Hidden Evidence Behind *Useless* Replications

Oscar Dieste
Universidad Politécnica de Madrid
odieste@fi.upm.es

Enrique Fernández
Universidad Nacional de La Plata
enriquefernandez@educ.ar

Ramón García
Universidad Nacional de Lanus
rgarcia@unla.edu.ar

Natalia Juristo
Universidad Politécnica de Madrid
natalia@fi.upm.es

ABSTRACT
Experiments that are run with few experimental subjects are often considered not to be very reliable and deemed, as a result, to be *useless* with a view to generating new knowledge. This belief is not, however, entirely correct. Today we have tools, such as meta-analysis, that we can use to aggregate small-scale experiments and output results that are equivalent to experiments run on large samples that are therefore reliable. The application of meta-analysis can overcome some of the obstacles that we come up against when running software engineering experiments (such as, for example, the practitioner availability problem).

Categories and Subject Descriptors
Empirical Software Engineering

General Terms
Experimentation

Keywords
Meta-analysis, statistical power, reliability, replications, sample size.

1. INTRODUCTION

Suppose that a hypothetical Dr. Smith is a university researcher working on testing techniques. Recently, Dr. Smith has read about a new inspection technique A that looks as if it might outperform other techniques, like, for example, technique B. And so, she decides to run an empirical study to test this hypothesis. To do this, she puts out a call for final-year BSc in Software Engineering students to participate in the study. As a result of the call, she manages to recruit 16 students, and 8 are trained in the new technique and the other 8 in the pre-existing technique. During the experiment, each group applies the respective technique to the same program. She measures the number of defects detected as the response variable. Table 1 shows the results (aggregated by group).

Technique A	**Technique B**
Means (Y_e) = 12.000 defects	Means (Y_c) = 11.125 defects
Standard Deviation (S_e) = 2.673	Standard Deviation (S_c) = 2.800

Table 1: Results of the experimental study by Dr. Smith

Based on these values, Dr. Smith runs a hypothesis test (a t-test assuming variances to be equal) with $\alpha = 0.05$. This test returns a p-value of 0.53. Therefore, technique A cannot be said to perform better than B.

Although the results are not promising, Dr. Smith decides to go ahead with their publication in the hope that the experiment will be replicated and the aggregation of data will better explain the comparison between A and B. Dr. Smith submits the paper and, at the end of the review process, receives the following assessment.

Originality:	*Neutral*
Importance:	*Strong Reject*
Overall:	*Reject*
Detailed comments:	*Your paper is interesting but has two major pitfalls. First, it was developed with very few experimental subjects (which, on top of this, are not practitioners). Second, the study results are not significant, meaning that it provides no useful information.*

Figure 1: Results of the paper review process

The above example, albeit fictitious, is representative of many real pieces of empirical software engineering (ESE) research. On the one hand, many researchers interpret hypothesis testing too restrictively, focusing on whether or not the results are significant. On the other hand, there is a tendency not to take experimental studies that were built with students as evidence, as this research is not considered to be extrapolable to real-world environments.

However, there is a shortage of subjects (be they practitioners or students) that are willing to participate in experimental studies. Additionally, the more subjects an experiment has, the more costly it will be in terms of workload, infrastructure, etc., and this can discourage researchers. On the other hand, the cost of experiments run with fewer subjects is likely to be more affordable. These factors clearly limit SE researchers' prospects of being able to generate new empirically validated knowledge.

Fortunately, there are some alternatives for exploiting the results of small-scale studies. In this paper we will focus on one: meta-analysis. Essentially, meta-analysis is a statistical technique for aggregating more than one study, thereby increasing the number of experimental subjects involved in the hypothesis testing and outputting more reliable results. In our research we have analyzed whether meta-analysis [1] could be applied in ESE to combine the results of several small-scale experiments, with the aim of increasing the power of experiments with small samples.

We will proceed as follows. Section 2 will describe how sample size affects hypothesis testing. In Section 3 we will outline how to use meta-analysis to combine the results of more than one small study and thus increase their power. Finally, Section 4 will discuss whether meta-analysis is reliable when applied to ESE.

2. SAMPLE SIZE AND ITS RELATIONSHIP TO TYPE I AND TYPE II ERRORS

Any statistical test is subject to two types of errors: α, or type I error, and β, or type II error [2]. These errors occur due to the uncertainty associated with estimating population parameters (means and standard deviation) from a sample of the population. Remember that an experiment observes what happens in a sample (the subjects that tested the techniques) to estimate what happens in a population (the reality of the tested techniques).

