and NPRR) remain reliable;
- In contexts where effect sizes are very high (d = 1.2), the trend for the high effects is unchanged, where the ratio-based methods (PRR and NPRR) are the best option.

We suggest that SE experimenters should use the following procedure to select the best meta-analysis method for a particular aggregation:

1. Create four different groups of experiments, one per meta-analysis technique. Notice that the studies that can be aggregated using WMD or PRR (which are the most restrictive methods) can also be aggregated using NPRR or SVC. Therefore, groups overlap and contain a different number of experiments.
2. Analyze the estimated effect sizes to establish whether the context of the effects is low (d = 0.2), medium (d = 0.5) or high (d > 0.8).
3. Analyze the global variance level with respect to the global mean of the first group of studies to be able to establish whether the variance is low (10%), medium (40%) or high (70%). If variances are not available, the variance can be assumed to be medium (40%), as this is the most frequent case in SE [25].
4. Calculate the total number of experiments and experimental subjects in each group.
5. Establish the levels of reliability and power of each group based on the information listed in Tables I and 2 or, alternatively, based on the analysis of the tables in Annex A.
6. Use the technique with the best reliability and power values to perform meta-analysis.

For example, suppose a researcher aims to aggregate the experiments presented in Table III.

TABLE III. Example of an Aggregation Situation

Exp.	Exp. Mean	Control Mean	Exp. Subjects	Control Subjects	Exp. Std deviation	Control Std deviation	Mean dif.
1	90	75	16	16	28	30	--
2	115	90	20	20	40	35	--
3	100	75	10	10	42	33	--
4	100	100	8	8	39	40	--
5	130	100	10	10	----	----	--
6	100	90	10	10	----	----	--
7	95	100	12	12	----	----	--
8	95	90	8	8	----	----	--
9	----	----	10	10	----	----	$Y^{E>}Y^{c}$
10	----	----	8	8	----	----	$Y^{E>}Y^{c}$

Following the above procedure, he or she would create four groups (see Table IV). The first and second groups are alike, and they contain the experiments reporting all statistical parameters (experiments 1, 2, 3 & 4); the third group contains the experiments that report averages and sample sizes but not variances (experiments 5, 6, 7 & 8); finally, the experiments in the fourth group report sample sizes and the direction of the effect size (experiments 9 & 10).

TABLE IV. GROUPS AND RELIABILITY/POWER CALCULATIONS

	WMD	PRR	NPRR	SVC
Effect size	Medium (0.504)	Medium (1.218)	Medium (1.159)	Medium (0.4)
Number of exp.	4	4	8	10
Number of subjects	54	54	94	112
Variance	Medium (approx. 40% of the mean)	Medium (approx. 40% of the mean)	Medium (approx. 40% of the mean)	Medium (approx. 40% of the mean)
Reliability	R	R	R	R
Statistical power	UP	UP	UP	P

Using the information in Tables I and II, it is immediately clear that (see Table IV):

- SVC is reliable (R) and powerful (P);
- WMD, PRR and NPRR are reliable (R) but not powerful (UP).

Therefore, we can say that the best method for conducting the meta-analysis of the above studies is SVC, as it has the lowest type I and type II error rate.

The information listed in Tables I and II is insufficient to discriminate whether or not WMD is preferable to PRR or NPRR. The detailed data shown in the Annex A provide a finer granularity. The findings according to Tables IX and X are:

- For four experiments (studies 1, 2, 3 and 4 in Table III) with a total of 54 experimental subjects per aggregation (approximately 16 subjects per study (54 / 4 = 13,5)) in a medium variance and medium effect setting, WMD has a reliability of 100% and a power of 57.1%, whereas PRR has a reliability of 98.6% and a power of 53.3%.
- For eight experiments (studies 1 to 8 in Table VII) with a total of 94 experimental subjects per aggregation (approximately 12 subjects per study (94 / 8 = 12.2)) in a medium variance setting, NPRR has a reliability of 100% and a power of 0%.

