2011 International Symposium on Empirical Software Engineering and Measurement

The risk of using the Q heterogeneity estimator for software engineering experiments

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

Enrique Fernández
Universidad Nacional de La Plata
Buenos Aires, Argentina
enriquefernandez@educ.ar

Ramón García-Martínez
Universidad Nacional de Lanus
Buenos Aires, Argentina
rgarcia@unla.edu.ar

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

***Abstract*— *Background:* All meta-analyses should include a heterogeneity analysis. Even so, it is not easy to decide whether a set of studies are homogeneous or heterogeneous because of the low statistical power of the statistics used (usually the Q test). *Objective:* Determine a set of rules enabling SE researchers to find out, based on the characteristics of the experiments to be aggregated, whether or not it is feasible to accurately detect heterogeneity. *Method:* Evaluate the statistical power of heterogeneity detection methods using a Monte Carlo simulation process. *Results:* The Q test is not powerful when the meta-analysis contains up to a total of about 200 experimental subjects and the effect size difference is less than 1. *Conclusions:* The Q test cannot be used as a decision-making criterion for meta-analysis in small sample settings like SE. Random effects models should be used instead of fixed effects models. Caution should be exercised when applying Q test-mediated decomposition into subgroups.**

Keywords- Meta-analysis, heterogeneity, reliability, statistical power, effect size, weighted mean difference (WMD).

I. INTRODUCTION

Meta-analysis is coming to be an important tool for aggregating the results of software engineering (SE) experiments, e.g., research carried out by Dyba et al. [1] and Ciolkowski [2]. To run a meta-analysis it is essential to check whether or not the primary studies are homogeneous, that is, we have to verify that the differences between the results of the studies are due to a random error and not to an effect caused by some uncontrolled external factor that is obscuring the final result [3].

There are several reasons why a set of studies can turn out to be heterogeneous. The most evident is the presence of moderator variables, but methodological issues related to experimental design and operation can also have an influence. In any case, heterogeneity must be tackled (either through decomposition into subgroups, meta-regression or applying random effects models). To do this, it has to have been detected beforehand.

There are several methods for evaluating the level of heterogeneity in a set of experiments. The most commonly used method is the Q test proposed by DerSimonian and Laird [4], which is generally recommended on the grounds of validity and computational simplicity [5]. The drawback is that the statistical power (capability of determining that a set of studies is heterogeneous) of the Q test is low when it is applied to a small number of experiments (as a general rule, the literature points to 10 experiments as being the lower bound [6] [3]). The Q test also appears to suffer from low power when there are few experimental subjects in the experiments to be aggregated [7]. In this latter case, the Q test's power may not improve even though more experiments are added to the meta-analysis.

There are some alternatives to the Q test for studying heterogeneity. The most popular is the visual examination of the overlap of the confidence intervals in a forest plot [8]. This type of analysis is sometimes recommended to offset the Q test's low power [9]. Unfortunately, it has been observed that visual examination is not very systematic, and the findings largely depend on the researcher applying the method [10].

Some researchers, such as [11] and [7], have examined the statistical power of the Q test. This research, conducted in the field of medicine, generally confirms that the Q test is not powerful when meta-analysis is applied to a small number of experiments or the experiments do not have many subjects. However, the above studies covered a much greater number of experiments and subjects per experiments than surveys conducted in empirical SE do nowadays.

This study is part of a series of reviews aiming to establish which statistical methods are best for the meta-analysis of SE experiments. The first of these papers is [12], where the reliability and statistical power of several fixed effects models was established. This study aims to analyse the power of the Q test in small sample settings, which are common in empirical SE today. Discovering the situations in which the Q test is powerful enough will help us to establish well-defined decision rules about the use of the fixed effects models studied in [12]), random effects models (which are to be examined in coming studies) and mechanisms for explaining heterogeneity (such as the above-mentioned decomposition into subgroups).

To analyse the power of the Q test, we used the Monte Carlo method to simulate multiple meta-analyses and calculated the power of the Q test in each case. The output results corroborate that the power of the Q test is low and establish lower bounds under which Q is simply not

978-0-7695-4604-9/11 $26.00 © 2011 IEEE
DOI 10.1109/ESEM.2011.15

powerful enough to positively determine whether a set of experiments is homogeneous or heterogeneous.

The article is structured as follows. Section 2 describes how heterogeneity detection methods work. Section 3 describes the existing studies on the power of the Q test. Section 4 specifies the goals of this research. Section 5 describes the applied research methodology. Section 6 presents the results of the Monte Carlo simulation. Section 7 discusses the results. Finally, Section 8 advances some provisional findings.

II. BACKGROUND

A. Concept of heterogeneity

A set of experimental replications that analyse the performance of a pair of treatments will always output different results due to random error [13]. This is because many aspects of an experiment (population, training, duration, etc.) can be neither randomized nor blocked absolutely satisfactorily. Apart from chance, differences between experimental replications could have a systematic cause. Possible grounds are the presence of moderator variables and discrepancies in experimental design and operation.