As Table 2 shows, α is the error associated with the alternative hypothesis (H_1: there is a difference between tested techniques) being accepted when the null hypothesis (H_0: there is no difference between the tested techniques) holds for the population, and β is the likelihood associated with the opposite event.

		Hypothesis that is true for the population	
		H_0	H_1
Test result	H_0	Correct decision	β (Type II error)
	H_1	α (Type I error)	Correct decision

Table 2: Decision of the statistical test

It is more dangerous for an experiment to lead to the belief that there actually is a difference between two tested techniques when there really is none (error α) than to believe that there is no difference (because none is observed in the experimental sample) when there really is (error β). Therefore, the value of α is set at extremely low values, such as 0.1, 0.05 or even 0.01 (10%, 5% and 1%, respectively).

Unfortunately, α and β are not independent: according to statistical theory, a hypothesis test is characterized by five factors [3]: α, β, the mean difference *d*, the level of variation of the response variable *s* (measured as the variance or standard deviation) and the number of experimental subjects, or, to be more precise, sample size, *n*. Equation 1 shows the relationship between these factors[1] [4], where *z* represents the typified normal distribution,

$$z_{1-\beta} = \sqrt{\frac{n}{2}}\frac{d}{s} - z_{1-\alpha}\,. \qquad (1)$$

These five factors form a closed system. This means that an increase or decrease in any one of the factors leads to increases or decreases in the others. In practice, the factor that really is affected is *n*, as type I (α) and type II (β) errors are set beforehand, and both *d* and *s* are circumscribed by the experimental context and cannot therefore be manipulated at liberty by the researcher [5].

This is perhaps the most important, albeit not the only, reason why experiments are required to have a large number of experimental subjects. When the number of experimental subjects is small and α is set at 0.05, β returns very high values.

[1] *d* and *s* are usually presented as a single factor $d \div s$, called *effect*.

Let us go back to the example of Dr. Smith's experiment. Applying Equation 1 we get β = 0.83, that is, the test will detect significant differences 17% of the time, whereas it will fail to do so in 83% of the cases, even though they possibly do exist in the population/reality (as in fact they do, see Table 2, row 1 and column 2).

The influence of the number of subjects on type II error is even clearer if we look at how β decreases as more experimental subjects join, all other factors being equal.

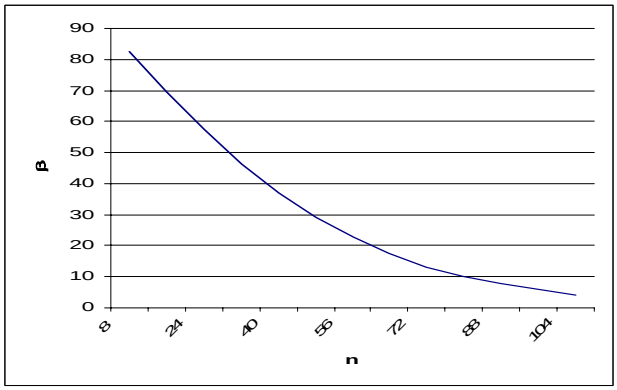

Figure 2: Decrease in type II error against the increase in n

It is usual practice to use the term *reliability* instead of α to refer to type I error and *statistical power* instead of β to refer to type II error. Reliability is calculated as 1 - α and power as 1 - β. For an experiment to be considered reliable, it is usual to set type I error at α = 0.05 (that is, a reliability of 0.95 or 95%) and type II error at β = 0.2 (that is, a power of 0.8 or 80%).

As Figure 2 shows, Dr. Smith would have needed a total of 120 subjects (60 in each group) for her experiment to be considered reliable.

Fortunately, there are several strategies designed to overcome the problems of low power caused by the use of experiments that have few experimental subjects. In this paper, we will look at meta-analysis.

3. HOW TO EXPLOIT SMALL-SCALE EXPERIMENTS

Meta-analysis is a statistical technique for combining the results of more than one experiment developed previously to achieve a greater statistical power than any of the individual experiments on their own [5].

Although usually associated with medicine, the term meta-analysis as it is now known was developed in psychology.