Therefore, WMD and PRR could be considered equivalent for aggregating the four experiments when we have access to all the information, whereas NPRR is the least suited method, as it has no power whatsoever (meaning its application never returns any significant effects between the treatments analyzed in the meta-analysis). In this example, then, it is better to aggregate studies 1, 2, 3 and 4 (covering a total of 54 subjects) using WMD or PRR than to aggregate studies 1 to 8 with 94 subjects using NPRR.

VIII. SCOPE AND LIMITATIONS

Note that our study considers only meta-analysis methods used with continuous variables, as they are the most widespread in SE. We have considered only fixed-effects models, that is, the statistical methods used in contexts of statistical homogeneity. There are two reasons for this decision. On the one hand, when the meta-analysis does not cover many studies (as is common in SE), statistical homogeneity is hard to detect [4]. On the other hand, when heterogeneity is detected in the context of few experiments, this is bound to be very pronounced, and their combination through meta-analysis is ill advised [27]. Therefore random-effects models are not suitable for aggregating heterogeneous SE experiments. This explains the use of just the fixed-effects models here.

IX. CONCLUSIONS

We have determined the values of reliability and statistical power for the WMD, PRR, NPRR and SVC meta-analysis methods under the usual conditions in SE (few experiments and few experimental subjects) across a range of effect sizes and variance levels. Under these conditions, as the large sample sizes condition does not hold, the analytical function for reliability and statistical power are not necessarily applicable, and they have to be studied by means of statistical simulation.

The values of reliability and statistical power have been tabulated with the aim of helping researchers to identify the reliability of WMD, PRR, NPRR, SVC in the context of their aggregation work. These tables enable the selection of the best meta-analysis method in each case. The tables are also useful for identifying how many experiments (or, to be more precise, experimental subjects) are required to achieve what are usually considered to be adequate levels of reliability and statistical power ($\alpha = 0.05$; $\beta = 0.2$).

The simulations that we have conducted are a means of answering questions regarding which meta-analysis method to apply in a specific aggregation situation. The decision is made depending on how reliable and statistically powerful the methods are for the set of experiments to be aggregated.

ACKNOWLEDGMENT

This research has been partially funded by Spanish Ministry of Science and Innovation grants TIN2008-00555 and HD2008-00046.

REFERENCES

[1] Goodman C.; Literature searching and evidence interpretation for assessing Health Care Practices; SBU; Stockholm, 1996.

[2] Chalmers I., Hedges L., Cooper H.; A brief history of research synthesis; *Eval Health Prof*, vol. 25, no. 1, pp. 12–37, 2002.

[3] Glass, G; Primary, secondary, and meta-analysis of research; *Educational Researcher*, vol. 5, pp. 3-8, 1976.

[4] Hedges, L.; Olkin, I.; 1985; Statistical methods for meta-analysis. Academic Press.

[5] Sjoberg, D.I.K.; Hannay, J.E.; Hansen, O.; Kampenes, V.B.; Karahasanovic, A.; Liborg, N.-K.; Rekdal, A.C.; A survey of controlled experiments in software Engineering; *IEEE Transactions on Software Engineering*, vol. 31, no. 9, pp. 733-753, 2005.

[6] Shercliffe, R.; Stahl, W.; Tuttle, M.; The use of meta-analysis in psychology; *Theory & Psychology*, vol. 19, no. 3, pp. 413-430, 2009.

[7] Gurevitch, J.; Hedges, L.; Meta-analysis: Combining results of independent experiments; Design and Analysis of Ecological Experiments (eds S.M. Scheiner and J. Gurevitch), pp. 347–369; Osford: Oxford University Press, 2001.

[8] Kopper, K.; McKenzie, D.; Peterson, D; The evaluation of meta-analysis techniques for quantifying prescribed fire effects on fuel loadings; United States Department of Agriculture; Forest Service; Pacific Northwest Research Station; Research Paper PNW-RP-582, 2009.