A forest plot is a particularly simple way of visualizing the differences among experimental replications. A forest plot represents the effect sizes of the experiments covered by the meta-analysis, as well as the global effect size, together with their respective confidence intervals [8]. For the purposes of the following discussion, *effect size* shall mean the standardized differences *g* between a treatment group and a control group, calculated according to Hedges and Olkin's equations [20], as this is the method most commonly used in SE (e.g., [1-2]). Note that *d* and *g* are used in the literature to denote the same effect size; notationally, we will use *d* as a general rule, and *g* where necessary in reference to the specific equations to be used.

Figures 1 and 2 are examples of homogeneous and heterogeneous forest plots, respectively. When the experiments included in a meta-analysis are homogeneous, their confidence intervals (with an arbitrary but similar significance level α in all experiments) tend to overlap, that is, the experiments return very similar effect sizes (the effect size is very close to 0.3 in all the experiments in Figure 1) and, consequently, the respective confidence intervals are aligned).

On the other hand, when any experiment does not overlap with the confidence intervals of the other experiments, we have a completely different scenario. For example, consider experiment 2 in Figure 2. The reported effect in experiment 2 is 0.8, very far removed from the 0.3 of the other experiments. The confidence interval of experiment 2 is centred on 0.8, but this is nowhere near the other confidence intervals. Consequently, experiment 3looks to be different from experiments 1 and 2 over and above the random variation to be expected in any set of experimental results (although more sophisticated tools than a mere visual examination are required to be able to confirm such a claim).

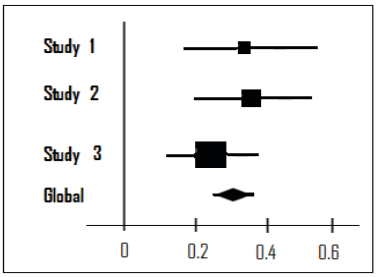

Figure 1. Forest plot showing a homogeneous set of studies

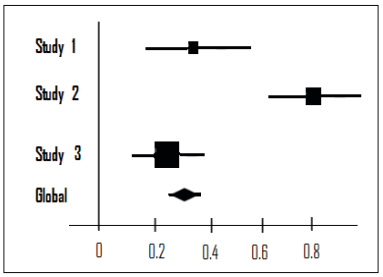

Figure 2. Forest plot showing a heterogeneous set of studies

B. Heterogeneity tests

Although the visual examination of a forest plot may suggest the presence of heterogeneity, the studies run in other disciplines argue against the use of this device [10], as the result of visual inspection has been found to largely depend on the subjective opinion of the researcher applying the technique [10]. To determine the homogeneity or heterogeneity of a set of studies, it is preferable to use statistical heterogeneity tests.

The best known and widely used method for determining the heterogeneity of a set of studies is the Q test proposed by DerSimonian and Laird [4], which is based on a test developed by Cochrane [14]. The analytical expression of the Q test is shown in (1).

$$Q = \sum_{i=1}^{k} w_i g_i^2 - \frac{\left(\sum_{i=1}^{k} w_i g_i\right)^2}{\sum_{i=1}^{k} w_i} = \sum_{i=1}^{k} w_i \left(g_i - \bar{g}\right)^2 \quad (1)$$

k: number of studies
w_i: weight of study i
g_i: effect of study i
$\bar{g}$: global effect

The symbols have been used intentionally to assure that (1) is familiar to SE researchers. *g* is Hedges and Olkin's effect size index [20], whereas w_i are the weights calculated using the same equation. However, the Q test is independent of the effect metric (effect size, odds ratios, etc.) [5], meaning that other equations, apart from Hedges and Olkin's [20], are applicable.

The Q test has a χ^2 distribution with *(k–1)* degrees of freedom. Q can be used in two different ways. In its simplest form, a significant result of the Q test denotes the presence of heterogeneity. Q can also be used to calculate the between-study variance τ^2. Both concepts are closely interrelated, although we focus on the first case in this research. The usual

significance level is $\alpha=0.05$, although some authors recommend the use of $\alpha=0.1$ to increase the power of the test [3].

There are many other methods for studying the heterogeneity of a set of experiments, such as Z^2_k [15] or LTR [16]. So far, however, these methods have not been used much at all. Although some of these methods are very promising [17], it seemed premature to address these tests in this paper.

Additionally, there are alternative formulations of Q, such as the well-known I^2 [18]. I^2 is very popular, as it is easier to interpret than Q. Generally, though, it suffers from the same weaknesses as the Q test from which it is derived [19]. For this reason, the I^2 test is not included in this study, and the constraints that are identified for the Q test will be equally applicable to the I^2 test.

C. Q test limitations

It is well documented in the literature that the power of Q is low when the number of experiments included in the meta-analysis is small [20]. The biggest problem from the viewpoint of SE, however, is that the Q test is unable to determine the heterogeneity of experiments run with few experimental subjects [7].

It is the high variance typically associated with small experiments that leads the number of subjects per experiment to have an influence on the Q test. The weights w_i in (1) are calculated as the inverse of the variance $w_i = 1/v_i$ [20]. This has the effect of smoothing the Q value and is an obstacle to statistical significance being achieved.

It is probably easier to understand the influence of sample size using the graphical analysis introduced earlier. The high variances associated with small experiments widen the confidence intervals, extending the overlap between studies and reducing the chances of detecting heterogeneity.

Consider the forest plot shown in Figure 3. It shows the meta-analysis of four heterogeneous experiments using the weighted mean difference method (WMD) [20].