In many cases of psychology, the treatments studied have very small effects on experimental subjects, meaning, as illustrated in Figure 2, that experiments need a very large sample size (usual guidelines are around 150). In many cases, however, not that many subjects are available for experiments and studies reporting insignificant effects predominate over others that do detect significant effects, as studies of low statistical power accumulate.

This was the way things were in psychotherapy, the specialized field with which Dr. G.V. Glass, creator of meta-analysis as we

know it today, was concerned. Using an argument very similar to the one brandished in ESE today (small studies are useless), psychotherapy was judged to be ineffective. Dr. Glass, who did not agree with this interpretation, took a different road to demonstrate his belief: instead of excluding studies (on the grounds of their size or statistical significance), he tried to consider as many studies as possible upon which to base his findings. Looking back, the hardest thing was to find a way of aligning the wide range of metrics used in the different replications to measure the response variables [6]. Appendix I sets out the technical details. The solution was to come up with what is today the well-known concept of effect size, briefly mentioned in Section 2. Effect size is a non-scalar measure calculated as the difference between the treatment means divided by the pooled standard deviation. After calculating the effect size of each experiment, all Glass had to do was average the results of the individual experiments to arrive at a *global effect* using a procedure dating back to the mid-19th century [7]. This value represents the effect that, theoretically, a single experiment having a greater sample size and, consequently, a smaller type II error than any of the original experiments would have achieved. This way he demonstrated that psychotherapy was indeed effective [8].

The parallelisms with ESE, in respect of the potential contribution of small studies, are evident. For example, suppose that Dr. Smith published her paper on her laboratory web site. Later Dr. Thomas visited the web site, found the experiment interesting and decided to replicate it. In this case, Dr. Thomas managed to recruit no more than eight advanced MSc in Software Engineering students, four of which he assigned to each of the experimental groups. His results are shown in Table 3.

Technique A	Technique B
Means (Y_e) = 13.000 defects	Means (Y_c) = 12.000 defects
Standard Deviation (S_e) = 1.800	Standard Deviation (S_c) = 1.700

Table 3: Results of Dr. Thomas' experimental study

Dr. Thomas ran a t-test on these results (assuming variances to be equal at $\alpha = 0.05$) and also found insignificant differences (p-value = 0.57). What would happen if these two studies were combined using meta-analysis to achieve a new result? Would the differences be significant? Would the test be more powerful?

In response to these questions, the sample size is still not big enough to return significant results (there are only 12 subjects per technique). Figure 3 (showing the statistical power of the meta-analysis for a population with an effect size [2] of 0.5 and α=0.05) indicates that about 70 experimental subjects would be necessary for a meta-analysis to achieve what is usually considered as a discriminative statistical power ($1-\beta = 0.8$).

[2] The example used throughout the article has an effect size of approximately 0.5. In conformity with the guidelines, we have assigned the largest variances to the experiments with the smallest sample sizes, which means that experiments have effect sizes of different magnitudes. Equation 4 (described in Appendix I), for example, can be used to calculate the effect size for the studies by Dr. Smith and Dr. Thomas:

$$d = 0.870 * \frac{13 - 12.250}{1.768} = 0.369$$

$$d = 0.954 * \frac{12 - 11.125}{2.737} = 0.302$$

The statistical power, however, has improved in part. Whereas Dr. Smith's and Dr. Thomas' tests had a power of 0.17 and 0.11, respectively, the meta-analysis achieved a power of 0.13. Note that if it had been possible to use 8 + 4 = 12 subjects per group in a single experiment, it would have been possible to achieve a power of 0.22.

Using meta-analysis it is possible to gradually increase the statistical power as more experiments are added. This way experiments with a small sample size can supplement each other.

The more experiments (no matter how small the number of subjects per experiment is) that are aggregated using meta-analysis, the more powerful the results and, consequently, the greater the possibility of detecting false-negatives will be.

Suppose that there are three more replications of Dr. Smith's research, whose results are shown in Table 4 (note that they all return insignificant results). Figure 4 charts how the power of the meta-analysis increases as more of these studies are added.

In this example, even though the test fails to achieve the desired power level of 80%, it does, in any case, manage to output significant differences at a power of almost 57% (which is much greater than the best experiment separately, estimated at 24%). This is noteworthy, as the example was designed based on the fact that the inspection technique efficiency actually IS different.