[9] Curtis J.; Kyong-Jee K.; Tingting Z.; Future directions of blended learning in higher education and workplace learning settings; In Bonk, C. J. & Graham, C. R. (Eds.). *Handbook of blended learning: Global Perspectives, local designs*. San Francisco, CA: Pfeiffer Publishing, 2004.

[10] Pickard, L.; Kitchenham, B.; Jones, P; Combining empirical results in Software Engineering; *Information and Software Technology*, vol. 40; pp. 811-821, 1998.

[11] Mohagheshi, P.; Conrradi, R.; Vote-Counting for combining quantitative evidence from empirical studies – An Example. In J.C. Maldonado and C. Wohlin; *Proceedings of the 5th ACM-IEEE Internet and Symposium an Empirical Software Engineering* (ISESE'06); Rio de Janeiro; IEEE press; pp24-26, 2006.

[12] Lajeunesse, M.; Forbes, M.; Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques; *Ecology Letters*, vol. 6, pp. 448-454, 2003.

[13] Friedrich, J.; Adhikari, N.; Beyene, J; The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study; BMC Medical Research Methodology, vol. 8, no. 32, 2008.

[14] Hunt, M.; *How Science takes stock: the story of meta-analysis*; New York: Russell Sage Foundation, 1997.

[15] Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available from www.cochrane-handbook.org.

[16] Miguez, E.; Bollero, G; Review of corn yield response under winter cover cropping systems using meta-analytic methods; *Crop Science Society of America*, vol. 45, no. 6, pp. 2318-2329, 2005.

[17] Sidhu, D.; Leung, T.; Formal methods for protocol testing: A detail study; *IEEE Transactions on Software Engineering*, vol. 15, no. 4, pp. 413-426, 1989.

[18] Vander Wiel, S.; Votta, L.; Assessing software design using capture-recapture methods; *IEEE Transactions on Software Engineering*, vol. 19, no. 11, pp. 1045-1054, 1993.

[19] El Emam, K.; Laitenberger, O.; Evaluating capture-recapture models with two inspectors; *IEEE Transactions on Software Engineering*, vol. 27, no. 9, pp. 851-864, 2001.

[20] Thelin, T.; Petersson, H.; Runeson, P.; Wohlin, C.; Applying sampling to improve software inspections; *Journal of Systems and Software*, vol. 73, pp. 257-269, 2004.

[21] Juristo N.; Moreno A.; Vegas S.; Towards building a solid empirical body of knowledge in testing techniques; *ACM SIGSOFT Software Engineering Notes*, vol. 29, no. 5, pp. 1-4, 2004.

[22] Davis, A.; Dieste O.; Hickey, A.; Juristo, N.; Moreno, A.; Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review; *14th IEEE International Requirements Engineering Conference* (RE'06), pp. 179-188, 2006.

[23] Dyba, T.; Arisholm, E.; Sjoberg, D.; Hannay J.; Shull, F.; Are two heads better than one? On the effectiveness of pair programming. IEEE Software, vol. 24, no. 6, pp. 12-15, 2007.

[24] Ciolkowski, M; What do we know about perspective-based reading? An approach for quantitative aggregation in software engineering; *3rd International Symposium on Empirical Software Engineering and Measurement* (ESEM'09), pp. 133-144, 2009.

[25] Kampenes, V.; Dyba, T; Hannay J.; Sjøberg, D.; A systematic review of effect size in software engineering experiments; *Information and Software Technology*, vol. 49, pp. 1073–1086, 2007.

[26] Cohen, J; Statistical power analysis for the behavioral sciences, Hillsdale, NJ: Lawrence Erlbaum Associates, 1977, 2nd Ed.

[27] Borenstein, M.; Hedges, L; Rothstein, H.; Meta-Analysis Fixed Effect vs. random effect; www.meta-analysis.com, 2007.

[28] Hedges, L.; Fitting categorical model to effect size from a series of experiments; *Journal of educational statistics*, vol.7, pp. 119-137, 1982.

[29] Harwell, M; An empirical study of the Hedges (1982) homogeneity test; *Psychological methods*, vol. 2, no. 2, pp. 219-231, 1995.