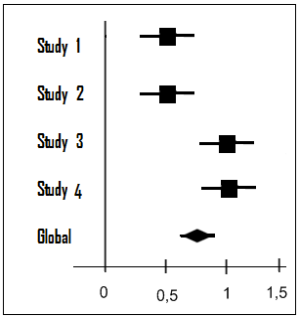

Figure 3. Forest plot resulting from aggregating four experiments each with 100 subjects

Experiments 1 and 2 have effect $d=0.5$, whereas experiments 3 and 4 have effect $d=1$. All four experiments have 100 subjects each, and the calculations are made using the significance level $\alpha=0.05$. The confidence intervals clearly do not overlap (the variances are small), and the heterogeneity test is plainly significant $Q=12.626$, p-value$=0.0056$. Now, if the experiment had been run with 25 instead of 100 subjects, the results would be as shown in Figure 4. The confidence intervals are visibly much greater than in the case above, and the overlap is more than evident. The heterogeneity test shows that $Q=3.087$, p-value$=0.378372$.

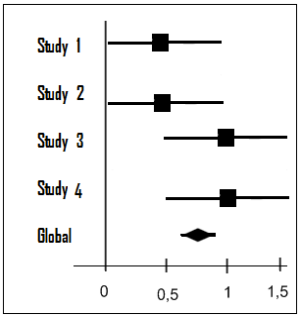

Figure 4. Forest plot resulting from aggregating four experiments each with 25 subjects

Note importantly that the addition of more experiments is, generally, not a solution to this problem, as one might think. For example, suppose that, instead of four, we had 20 experiments (10 with effect $d=0.5$ and 10 with effect $d=1$). The result of the meta-analysis would be as shown in Figure 5, where $Q=16.22$ (as Q is less than k-19 in this case-, it is pointless to estimate the p-value). This result is even less statistically significant than in the above case (Figure 4).

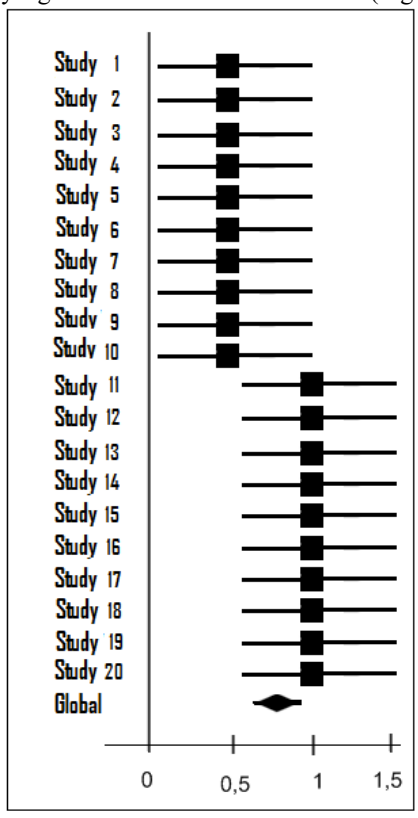

Figure 5. Forest plot resulting from aggregating 20 experiments each with 25 subjects

III. RELATED WORK

A number of researchers, such as Liang and Self [21], Jones et al [17], or Takkouche et al [22], have studied the power of the Q estimator. However, these studies are based on arrangements that are far removed from real-world SE experimentation today, especially in terms of the high number of experimental studies. There are, however, two studies that are fairly close to SE: Kim [11] and Hardy and Thompson [7].

Kim [11] analyses the power of Q through a Monte Carlo simulation varying the following parameters: number of experiments, number of experimental subjects per experiment and differences in the effect size. The parameters are set as follows:

- Number of experiments to be included in meta-analysis: 5, 10 and 30 experiments
- Number of subjects per experiment: 10, 30 and 300 subjects
- Difference between the effect sizes of the experiments included in the meta-analysis: 20%, 40% and 60% of the baseline effect size. In other words, let us suppose that we set the effect size of a baseline subset of experiments at $\delta=1$. These experiments are homogeneous. To produce heterogeneity, Kim [11] then generates another subset of experiments with effect sizes that are 20%, 40% and 60% greater than the effect size of the baseline subset (that is, $\delta=1.2$, 1.4 and 1.6).

As a result of this process, the author concluded that the power of the Q test is high (almost 100%) when the studies have 300 subjects, irrespective of the number of experiments that are aggregated or the difference of effect. On the other hand, the author considers the power of the Q test to be unacceptable when the experiments contain 10 or 30 subjects.

Hardy and Thompson [7] also analysed the power of the Q test by means of a Monte Carlo simulation. Unlike [11], though, they used the statistical power function of the Q test rather than (1) directly. The parameters used in the simulation are the number of experiments, the between-study variance (τ^2) and the relative weight of the experiment in the finding, where the values were as follows:

- Number of experiments: 5, 10 and 20 experiments
- Between-study variance (which defines the effect difference between studies): from 0.1 to 0.5.
- Relative of the weight of the experiment: from 10% to 90%.

In this paper, Hardy and Thompson try to determine how powerful Q is in terms of the above parameters, for example, to analyse how the weight of an experiment influences the power of Q. To do this, they set, as the most extreme case, a weight of 90% for one of the experiments in a 10-experiment meta-analysis and distributed the other 10% among the other 9, which were weighted as 1.1.