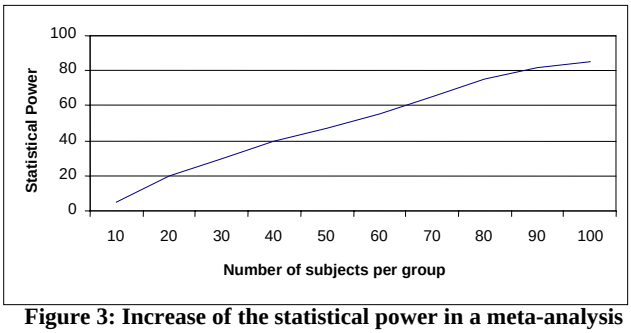

Figure 3: Increase of the statistical power in a meta-analysis

Study	Ne	Me	Se	Nc	Mc	Sc	p-value	Power
3	9.00	11.00	1.80	9.00	10.10	1.70	0.30	17.58
4	10.00	10.00	1.40	10.00	9.10	1.50	0.20	20.34
5	12.00	9.00	1.60	12.00	8.10	1.80	0.20	24.63

Table 4: Results of replications[3]

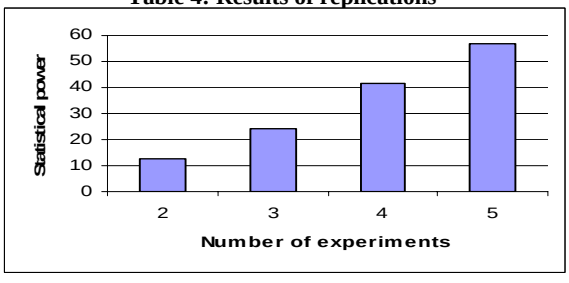

Figure 4: Increase in the statistical power by accumulating small-scale replications

[3] N = number of subjects; M = mean; S = standard deviation; e = results linked to the new technique; c = results linked to the perspective-based technique.

So far we have given an example of how meta-analysis can be used to take advantage of studies with small sample sizes that, separately, return results that are insignificant but, together, could provide valuable evidence. The question now is, are these results possible in the real world?

4. IS META-ANALYSIS RELIABLE?

Nowadays meta-analysis has a sound reputation as a statistical technique (although this has not always been the case [7]). In eminently experimental disciplines, like medicine or physics, for example, meta-analysis is regularly used according to their respective traditions, and its results are considered to be highly reliable within both disciplines [9].

Nonetheless, we have compiled possible objections to a researcher using meta-analysis in SE with the aim of increasing the power of his or her experimental results:

1. **There are many co-variants in SE**. This leads to studies that return inconsistent results and cannot be combined with each other.

   This argument can be traced back to Miller [10]. Miller tried in [10] to apply meta-analysis to a small set of SE experiments. The experimental data were very heterogeneous, that is, appeared to be taken from different populations (for a better definition of heterogeneity, see Appendix I). Miller's finding was:

   "the discipline must embark upon a period of improvement to reduce the variability between replicated experiments or experiments examining the same hypothesis."

   This is quite true, but, as Miller himself said, this problem should be viewed as a challenge but not as a limitation. There is more than one alternative for dealing with experimental heterogeneity in meta-analysis [11], but they all involve having a large experimental database. Precisely because of this, we propose in this paper that experiments with small sample sizes can also be useful. In other words, instead of shying away from running experiments, we should experiment more.

2. **Articles on experiments in SE are of poor quality**

   Miller [10] again could be considered the source of this criticism, although it has been repeated in several works since [12]. There can be no doubt that, as Miller himself says, we have to learn to experiment better in SE. This is precisely what ESE is all about. But, again, this limitation can be construed as a challenge. Our research [13], as well as investigations conducted in other disciplines [14], shows that the concept of experimental quality is very elusive. It is not easy to determine when an experiment returns true or false results. The really important thing is to have enough experiments to be able to detect which are biased and which are not. Again, the key is the availability of replications irrespective of how alike and good they are.