ANNEX A. SIMULATION RESULTS

TABLE V. COMPARISON OF AGGREGATION METHOD RELIABILITY, $\alpha = 0.05$

			WMD					PRR					SVC					NPRR				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Reliability-Low Variance	EF: 0.2	4	98.4	100	100	99	100	93.9	89.7	85.5	93.3	93.5	0.9	0.7	0.3	0.3	0.2	100	100	100	100	100
		8	98.1	100	100	100	100	98.1	100	100	100	100	1.2	0.5	0.2	0.3	0.2	100	100	100	100	100
		10	97.6	100	100	100	100	96.6	99.3	100	100	100	0.7	0.4	0.3	0.3	0	100	100	100	100	100
		14	97.6	100	100	100	100	97.6	97.6	100	100	100	1.2	0.5	0.3	0.3	0.3	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	1.1	0.5	0.3	0.3	0.2	100	100	100	100	100
	EF: 0.5	4	97.3	100	100	99.4	96.3	91.6	90.2	89.5	92.3	93	57	75.6	85.6	92.6	79	100	100	100	100	100
		8	96.3	100	98.9	98.1	98.1	97.5	100	100	100	100	53.7	80.8	83.7	92.7	97.2	100	100	100	100	100
		10	92.1	97.9	100	99	93.1	96.5	100	100	100	100	68.9	86.5	89.7	93.9	94	100	100	100	100	100
		14	97.4	99.4	98.5	98.4	98.2	96.1	98.3	100	100	100	83	92.1	99	100	100	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	92.9	98.9	99.1	98.6	100	100	100	100	100	100
	EF: 0.8	4	95.6	99.1	99.1	97.1	94.1	93.6	90.4	90.8	94.2	95.2	0.6	0.4	0.2	0.2	0.2	100	100	100	100	100
		8	93.7	95.6	96	92.8	93.4	96.3	100	100	100	100	0.5	0.4	0.2	0.2	0.2	100	100	100	100	100
		10	87.6	92.1	95.9	92.2	83.5	96	100	100	100	100	0.6	0.4	0.2	0.2	0.2	100	100	100	100	100
		14	98.4	91.8	94.2	90.5	90.3	96.5	98.3	100	100	100	0.6	0.4	0.2	0.2	0.2	100	100	100	100	100
		20	100	100	96.4	92.5	100	100	100	100	100	100	0.6	0.4	0.2	0.2	0.2	100	100	100	100	100
	EF: 1.2	4	95.4	96.3	92.4	88.1	82.2	91.9	91.6	89.6	94.3	95	0.2	0.2	0	0	0	100	100	100	100	100
		8	90.5	91.4	83.8	73.1	80.7	97.3	100	100	100	100	0.2	0.2	0	0	0	100	100	100	100	100
		10	81.3	81.1	85.7	79.5	62.8	96.1	100	100	100	100	0.2	0.2	0	0	0	100	100	100	100	100
		14	94.7	79.1	81.6	68.3	55	96.3	97.6	100	100	100	0.2	0.2	0	0	0	100	100	100	100	100
		20	98.2	91.6	72.5	63.1	51.5	100	100	100	100	100	0.2	0	0	0	0	100	100	100	100	100