As a result of this simulation, they found that the power of the Q test was generally low. They also found that the number of experiments does not appear to be a factor that has a big influence on the power of Q, especially when there are experiments that account for most of the weight of the result. They concluded that this weakness of the Q method conditions the use of the fixed effects meta-analysis model, as, in principle, almost all the groups of studies will appear as heterogeneous.

IV. OBJECTIVES

In this paper we will study the power of the Q test in settings closer to experimental practice in SE than the contexts addressed by Kim [11] and Hardy and Thompson [7], that is:

- Meta-analysis with few experiments. For example, Dyba et al. [1] report 15 experiments per meta-analysis. Ciolkowski [2] reports meta-analyses with 5, 7 and 9 experiments; in the case of Dieste et al. [23] the number of experiments per aggregation is even smaller.
- Few subjects per study. For example, Dyba et al [1] identify 20 experiments linked to pair programming, where the smallest study contains four experimental subjects per experimental group, the biggest 35 subjects per experimental group, and the average amounts to 13 subjects per experimental group. Ciolkowski [2] identified 21 studies of varied sizes, the smallest containing three experimental subjects per experimental group, the biggest with 45 subjects per experiment, whereas the average was six subjects per experimental group. Dieste et al. [23] identifies 30 experiments, where the smallest contains two experimental subjects per experimental group, the biggest 21, and the average amounts to 11 subjects per experimental group.

Our goal is to detail the conditions under which the Q test is powerful or not powerful enough to positively determine whether or not there is heterogeneity in the meta-analysis of SE experiments. The importance of this goal is that the meta-analysis depends on the Q test both for deciding which statistical model to use (fixed or random effects), as well as for identifying the possible sources of heterogeneity. Therefore, a low power of the Q test would have a major impact on the use of meta-analysis in SE.

This work is part of a wider research agenda where we aim to determine under what conditions meta-analysis techniques are applicable to SE. In Dieste et al. [12] we established the conditions for using the fixed effect methods. This paper sets out to establish when the Q test reliably determines that a set of experiments is homogeneous, meaning that the fixed effects models for meta-analysis methods can be used.

In the future we will analyse the conditions for using the random effects models (which should be used when there is heterogeneity) to complete the characterization of the meta-analysis techniques.

V. RESEARCH METHODOLOGY

Like Kim [11] and Hardy and Thompson [7], we will use a Monte Carlo simulation to study the behaviour of the Q test. This simulation will generate two sets of experiments

with different effect sizes. The Q test will then be applied to these sets to determine their heterogeneity (which, from the construction of the simulation, is known to exist).

Because our simulation is to be run on small samples, we will not directly generate the effect sizes of the primary and secondary treatments. Instead, we will generate instances of experimental subjects that will later be combined in primary and secondary treatment groups (that is, experiment and control groups according to the medical terminology used in many SE experiments). With this, we will assure that our simulation only depends on the probability distribution of the baseline populations and not on the theoretical distribution of Hedges and Olkin's *g* [20].

We assume that the baseline populations of the treatment groups (primary and secondary) have normal distributions. The simulation, including the generation of random numbers (from which the instances of the above-mentioned experimental subjects are generated), was run using a .NET program that we developed.

With respect to the simulation parameters, we will use the same values as in Dieste et al [12] because they are both adequate for experimental SE today and they assure that this study and [12] will be compatible. These values are:

- For the number of subjects per experiment, we consider the range of 4 to 20. 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 to be aggregated in each meta-analysis will range from 2 to 10, as these are typical values of the aggregations in SE, e.g., Ciolkowski [2], Dyba et al [1], Dieste et al. [23].
- The population effect sizes (δ) are the typical values as defined in Cohen [25] (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 [26].

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 Friedrich et al [24] and Dieste et al. [12], 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: $\mu^e = \mu^c + \delta * \sigma$.

The strategy for combining results will be as follows: each meta-analysis will contain *n* heterogeneous experiments (*n=2, 4, 6, 8* and *10*), which will be divided into two subgroups of homogeneous experiments, each containing $n/2$ experiments. Each subgroup will be assigned a different effect size (e.g., the first subgroup may have an effect size of *δ=0.2*, whereas the effect size of the second might be *δ=0.5*). The bigger the difference between the effect sizes, the more heterogeneous the full set of studies will be, and the more feasible it should be to detect heterogeneity.

We will analyse the statistical significance at levels α=0.05 and α=0.10. We will run 10000 simulations for each combination of parameters and then calculate the values of the Q test power. This power will be calculated as the fraction of times the Q test returns a p-value greater than 0.05 (or 0.10) over the total number of generated simulations.

VI. RESULTS

Tables I and II show a summary of the results output in the simulation process. The detailed results are shown in Tables III, IV and V of the Appendix.