3. **Meta-analysis requires more than 4 or 8 subjects per study**

   Very true. One of the hypotheses underlying most statistical techniques (including meta-analysis) is large sample theory. This means that meta-analysis does not achieve the levels of reliability and real power specified levels by the theory unless it includes a high number of subjects (typically 30). Fortunately, research conducted recently has been able to reduce this number considerably. Back in 1986, Larry V. Hedges (one of the most influential researchers in the meta-analysis world) demonstrated that meta-analysis techniques are reliable as of 10 subjects per group [15]. Now, in even worse conditions than set by Hedges, we have been able to demonstrate using Monte Carlo simulations that meta-analysis (specifically, its most popular variant WMD, see Appendix I) is reliable even when aggregating studies with as few as four subjects per group [16]. This is a key finding, as it opens the door to the aggregation of very small-sized studies.

4. **SE experiments do not report enough data to run meta-analyses**

   In many cases this is true, and is also perfectly understandable. It has happened in other disciplines as well [17]. Nobody told SE researchers that they had to report the number of subjects, means and variances. On the contrary, for a long time we thought that hypothesis testing and p-values were the really important results of our experiments. However, this problem is easy to solve. There are now reporting guidelines, such as [18], which, once interiorized by researchers, will remedy this defect. In the meantime, we can use non-parametric techniques (see Appendix II) or simply e-mail the authors of a poorly reported experiment and request the raw or aggregated data.

5. **Experimentation in SE is not as rigorous as in other sciences**

   We are inherently "softer", because of the multiple factors involved and the low quality of the data. This makes it impossible to aggregate separate experimental results using meta-analysis.

Many SE experimenters agree with the above statement. Those that do will be quite surprised to find out that it was more or less what L.V. Hedges said in 1987 about a completely different discipline: education. The literal citation is [19]:

> *"Those of us […] know intuitively that there is something 'softer' and less cumulative about our research than about those of physical sciences. […] distinguished researchers have cited the pervasive presence of interactions or historical influences as reasons not to expect a cumulative […] science. Still others have cited the low quality of data […] as a barrier."*

ESE researchers would do well to read the above article. While it is true that other articles cause similar impressions (for example, papers on experimental economics [20]), the above adds to the strength of data argument to the discourse. Specifically, Hedges

compares the reliability of the data collected in experiments in the field of education with data gathered in physics. His finding was:

> *"What is surprising is that the research results in the physical sciences are not markedly more consistent […] data do suggest that results from replicated experiments do not always tend to be consistent"*

The question is, why should experiments in education be more consistent than in SE? Could it be that we have a very pessimistic view of ourselves?

5. CONCLUSIONS

In this paper we have shown that there are options open to researchers to generate pieces of empirical SE knowledge more efficiently than they do today. We have shown that meta-analysis is able to increase the power of experiments, enabling a set of small studies that individually do not return statistically significant differences to do so, if taken together. This way we can solve some of the problems related to the accumulation of a sizeable number of experimental subjects by a single researcher, as we can put together a large-scale experiment by meta-analyzing replications of small studies.

In summary, we can say that:

1) It is worthwhile running experiments even if they do not have many experimental subjects, as they can be combined to form a larger scale study;
2) It is worthwhile publishing studies even if they do not return significant results, as this can be very often due to the low power of the statistical method.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sörumgård, S., Zelkowitz, M.; 1996; *The empirical investigation of perspective-based reading*, International Journal on Empirical Software Engineering, Vol. 1, No. 2; pp. 133–164.

[2] Everitt, B.; 2003; *The Cambridge Dictionary of Statistics*, CUP; ISBN: 0-521-81099-x

[3] Cohen, J.; *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.) 1988. ISBN 0-8058-0283-5.

[4] Cohen J.; 1988; *Statistical Power Analysis for the Behavioral Sciences*. (2nd ed.); ISBN 0-8058-0283-5.

[5] Lawrence D., Betsy J. Becker; 2003; *How Meta-Analysis Increases Statistical Power; Psychological Methods by the American Psychological Association*; Vol. 8, No. 3, 243–253

[6] Hunt, M.; 1997; *How science takes stock: the story of meta-analysis*; New York, Russell Sage Foundation; ISBN: 0871543893

[7] Chalmers I., Hedges L., Cooper H.; 2002; *Abrief history of research synthesis*; Eval Health Prof March;25(1):12–37.

[8] Fischer, L.; Wilson, G.; 1985; *Clinical Psychology Review*; Elsevier

[9] Cochrane; 2008; *Curso Avanzado de Revisiones Sistemáticas*; www.cochrane.es/?q=es/node/198

[10] J. Miller. *Applying meta-analytical procedures to software engineering experiments. University of Strathclyde*, Research Report EFOCS-30-98.