TABLE VI. COMPARISON OF AGGREGATION METHOD STATISTICAL POWER, $\beta = 0.2$

			WMD					PRR					SVC					NPRR				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Power- Low Variance	EF: 0.2	4	1.0	2.0	1.1	1.2	2.7	3.9	5.1	10.3	9.1	11.8	40.9	51.6	57.8	65.6	70.0	0	0	0	0	0
		8	2.8	5.2	1.2	3.1	5.9	2.8	7.7	2.6	4.2	7.7	37.5	59.7	73.9	77.3	87.7	0	0	0	0	0
		10	6.9	7.1	1.9	8.5	11.9	8.9	7.9	4.9	10.1	16.5	47.3	62.7	64.9	75.7	81.2	0	0	0	0	0
		14	8.9	7.7	9.7	1.8	13.1	8.9	11.7	15.1	5.9	18.3	38.6	55.5	74.0	77.9	87.9	0	0	0	0	0
		20	4.6	1.7	0.0	19.1	0.0	4.6	1.7	0.0	22.2	5.9	54.4	77.0	81.5	83.8	100.0	0	0	0	0	0
	EF: 0.5	4	5.0	8.5	16.5	19.8	25.0	9.7	24.1	33.6	40.8	47.9	56.0	75.1	85.1	92.6	93.7	0	0	0	0	0
		8	5.6	25.8	45.2	61.1	82.6	15.6	34.4	56.7	71.2	85.7	53.1	80.3	83.2	92.7	97.2	0	0	0	0	0
		10	28.2	40.8	52.5	77.9	84.8	32.2	44.3	66.0	84.7	90.8	68.4	86.0	89.3	93.9	94.0	0	0	0	0	0
		14	31.5	57.6	83.6	98.4	100.0	36.8	60.1	92.7	98.4	100.0	82.3	91.6	98.5	100.0	100.0	0	0	0	0	0
		20	41.9	100.0	100.0	100.0	100.0	47.7	100.0	100.0	100.0	100.0	92.2	98.4	99.1	98.6	100.0	0	0	0	0	0
	EF: 0.8	4	11.6	32.5	44.6	70.0	79.8	22.4	42.0	65.8	80.0	90.3	78.1	94.6	99.0	100.0	100.0	0	0	0	0	0
		8	33.4	70.9	94.4	98.9	100.0	43.1	78.5	95.2	100.0	100.0	78.6	93.9	94.5	100.0	100.0	0	0	0	0	0
		10	53.4	81.2	100.0	100.0	100.0	54.4	88.2	100.0	100.0	100.0	85.0	97.6	99.8	100.0	100.0	0	0	0	0	0
		14	66.3	98.0	100.0	100.0	100.0	67.8	98.0	100.0	100.0	100.0	99.8	99.8	99.8	100.0	100.0	0	0	0	0	0
		20	97.7	100.0	100.0	100.0	100.0	97.7	100.0	100.0	100.0	100.0	98.8	99.8	100.0	100.0	100.0	0	0	0	0	0
	EF: 1.2:	4	30.9	73.3	95.1	98.8	98.1	46.3	88.5	97.5	98.8	99.2	94.0	99.8	99.8	100.0	100.0	0	0	0	0	0
		8	73.5	100.0	100.0	100.0	100.0	77.5	100.0	100.0	100.0	100.0	94.1	99.2	99.8	100.0	100.0	0	0	0	0	0
		10	78.8	100.0	100.0	100.0	100.0	81.3	100.0	100.0	100.0	100.0	99.8	99.8	99.8	100.0	100.0	0	0	0	0	0
		14	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.8	99.8	99.8	100.0	100.0	0	0	0	0	0
		20	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.8	99.8	100.0	100.0	100.0	0	0	0	0	0