TABLE I. Q TEST POWER (*α=0.05*)

Effect difference	Experiments	Subjects	Power
< 1	---	---	[0, 0.6)
1	10	[10, 20)	[0.6, 0.8)
1	10	[20, ∞)	[0.8, 1]

TABLE II. Q TEST POWER (*α=0.10*)

Effect difference	Experiments	Subjects	Power
< 1	--	--	[0, 0.6)
1	8	(0, 10)	[0, 0.6)
1	8	[10, 20)	[0.6, 0.8)
1	10	[4, 8)	[0, 0.6)
1	10	[8, 14)	[0.6, 0.8)
1	10	[14, ∞)	[0.8, 1]

Table I indicates the power of the method when reliability is 95% (*α=0.05*, recommended value for most statistical tests), and Table II indicates the power of the method when reliability is 90% (*α=0.1*,value suggested by Schmidt and Hunter [3] as an alternative for improving the power of the Q method).

Tables I and II contain the following information:

- The "Effect difference" column specifies the difference of effect size between the two subgroups of studies included in the simulation: 0.3, 0.4, 0.6, 0.7 and 1, although, as we will discuss later, we have only achieved acceptable Q test power with effect size differences of around 1.
- The "Experiments" column specifies the minimum number of experiments necessary to achieve the specified power.
- The "Subjects" column specifies the minimum number of subjects per experiment necessary to achieve the specified power.
- Finally, the "Power" column specifies the empirical values of power output for the Q test and for the above-mentioned parameters. To minimize the size of the tables, we have established cut-off points for power at 0.6 and 0.8 (60% and 80%). 80%is the power value typically recommended for statistical tests, whereas Schmidt and Hunter [3] recommended

powers from 60% to 80% for working with small samples.

The summarized results do not include the population variance, because it did not affect the power of Q at any time (for more details, see Tables III, IV and V in the Appendix).

The ideal power (80%) when using the Q test at a confidence level of 95% is only achieved when the effect size difference is 1 (1.2 – 0.2) and at least 10 experiments with 20 experimental subjects each are aggregated. With the same number of experiments we can achieve a power of 60% when the experiments contain at least 14 experimental subjects. For the other cases, power is low and often zero (for more details, see the Appendix).

On the other hand, if we relax the reliability of the Q test to 90%, there is an increase in the power, but this is not strong, as the effect size difference still has to be 1 in all cases for power to be greater than 60%. As regards the number of experimental subjects, however, this reduction in the reliability can achieve powers of 60% with eight experiments with 10 or more experimental subjects.

VII. Discussion

We have corroborated the findings of Kim [11] and Hardy and Thompson [7] concerning the power of the Q test being low in small sample settings. However, the results do more than just corroborate their findings, as, thanks to the thoroughness of the simulation, we were able to establish more clearly the regions in which the Q test is and is not reliable. This way, for example, we observed that the Q test has a power of 80% in settings where there are 10 experiments with 20 subjects per experimental group and effect size differences of 1, a possibility that [11] categorically ruled out.

While it is true that we have detected regions where the Q test is powerful (≥80%) or almost powerful (60%-80%), the requirements in terms of experiments, subjects and effect differences are in fact very demanding. None of the meta-analyses run to date in SE [1] [2] simultaneously meet all the requirements.

In practice, the low power of the Q test implies that is not possible to rely on the insignificant results in the current context of SE; that is, we have to suppose that all the experiments included in a meta-analysis are potentially heterogeneous, irrespective of their p-value. Note that this is not the case for the significant results, i.e., the experiments are almost certain to really be heterogeneous (at the respective level) when the Q test detects heterogeneity.

The supposition that the whole set of experiments included in a meta-analysis is heterogeneous automatically rules out the use of the fixed effects models, and the random effects models should be applied instead, as Hardy and Thompson [7] recommend in the field of medicine. Fixed effects models assume that the experiments to be combined are homogeneous and they are rather imprecise in the presence of heterogeneity [3]. As homogeneity cannot be reliably ensured, the imprecision of the fixed effects model is unacceptable.

The use of a random effects model assumes that the heterogeneity is caused by chance or by changes in the experimental methodology (design, response variables, etc.). However, moderator variables are another cause of heterogeneity. In this case, random effects models should not be used, and the sample has to be decomposed into subgroups (assuming that the moderator variable is categorical, which is a reasonable premise in SE). One frequently used decomposition method is to identify and exclude heterogeneous studies from the original group and then re-estimate the global effect size as suggested by Hedges and Olkin [20]. Now, the Q test should not be used as a guide for decomposing a set of heterogeneous experiments into subgroups, and, if it is, this should be done with extreme care. The resulting subgroups contain fewer experiments, and this reduces the power of the Q test. The output results are therefore more likely to be insignificant. This means that any decomposition into subgroups could be accepted as valid, irrespective of whether these subgroups do or do not correspond to a reliable moderator variable.

VIII. Limitations

In the simulation, we have used experiments of equal size, that is, all the experiments included in the meta-analyses had the same number of experimental subjects. We believe that this limitation poses no threat at all to the validity of the findings, as what little evidence there is in the literature [7] suggests that the power of the Q test drops even further when the experiments to be aggregated differ in size. Obviously, this point needs to be further investigated, but we have reason to believe that the power values specified in this paper are upper bounds, and the power of the Q test is likely to be even smaller in real meta-analyses. This offers further support, if possible, for the recommendations made in the discussion section.

IX. Conclusions

From the simulation run we have found that, although it is important to detect when a set of experiments is heterogeneous, there is no way of doing so with any guarantee in the current setting of SE experimentation. In fact, the requirements for the Q test to be acceptably powerful are so demanding, especially in terms of effect differences, that we could gauge the heterogeneity of a set of experiments much more readily visually by inspecting a forest plot than from statistical tests.