[11] Glass, G; 2000; *Meta-Analysis at 25*; http//glass.ed.asu.edu/gene/papers/meta25.html

[12] Kampenes, B.; 2007; *Quality of design, analysis and reporting of software engineering experiments: A systematic review*; PhD thesis, University of Oslo; http://simula.no/research/se/publications/Simula.SE.178

[13] Saxena, H.; 2009; *Prospective study for the quality assessment of experiments included in systematic reviews;* Master Thesis; Universidad Politécnica de Madrid.

[14] Lee, Y.; Nelder J; 1998; *Generalizad linear models for the analysis of quality-improvement experiments*; the Canadian Journal of Statistics; Vol. 26, No. I; pages 95-105

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

[16] Pollo-Cattáneo, F. 2009. *Análisis de Precisión de Técnicas de Agregación en Contextos Experimientales Poco Maduros. Master Thesis*; Instituto Tecnológico de Buenos Aires.

[17] Moher D.; Jones A.; Lepage L.; 2001; *Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation*; JAMA;285:1992-5.

[18] Runeson, P. and Höst, M.; 2009; *Guidelines for conducting and reporting case study research in software engineering*; Empirical Software Engineering; Springer Netherlands; ISSN 1573-7616; Volume 14, Number 2.

[19] L. Hedges. *How hard is hard science, how soft is soft science? The empirical cumulativeness of research*; American Psychologist, 42(2):443-455, 1987

[20] Smith, V.; 2008; experimental economics; *The New Palgrave Dictionary of Economics*, 2nd Edition, Edited by Steven N. Durlauf and Lawrence E. Blume

[21] Hunt, Morton; H*ow Science takes stock: the story of meta-analysis*; Russell Sage Foundation: New York; 1997

[22] Glass, G; 1976; P*rimary, secondary, and meta-analysis of research*; Educational Researcher 5: 3-8

[23] Goodman C.; 1996*; Literature Searching and Evidence Interpretation for Assessing Health Care Practices;* SBU; Stockholm.

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

[25] Kitchenham, B. A.; 2004; *Procedures for performing systematic reviews*. Keele University; TR/SE-0401. Keele University Technical Report.

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

[27] Hedges, L. V. & Pigott, T. D. (2001). *The power of statistical tests in meta-analysis*. *Psychological Methods, 6(3),* 203-217.

[28] Miguez, E. & Bollero, G; 2005; *Review of Corn Yield Response under winter cover cropping systems using Meta-Analytic Methods*; Crop Science Society of America

APPENDIX I: Meta-analysis statistics

As mentioned in Section 3, the quantitative synthesis of experimental studies involves the aggregation of the results of a set of previously identified experiments analysing the performance of a pair of treatments (circumstances or interventions [21]) with the aim of giving a synthetic quantitative estimate of all the available studies [15]. As a quantitative synthesis aggregates studies previously developed and analyzed by their authors, this type of study also goes by the name of meta-analysis, a term coined by Glass [22] in the field of psychology and used in most sciences.

If all the studies included in a meta-analysis process were equally accurate, it would suffice just to average the results of each study to arrive at a final conclusion. In practice, however, not all studies have the same accuracy. For this reason, when they are combined, a greater weight must be assigned to the studies from which more reliable information can be gained. To do this, we combine the results using a weighted mean [15]. Also as the results of the different studies are sometimes measured differently, the dependent variable in a meta-analysis must be able to combine the different metrics used. This is done by estimating the effect size. *Effect size* is a standardized, non-scalar estimator of the relationship between treatments (for example, the number of defects detected by an inspection technique). The weighted mean difference (WMD) is the method par excellence for use with continuous variables (commonly used in SE) [23]. This method is conceptually simple [24]: the effect size of each study is estimated as the mean difference divided by the pooled variance of both treatments (Equations 2 and 3):

$$g = \frac{Y^E - Y^C}{S_P}$$

g is the effect size

Y is the mean of the experimental (E) & control (C) groups

S_p is the pooled standard deviation (2)

$$S_P = \sqrt{\frac{(n^E - 1)(s^E)^2 + (n^C - 1)(s^C)^2}{n^E + n^C - 2}}$$

S is the standard deviation of the experimental (E) & control (C) groups

n is the number of experimental subjects in the experimental (E) & control (C) groups (3)