TABLE VII. COMPARISON OF AGGREGATION METHOD RELIABILITY, $\alpha = 0.05$

			WMD					PRR					SVC					NPRR				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Reliability – Medium Variance	EF: 0,2	4	97.8	100	100	99.2	100	96.4	92	90	93.2	92	0.8	0.6	0.4	0.2	0.4	100	100	100	100	100
		8	97	100	100	100	100	97	100	100	100	100	1	0.8	0.4	0.2	0.2	100	100	100	100	100
		10	97.4	100	100	100	100	96	100	100	100	100	1	0.4	0.4	0.4	0.2	100	100	100	100	100
		14	96	100	100	100	100	96	98.6	100	100	100	1	1	0.2	0.2	0.2	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	1	0.4	0.2	0.2	0.2	100	100	100	100	100
	EF: 0,5	4	97	100	100	99.1	96.6	90.7	92.3	90	93.6	94.3	56.7	76	85.4	91	80.6	100	100	100	100	100
		8	96.3	100	98.6	96.8	97.8	97.2	100	100	100	100	52.9	81	85.2	91.7	97	100	100	100	100	100
		10	93.3	97.7	100	98.9	91.8	95.7	100	100	100	100	68.1	85.9	89.9	94.6	94.4	100	100	100	100	100
		14	97.6	99.5	98.7	99.1	98	96.2	98.8	100	100	100	82.4	90	99	100	100	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	92.5	97.8	98.5	99	100	100	100	100	100	100
	EF: 0,8	4	96.7	98.8	98.5	95.5	95	93.5	90.6	93.3	92.2	90.7	0.2	0.2	0.1	0.1	0.1	100	100	100	100	100
		8	94.7	95.3	94.6	90.7	93.9	97.5	100	98.7	98.7	100	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
		10	85.7	91.9	95.4	95	81.7	95.6	99.4	100	100	100	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
		14	96.9	92.2	95.6	89.7	88.9	98.4	100	99.3	100	100	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
		20	100	100	94.8	89	100	100	100	100	100	100	0.3	0.2	0.1	0.1	0	100	100	100	100	100
	EF: 1,2	4	96.7	94.1	92.6	87.9	81.6	93.7	89.9	92.8	92.7	90.8	0.1	0.1	0	0	0	100	100	100	100	100
		8	92	91.7	81.6	76.8	79.7	97.5	100	98.8	99.2	100	0.1	0.1	0	0	0	100	100	100	100	100
		10	81.5	81.5	84.9	78.8	59.8	96.8	99.4	100	100	98.6	0.1	0.1	0	0	0	100	100	100	100	100
		14	94.8	79.2	80.3	60.9	55.4	99.4	100	99.8	100	100	0.1	0	0	0	0	100	100	100	100	100
		20	97.2	91.9	67.8	58.6	47.9	100	100	100	100	100	0.1	0	0	0	0	100	100	100	100	100

TABLE VIII. COMPARISON OF AGGREGATION METHOD STATISTICAL POWER, $\beta = 0.2$

			WMD					PRR					SVC					NPRR				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Power – Medium Variance	EF: 0.2	4	0.4	1.6	1.8	1.2	1.6	0.6	3.6	5	4.6	8	37.8	51.4	54.4	68.2	68.4	0	0	0	0	0
		8	2.6	5.2	0.6	2.4	5.4	2.6	4.4	0.6	1.8	2.6	37.2	55.6	70.2	77.2	87.6	0	0	0	0	0
		10	7.8	9.2	3	7	11.6	7.8	6.8	1.8	5.6	9.4	49.8	59.8	65.2	78.8	80.6	0	0	0	0	0
		14	8.6	6.8	8.6	1	16.8	8.4	6	6.4	1	10.6	42.4	55.2	76	78.8	88.4	0	0	0	0	0
		20	2.8	2	0	16.4	0	2.8	1.2	0	15.6	0	53.4	76.8	81	84.6	100	0	0	0	0	0
	EF: 0.5	4	5.3	10.3	16.4	17.7	25.3	6.1	18.4	22.3	23.4	34.8	56.1	75.7	85.1	91	93.2	0	0	0	0	0
		8	5.1	25.4	48.4	59.3	81.1	6.8	25	42.4	53	74.8	52.4	80.7	84.9	91.7	97	0	0	0	0	0
		10	27.9	39.5	55	78.3	84.7	28	37.7	50.2	79.8	76.1	67.8	85.6	89.6	94.6	94.4	0	0	0	0	0
		14	32.1	57.1	83.7	99.1	100	31	53.3	82.9	97.3	99	82.1	89.7	98.7	100	100	0	0	0	0	2
		20	44.9	100	100	100	100	42.5	97.6	100	100	100	92.2	97.5	98.5	99	100	0	0	0	0.9	0
	EF: 0.8	4	13	33.2	44.2	68.6	84.1	18.2	35.2	49.5	68.6	76.2	77.6	95.6	99.2	100	100	0	0	0	1.1	1.9
		8	37.1	70.7	93.6	98.3	100	36.1	72.8	91.9	98.7	100	79.1	94.9	94.3	100	100	0	0	0	8.6	19.6
		10	54	81.1	100	100	100	52.8	80.8	100	100	100	83	97.5	99.9	100	100	0	1.3	1.5	28.1	56.4
		14	66.8	98.7	100	100	100	59.2	93.4	100	100	100	99.9	99.9	99.9	100	100	0	4.8	31.8	59.2	94.1
		20	97.6	100	100	100	100	97.6	100	100	100	100	99.4	99.9	100	100	100	0	7.8	81.3	100	100
	EF: 1.2	4	33	75.8	96.6	99.3	97.8	39.7	79.2	97.9	99.3	97.8	94.3	99.9	99.9	100	100	0	0	0.8	5.5	20.3
		8	76.9	100	100	100	100	75.3	100	100	100	100	96.7	99.6	99.9	100	100	0	5.6	32.9	69.7	96.1
		10	79.5	100	100	100	100	80.5	100	100	100	100	99.9	99.9	99.9	100	100	0	20.1	68.7	95.6	100
		14	100	100	100	100	100	100	100	100	100	100	99.9	99.9	99.9	100	100	5	54.2	98.9	100	100
		20	100	100	100	100	100	100	100	100	100	100	99.9	99.9	100	100	100	16.6	100	100	100	100