As the Q test is not very powerful, we have to be extremely cautious about its use. A significant result means that a set of experiments is heterogeneous, but a negative result cannot guarantee that it is not. In this scenario, caution dictates that meta-analyses should be run using only random effects models.

Although the recommendation to use random effects models is well founded, it is at the same time debatable. Random effects models differ from fixed effects models merely in that they include between-study variance (τ^2), but the estimation of between-study variance is not very precise in contexts where there are few experiments [6]. In other words, it is not reasonable to use random effects models when there is a risk of the τ^2 variance being incorrect. We

intend to study the precision of τ^2 estimation in future research to gather more evidence than an authoritative reference to the literature about this question.

Finally, let us add that we are not satisfied with the empirical power tables generated in this research. In all truth, we expected to get higher powers for the Q test. Note that, at a significance level of 90% (α=*0.10*), we need a total of 160 subjects (10 experiments with at least eight subjects per experimental group) and an effect difference of 1 to achieve a quasi-acceptable power (60%-80%). While it is practicable, we think, to amass this number of experiments and subjects per experiment in SE, it is less reasonable to suppose that it will be possible to design technologies that produce such effect differences. We aim to extend our simulation to a greater number of experiments and subjects per experiment to find out if it is possible to identify regions where the Q test is powerful with smaller effect differences. To do this we will be obliged, in future studies, to extend the tables output in [12] (in order to assure that the research is comparable), as well as to extend the range of the Monte Carlo simulation parameters.

ANNEX A. SIMULATION RESULTS

Tables III, IV and V show the result of the simulation process. Column 1 indicates the number of subjects in each experiment; column 2 indicates the total number of experiments in the aggregation process; columns 3 to 8 indicate the effect sizes for comparison. These columns are divided into three subcolumns, which indicate the percentage of times that the simulation process output $p \leq 0.05$ (meaning that the Q test managed to detect heterogeneity with a reliability greater than or equal to 95%), a $0.05 < p \leq 0.1$ (meaning that the Q test managed to detect heterogeneity with a reliability of from 90 to 95%) and $p > 0.1$ (meaning that the Q test did not detect heterogeneity).

TABLE III. RELIABILITY OF P ASSOCIATED WITH Q FOR LOW VARIANCE SETTINGS (10% W.R.T. THE MEAN)

Subjects	Exp.	Effect size 0.2 vs. 0.5			Effect size 0.2 vs. 0.8			Effect size 0.2 vs. 1.2			Effect size 0.5 vs. 0.8			Effect size 0.5vs.1.2			Effect size 0.8vs. 1.2		
		p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1
4	2	1	2	97	1	2	97	1	10	89	1	1	98	2	3	94	3	2	95
8		0	0	100	0	1	99	1	3	96	0	0	100	0	2	98	0	1	99
10		0	0	100	0	0	100	3	3	94	0	0	100	0	2	98	0	0	100
14		0	0	100	0	1	99	0	0	100	0	0	100	1	0	99	0	0	100
20		0	0	100	0	0	100	3	2	95	0	0	100	0	0	100	0	0	100
4	4	4	6	90	11	12	77	19	4	77	4	5	91	17	9	74	6	8	86
8		2	1	97	7	5	88	12	11	77	3	1	96	7	7	85	1	3	96
10		1	4	94	5	4	91	12	12	76	2	2	96	5	2	93	5	2	92
14		2	2	96	7	8	85	18	7	75	2	3	95	9	7	84	8	1	91
20		2	3	95	5	4	91	20	11	69	2	3	95	11	5	84	5	1	94
4	6	0	0	100	9	8	83	24	18	58	0	3	97	11	15	74	0	3	97
8		2	2	96	4	6	89	16	14	69	2	3	95	12	5	83	2	10	88
10		0	2	98	4	3	93	19	10	71	0	2	98	8	9	83	0	3	97
14		2	0	98	7	3	90	19	19	62	2	1	97	10	3	87	1	4	96
20		3	2	95	11	5	84	30	14	56	2	1	97	12	4	84	4	1	95
4	8	6	7	87	14	12	74	18	20	62	6	11	83	12	13	75	5	1	94
8		2	5	93	7	8	85	39	15	46	4	5	91	11	11	78	6	3	91
10		3	3	94	6	11	84	46	18	36	4	3	93	14	9	77	4	3	94
14		1	5	94	10	7	83	37	21	42	2	4	94	8	8	84	1	6	93
20		0	3	97	14	5	81	51	17	32	5	2	93	12	9	79	1	5	94
4	10	6	7	88	24	18	58	34	18	49	0	0	100	16	8	76	13	13	74
8		4	4	93	10	12	78	50	13	38	7	6	87	16	19	64	7	6	86
10		4	2	94	7	7	86	65	10	24	7	1	92	12	16	72	6	4	91
14		6	4	90	10	8	82	71	19	10	6	3	91	19	14	67	10	2	88
20		6	6	88	12	0	88	100	0	0	6	6	88	12	9	79	12	0	88