Note that Hedges and Olkin [15] optimized Equation (2) by adding a correction factor "J" (4). This factor is used to increase the reliability of the method when the studies to be aggregated do not have many experimental subjects. The new equation is known as *"d"* and is recommended in [25] for use in SE.

$$d = J\frac{Y^E - Y^C}{S_P}$$

d is the effect size

Y is the mean of experimental (E) & control (C) groups

S_p is the pooled standard deviation

J is Hedges' correction factor (4)

After estimating the effect size, we have to estimate the confidence interval (see Equations 5 and 6). The confidence interval provides a range of values (minimum and maximum) within which the mean population is located. When the difference between the two means is significant, the confidence interval of the effect size must not contain the value 0 (value for which both means are equal), that is, equality is not a possibility within the range of possible effect size values.

$$d - Z_{\alpha/2}\sqrt{v} \le \lambda \le d + Z_{\alpha/2}\sqrt{v}$$

d is the effect size

Z is the number of standard deviations that separate, at a given significance level, the mean from the endpoint. Generally, 1.96 ($\alpha = 0.05$) is used.

v is the standard error. (5)

$$v = \frac{\tilde{n} + d^2}{2(n^E + n^C)}$$

$\tilde{n} = (n^E + n^C) / (n^E * n^C)$ (6)

After estimating the effect size for each study, we can estimate the general or global effect. To do this, we use the following Equation (7):

$$d^* = \frac{\sum d_i / \sigma^2{}_i(d)}{\sum 1/\sigma^2{}_i(d)}$$

d^* is the global effect size

$\sum d_i / \sigma^2{}_i(d)$ is the sum of the individual effects

$\sum 1/\sigma^2{}_i(d)$ is the sum of the inverse variance (7)

To make the results easier to interpret, they are generally charted as a forest plot. The forest plot shows the values of the confidence intervals on the x-axis, whereas the y-axis plots the different meta-analyzed experiments, together with the global effect. This way, it is extremely easy to see which confidence interval contains the value 0 and how near to or far from 0 it is.

Equation (7) is also known as the fixed effects model, because it assumes that the variation between the results of the experiments is due exclusively to experimental error [26]. There are improvements on this formula for use when there are other types of variations, such as publication bias and experimental heterogeneity. In this paper, however, we will deal with the fixed effects model only, because, as SE meta-analyses typically include few experiments, it is very hard to precisely estimate the variables (e.g. between study variance) required by such improvements. For more details on how to apply the above formulae, interested readers are referred to [15].

One of the drawbacks of the meta-analysis 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) [15]. As the effect size estimated using WMD is symmetrical with respect to the treatments, a positive value indicates that the experimental treatment mean is greater than the control treatment mean, whereas a negative value means the opposite.

Finally, let us look at Equations 8 and 9 [27]. They are the functions of statistical power estimation for a meta-analysis used in Section 3:

$$power = 1 - \phi(C_{\alpha/2} - \lambda) + \phi(-C_{\alpha/2} - \lambda) \quad (8)$$

$\phi(x)$ is the standard normal cumulative distribution function

$C_{\alpha/2}$ is the standard normal critical value for the two-sided test at level α

$$\lambda = \frac{(\theta_1 - \theta)}{\sqrt{V}} \quad (9)$$

θ_1 is the sample effect size

θ is the endpoint effect size

v is the standard error

The functions of meta-analysis power estimation can accurately estimate the statistical power of a meta-analysis when the set of experiments that are part of the aggregation process are homogeneous [15, 26]. As ESE studies are generally small, however, the variations in the results are, as a rule, large due to experimental error. Such variations tend to lead to false heterogeneity.

When this happens, the power of the meta-analysis (we mean a fixed effects model) tends to decline, as, incidentally, Hedges and Olkin found [15] and our experiments confirmed [16]. Figure 5 illustrates a chart showing the estimated power for the studies with low (0.2), average (0.5) and large (0.8) effect sizes, where $\alpha = 0.05$. The power values are systematically smaller than the values calculated using Equations 8 and 9.