TABLE IX. COMPARISON OF AGGREGATION METHOD RELIABILITY, $\alpha = 0.05$

			WMD					PRR					SVC					NPRR				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Reliability – High Variance	EF: 0.2	4	98.2	100	100	98.7	100	96.2	90.2	89.7	89.7	87.2	0.5	0.3	0.1	0.2	0.1	97.5	96.5	99.3	95.8	94.5
		8	95.8	100	100	100	100	99	100	100	100	100	0.6	0.4	0.1	0.1	0.1	97.4	100	98.8	100	100
		10	97.2	100	100	100	100	95.7	100	100	100	100	0.6	0.2	0.1	0.3	0.1	99.2	100	100	100	100
		14	96.9	100	100	100	100	98.5	100	100	100	100	0.7	0.3	0.1	0.1	0.1	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	0.4	0.2	0.1	0.2	0.1	100	100	100	100	100
	EF: 0.5	4	97.7	100	100	98.5	96.3	94.6	93.2	92.8	88.8	83	59	76	84.1	90.5	77.2	98.8	100	100	97.5	96.3
		8	95.2	100	99.6	95.8	97.5	96.6	99	98.3	95.9	100	54.3	77.8	87.8	89.2	97.2	97.8	100	100	100	100
		10	92.3	98.4	100	98.9	93.8	96.9	99.2	100	98.4	98.1	70.8	85.5	89.1	96	94.3	100	100	100	100	100
		14	97.6	98.8	97.7	98.6	98	100	98	98.9	100	100	80.3	89.7	98.6	100	100	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	91.9	96.8	98.8	99	100	100	100	100	100	100
	EF: 0.8	4	96.4	99.1	99.1	96.2	94	95.3	93.4	94.4	89.7	87.6	0.4	0.2	0.1	0.1	0.1	98.8	100	100	99.5	100
		8	94.2	95.7	96.9	92.5	94.2	95.5	94.2	97.2	96.3	98	0.3	0.2	0.1	0.1	0.1	99	100	100	100	100
		10	86.4	91.8	96.7	93.3	79.6	95.1	98.6	98.4	96.4	94.2	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
		14	97.7	92.5	95.6	91.6	88.9	100	96.3	99.3	100	99	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
		20	100	100	94.7	90.1	100	100	100	100	100	100	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
	EF: 1.2	4	94.9	96.4	93.6	89.7	80.4	95	90.6	92.1	89.3	84.7	0.1	0.1	0	0	0	98.8	100	100	99.4	100
		8	90.6	91.4	80.4	73.4	80.3	95.1	95.3	94.7	94	97	0.1	0.1	0	0	0	99	100	100	100	100
		10	81.8	83.2	85.3	81.4	60.1	94.6	95.6	97.2	96.6	87.1	0.1	0.1	0	0	0	100	100	100	100	100
		14	96.1	80	80.1	66.2	57.9	100	95.8	99.4	99	97.1	0.1	0.1	0	0	0	100	100	100	100	100
		20	98.1	91.7	67.2	60.7	52.3	100	100	100	100	100	0.1	0	0	0	0	100	100	100	100	100