TABLE IV. RELIABILITY OF P ASSOCIATED WITH Q FOR MEDIUM VARIANCE SETTINGS (40% W.R.T. THE MEAN)

Subjects	Exp.	Effect size 0.2 vs. 0.5			Effect size 0.2 vs. 0.8			Effect size 0.2 vs. 1.2			Effect size 0.5 vs. 0.8			Effect size 0.5vs.1.2			Effect size 0.8vs. 1.2		
		p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1	p≤ 0.05	0.1≤ p >0.05	p > 0.1
4	2	1	2	97	1	2	97	2	10	88	1	1	98	2	3	95	3	2	95
8		0	0	100	0	1	99	1	3	96	0	0	100	0	2	98	0	1	99
10		0	0	100	0	0	100	3	4	93	0	0	100	0	2	98	0	0	100
14		0	0	100	0	1	99	0	0	100	0	0	100	1	0	99	0	0	100
20		0	0	100	0	0	100	3	2	95	0	0	100	0	0	100	0	0	100
4	4	4	6	90	11	12	77	20	5	75	4	5	90	17	9	74	6	8	86
8		2	1	97	7	6	87	12	12	76	3	1	96	8	8	84	1	3	96
10		1	4	95	5	4	91	14	10	76	2	2	96	6	5	89	5	3	92
14		2	2	96	7	8	85	19	9	73	3	2	95	9	7	84	8	2	90
20		2	3	95	5	4	91	22	10	68	2	3	95	11	5	84	5	1	94
4	6	0	0	100	9	8	83	26	18	55	0	4	96	13	14	72	0	3	97
8		2	2	96	6	5	89	18	12	70	2	3	95	12	7	81	1	10	88
10		0	2	98	4	4	92	23	10	67	0	2	98	8	9	83	0	3	97
14		2	0	98	7	3	90	23	18	58	2	1	97	11	2	87	1	3	96
20		3	2	95	12	4	84	34	11	55	2	1	97	13	5	82	4	1	95
4	8	6	7	87	14	11	75	20	23	57	6	11	83	12	13	75	5	2	93
8		2	5	93	6	8	86	42	14	43	4	6	90	12	12	76	6	3	91
10		3	3	94	6	11	83	46	19	35	4	4	92	13	11	76	4	4	92
14		1	5	94	10	8	83	42	19	38	2	5	93	9	8	83	1	6	93
20		0	3	97	15	5	80	56	18	26	6	3	91	14	9	78	2	5	93
4	10	6	7	87	26	17	57	35	17	48	0	0	100	16	9	75	14	13	73
8		4	4	92	10	13	77	51	13	36	8	5	87	18	18	64	8	7	85
10		4	2	94	9	8	83	70	10	20	7	1	92	14	19	67	6	4	91
14		6	4	90	11	6	82	79	12	9	6	3	91	19	17	64	10	2	88
20		6	6	88	12	0	88	100	0	0	9	3	88	12	18	70	12	0	88

TABLE V. RELIABILITY OF P ASSOCIATED WITH Q FOR HIGH VARIANCE SETTINGS (70% W.R.T. THE MEAN)

Subjects	Exp.	Effect size 0.2 vs. 0.5			Effect size 0.2 vs. 0.8			Effect size 0.2 vs. 1.2			Effect size 0.5 vs. 0.8			Effect size 0.5vs.1.2			Effect size 0.8vs. 1.2		
		p≤ 0.05	0.1≤ p > 0.05	p> 0.1	p≤ 0.05	0.1≤ p > 0.05	p > 0.1	p≤ 0.05	0.1≤ p > 0.05	p > 0.1	p≤ 0.05	0.1≤ p > 0.05	p > 0.1	p≤ 0.05	0.1≤ p > 0.05	p > 0.1	p≤ 0.05	0.1≤ p > 0.05	p > 0.1
4	2	1	2	97	1	2	97	1	11	88	1	1	98	2	3	95	3	2	95
8		0	0	100	0	1	99	1	3	96	0	0	100	0	2	98	0	1	99
10		0	0	100	0	0	100	3	3	94	0	0	100	0	2	98	0	0	100
14		0	0	100	0	1	99	0	0	100	0	0	100	1	0	99	0	0	100
20		0	0	100	0	0	100	3	2	95	0	0	100	0	0	100	0	0	100
4	4	4	6	91	11	12	77	19	5	76	4	5	91	17	9	74	6	8	86
8		2	1	97	7	6	87	11	11	77	3	1	96	8	8	84	1	3	96
10		1	4	95	5	4	91	14	12	75	2	2	96	5	3	92	5	2	93
14		2	2	96	7	8	85	18	10	72	2	3	95	9	7	84	8	2	90
20		2	3	95	5	4	91	21	11	68	2	3	95	11	5	84	5	1	94
4	6	0	0	100	9	8	84	25	20	56	0	3	97	12	15	72	0	3	97
8		2	2	96	5	6	89	17	13	70	2	3	95	13	5	82	2	10	88
10		0	2	98	4	3	93	22	9	69	0	2	98	8	8	83	0	3	97
14		2	0	98	7	3	90	19	18	63	2	1	97	11	2	87	1	3	96
20		3	2	95	11	5	84	33	12	55	2	1	97	12	4	84	4	1	95
4	8	6	7	87	14	12	74	19	21	60	6	11	83	12	13	75	5	2	93
8		2	5	93	6	7	87	41	16	43	4	5	91	12	10	78	6	3	91
10		3	3	94	6	11	83	46	19	35	4	3	93	13	9	78	4	4	92
14		1	5	93	10	7	83	41	20	39	2	4	94	9	8	84	1	6	93
20		0	3	97	15	5	81	54	17	29	6	2	92	13	10	77	1	5	94
4	10	5	7	88	24	18	58	35	18	47	0	0	100	16	8	76	13	14	73
8		4	4	92	10	13	78	51	13	36	7	6	87	17	19	64	8	6	86
10		4	2	94	8	7	85	67	10	22	7	1	92	14	18	68	6	4	90
14		6	4	90	10	7	83	76	15	9	6	3	91	19	14	67	10	2	88
20		6	6	88	12	0	88	100	0	0	6	6	88	12	15	73	12	0	88