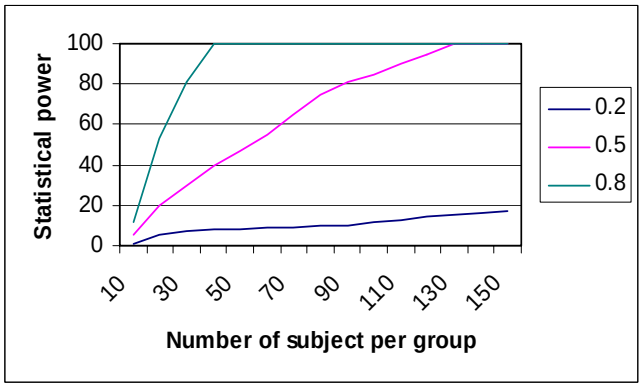

Figure 5: Simulated power for a meta-analysis

APPENDIX II: Alternative meta-analysis methods

Although WMD is the most commonly used meta-analysis technique, there are other alternative techniques with beneficial properties for certain experimental settings (for example, when the experiments have reporting defects). These techniques are: the parametric and non-parametric response ratio, and statistical vote counting. These methods are described in the following.

Response Ratio

The response ratio is the method of meta-analysis recommended for synthesis processes enacted within the field of ecology [24]. It involves estimating an effect index or ratio between two treatments by calculating the quotient of the two means. This quotient estimates the how much better one treatment is than the other [24]. For example, a ratio of 1.3 will indicate that the experimental treatment is 30% better than the control treatment, whereas a ratio of 1 will signify that there is no difference in the performance of the two treatments. A ratio of less than 1 means that the control treatment is better than the experimental treatment.

The method is applied similarly to WMD. First, we have to estimate the ratio of each experiment and then, based on these ratios, estimate the global ratio using a weighted mean of the individual ratios (10), where each study is weighted based on its inverse variance (11):

$$RR = \frac{\sum_{i=1}^{k} W_i^* RR_i}{\sum_{i=1}^{k} W_i^*} \quad (10)$$

RR is the global effect size

RR_i is the individual effect size

W_i is the weight factor = $1/v_i$

$$v_i = \frac{S^{E2}}{n^E Y^E} + \frac{S^{C2}}{n^C Y^C} \quad (11)$$

v is the standard error

S^2 is the standard deviation of the experimental (E) & control (C) groups

Y is the mean of the experimental (E) & control (C) groups

n is the number of experimental subjects in the experimental (E) & control (C) groups

To make the combination of a set of studies more accurate, the natural logarithm was added to the method. Applied to the effects of the individual studies, this logarithm linearizes the results and normalizes their distribution. Note importantly that after estimating the global effect index, the anti-logarithm must be applied to the result to calculate the final effect size. For more details on how to apply the above formulae, readers are referred to [24].

There is a non-parametric version of the method. This method is essentially the same as the above parametric version, the difference being that the studies are weighted by the number of subjects (12) and not by the inverse variance [28]. The main advantage of this version of the technique is that no knowledge of the experiment variances is required for its application. This is especially useful when reports are incomplete, as is often the case in SE.

$$v = \frac{n_C + n_E}{n_E n_C} + \frac{Ln(RR^2)}{2(n_C + n_E)} \quad (12)$$

v is the standard error

n is the number of experimental subjects in the experimental (E) & control (C) groups

Statistical Vote Counting

The vote counting method requires very little information to be applied. All we need to know in this case is whether or not there is a difference between the treatment means (which we will call "vote") and the number of experimental subjects used in each study (used as a weighting factor of the "vote") [15]. Based on these data, a maximum likelihood estimation process is enacted. The aim of this process is to determine the effect size (generally

selected from a list ranging from -0.5 to 0.5) that WMD would most likely have estimated if all the data had been available. The main estimation function is:

$$L(\delta \mid X_1,\ldots,X_i) = \sum_{i=1}^{k}\left\{\begin{array}{l} X_i \ln\left[1-\phi\left(-\sqrt{\tilde{n}}\delta\right)\right]+ \\ (1-X_i)\ln\phi\left(-\sqrt{\tilde{n}}\delta\right)\end{array}\right\} \quad (13)$$

$L(\delta|X_1,\ldots.X_n)$ is the likelihood of the effect size

δ is the effect size to be tested

X_i is the value of the vote in each study

$\tilde{n} = (n^E+n^C) / (n^E*n^C)$

Note that the effect size returned by SVC is similar to WMD and, therefore, has the same meaning. For more details on how to apply the specified formulae, see [15].