TABLE X. COMPARISON OF THE AGGREGATION METHOD STATISTICAL POWER, $\beta = 0$

			WMD					PRR					SVC					NPRR				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Power – High Variance	EF: 0.2	4	1.1	1.8	0.7	0.7	1.4	2.2	3.5	4.2	5.1	2.2	41.1	50.2	59.6	64.9	69.5	1.6	0.5	0.5	1.7	1.4
		8	2.2	5.2	0.9	3.1	6	0	1.3	0	1	2.6	40.4	59.8	71.4	75.3	87.3	0.9	0	0	0	2.6
		10	10	10.1	1.9	8.3	13.7	2.7	6.6	0.9	6.5	5.8	52.1	64.4	63.8	74.8	80.9	0	1.2	0	1.1	3.2
		14	10.4	8.3	9.9	2.3	14.1	5.9	3.2	4.2	1	1.1	41.7	55.4	73.5	77.9	88.9	3.1	1.7	0	0	0
		20	3.1	1.9	0	17.2	0	2.5	0	0	8.9	0	58.4	76.3	83	84.6	100	0	0	0	0	0
	EF: 0.5:	4	5.4	8.4	14.6	20	22.9	2.4	7.3	10.5	11.6	15.8	58.8	75.9	84	90.5	90.9	1.2	1.6	5.1	9.4	11.4
		8	6	22.8	48	56	79.6	5.3	14.6	23.2	30.1	52.8	54.2	77.7	87.7	89.2	97.2	2.2	7.2	12.1	22.3	32.5
		10	28	42.5	50.7	80.5	86.8	21.1	30.1	31.9	56.6	69.4	70.7	85.4	89	96	94.3	1.4	14.2	16.5	34.7	57.9
		14	30.2	54.1	83.4	98.6	100	22.8	34.1	68.4	86.2	96.3	80.1	89.6	98.5	100	100	7.7	19.1	42.2	56.8	83.5
		20	42.5	100	100	100	100	33.7	86.2	100	100	100	91.7	96.7	98.8	99	100	6.2	37.8	80.1	95.9	100
	EF: 0.8	4	14.4	32.9	42.7	65.7	81.3	9.7	25.4	36.3	48.5	56.8	78.1	94.6	98.7	100	100	3.5	7.1	18.3	31.1	38.7
		8	36.5	69.8	95.7	98.4	100	22.4	59.2	78.5	93.5	100	79.4	93.8	94.7	100	100	5.8	25.7	53.4	75.1	92.4
		10	55.1	82.6	100	100	100	49.4	71	93.7	100	100	86.6	98.2	99.9	100	100	24.2	46.5	75.6	96	94.2
		14	66.3	98.6	100	100	100	54.7	88	99.3	100	100	99.9	99.9	99.9	100	100	24.2	62.7	95.6	100	100
		20	97.4	100	100	100	100	94.8	100	100	100	100	99.1	99.9	100	100	100	44.3	100	100	100	100
	EF: 1.2 -RR:	4	32	72.6	95.8	99	97.6	26.2	64.4	90	94.7	95.4	94	99.9	99.9	100	100	5	26.4	61.6	80.8	88
		8	74.6	100	100	100	100	55.6	96.6	100	100	100	95.1	99.3	99.9	100	100	17.8	76.6	96.8	100	100
		10	78.9	100	100	100	100	75	100	100	100	100	99.9	99.9	99.9	100	100	45.6	84.9	100	100	100
		14	100	100	100	100	100	100	100	100	100	100	99.9	99.9	99.9	100	100	67.5	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	99.9	99.9	100	100	100	98.1	100	100	100	100