ACKNOWLEDGMENTS

This research has been partially funded by the grants TIN2008-00555 and HD2008-0048 of the Spanish Ministry of Science and Innovation.

REFERENCES

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

[2] Ciolkowski, M., (2009), "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, pp. 133-144.

[3] Schmidt, F. and Hunter, J. (2003) "Handbook of Psychology, Research Methods in Psychology", Chapter 21, "Meta-Analisis", Schinka, J., Velicer, W., Weiner, I. Editors, Volume 2

[4] DerSimonian, R. and Laird, N. (1986) "Meta-analysis in clinical trials", Controlled Clinical Trials, Vol. 7, no. 3, pp. 177-188

[5] Cochrane colaboration (2011), "Open learning material", http://www.cochrane-net.org/openlearning/html/mod0.htm

[6] Borenstein, M., Hedges, L., Rothstein, H., (2007) "Meta-Analysis Fixed Effect vs. random effect", http://www.meta-analysis.com/downloads/Meta%20Analysis%20Fixed%20vs%20Random%20effects.pdf.

[7] Hardy, R. and Thompson, S., (1998) "Detecting and Describing Heterogeneity in Meta-Analysis", Statistics in Medicine, vol. 17, pp. 841-856

[8] Song, F., Sheldon, T., Sutton, A., Abrams, K., Jones, D., (2001) "Methods for Exploring Heterogeneity in Meta-Analysis", Evaluation and The Health professions, vol. 24 no. 2, pp. 126-151.

[9] Kitchenham, B., (2004), "Procedures for performing systematic reviews", Keele University; TR/SE-0401, Keele University Technical Report.

[10] Ioannidis,J., Patsopoulos, N., Evangelou, E.(2007) "Uncertainty in heterogeneity estimates in meta-analyses", BMJ 335:914, doi: 10.1136/bmj.39343.408449.80.

[11] Kim, J., (2000) "An Empirical Study of the Effect of Pooling Effect Sizes on Hedges's Homogeneity Test", Annual meeting of the American Educational Reseerch Association. New Orleans.

[12] Dieste, O., Fernández, E., García, R., Juristo, N. (2011) "Comparative analysis of meta-analysis methods: when to use which?" 6th EASE Durham (UK).

[13] Hunter, J. (2001), "The desperate need for replications", Journal of Consumer Research, vol. 28, pp. 149-158.

[14] Cochran, W. G. (1954), "The combination of estimates from different experiments", Biometrics, vol. 10, pp. 101–129.

[15] Lipsitz S., Dear K., Laird N., Molenberghs, G. (1998), "Tests for homogeneity of the risk difference when data are sparse", Biometrics, vol. 54, pp. 148-160.

[16] Strain, D. and Lee, J. (1984) "Variance Component Testing in the Longitudinal Mixed Effects Model", Biometrics, vol. 50, pp. 1171-1177.

[17] Jones, M., O'Gorman, T., Lemke, J., Woolson, R. (1989), "A Monte Carlo Investigation of Homogeneity Tests of the Odds Ratio under Various Sample Size Configurations", Biometrics, vol. 45, no. 1, pp. 171-181.

[18] Gavaghan D., Moore A., McQay H. (2000) "An evaluation of homogeneity tests in meta-analysis in pain using simulations of patient data", Pain, vol. 85, pp. 415-424.

[19] Higgins J., Thompson S. (2002) "Quantifying heterogeneity in a meta-analysis", Statistics in Medicine, vol. 21, pp. 1539-1558

[20] Hedges, L. and Olkin, I. (1985) "Statistical methods for meta-analysis", Academic Press.

[21] Liang, K. and Self, S. (1985) "Tests for Homogeneity of Odds Ratio When the Data are Sparse", Biometrika, vol. 72, No. 2, pp. 353-358.

[22] Takkouche B., Cadarso-Suarez C. & Spiegelman D. (1999) "Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis", Am. J. Epidemiol, 150, pp. 206–215.

[23] Dieste, O., Juristo, N. (2011) "Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques", IEEE Transactions on Software Engineering, vol 37, no. 2, pp. 283-304.

[24] Friedrich, J., Adhikari, N., Beyene, J., (2008) "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.

[25] Cohen, J., (1977) "Statistical power analysis for the behavioral sciences", Hillsdale, NJ: Lawrence Erlbaum Associates, 2nd Ed.

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