Applying Meta-Analytical Procedures to Software Engineering Experiments

James Miller
Dept. Computer Science, University of Strathclyde,
Livingstone Tower, Richmond Street, Glasgow G1 1XH, Scotland

EFOCS-30-98

Abstract

Deriving reliable empirical results from a single experiment is an unlikely event. Hence to progress multiple experiments must be undertaken per hypothesis and the subsequent results effectively combined to produce a single reliable conclusion. Since results are quantitative in nature, a quantitative conclusion would be the optimal solution. Other disciplines use meta-analytic techniques to achieve this result. The treatise of this paper is: can meta-analysis be successfully applied to current Software Engineering experiments? The question is investigated by examining a series of experiments, which themselves investigate — which defect detection technique is best? Applying meta-analysis techniques to the Software Engineering data is relatively straightforward, but unfortunately the results are highly unstable, as the meta-analysis shows that the results are highly disparate and don't lead to a single reliable conclusion. The reason for this deficiency is the excessive variation within various components of the experiments. The paper outlines various ideas from other disciplines for controlling this variation and describes a number of recommendations for controlling and reporting empirical work to advance the discipline towards a position where meta-analysis can be profitably employed.

1 Introduction

Single experiments or studies in Software Engineering rarely provide definitive answers. Hence if scientific investigation is to progress in Software Engineering, it must be through the discovery of trends and ideas derived from a large number of studies. This has given rise to calls in many papers for "further work on this topic" or "replications of this study are required" while these are necessary, they are not the end of the story. For success the material from these studies must be accumulated, summarised and a clear description of the current 'state of the art' must be produced. Traditional reviews suffer from their dependence on subjective judgements, preferences, and biases of the reviewers and from ill-defined procedures for undertaking the process.

Hence many disciplines have turned to more quantitative methods of producing such essential summaries, and in particular have adopted a meta-analytic approach to producing unbiased statements. The main purpose of this paper is to pose the question: "Is Software Engineering ready for meta-analysis?". This question will be answered by analysing a well-known set of experiments on defect detection.

2 Meta-Analysis

Meta-analysis has existed for a long time, the first recorded use was by Pearson[30] in 1904. Despite its application to various disciplines for nearly a century, its application within Software Engineering[1] remains unexamined. Having said this, several authors have produced work with some parallels to the meta-analytic approach:

- Porter and Johnson[33] describe using "meta-analysis" in their work. They use a process they call "reconciliation" rather than traditional meta-analysis.
- Hu[21] describes the use of the Davidson and MacKinnon method[10] to analyse different software cost estimation models.
- Banker and Kemerer[1] reconcile two opposing views regarding the presence of economies or diseconomies of scale in software development by informal vote counting[15]. This technique is not commonly used because of its independence from the size (in subjects) of the studies.
- Also Brooks[4] gives a short discussion about the possibilities for applying meta-analysis to Software Engineering projects.

A traditional meta-analysis starts with an exhaustive search of the literature to find all the articles describing empirical evaluations of the concept under investigation. In conjunction with this the researcher should only identify the relevant variables. The unit of analysis in a meta-analysis should be the impact of Variable X on Variable Y. It should be noted that this is a tightly focussed investigation, the concept of exploring multi-variables and their inter-relationships via a single meta-analysis is a topic of hot debate in many experimental disciplines. Many experts express the viewpoint that a less focussed meta-analysis runs too great a risk of producing results, which suffer from inflated error rates and increased conceptual confusion. For example, Kulik[25], Rosenthal[36] and many others recommend performing separate meta-analyses on each type of dependent variable; Gilbert *et al.*[12] go further and suggest that the number of results, from a single study, be restricted to two when subsequently conducting a meta-analysis. Since this topic is the scene of continuing open-debate about which forms are valid and which are invalid, the author plans to adopt a conservative approach in examining the case study, to ensure that the study conforms to the model of currently accepted practice; see Smith and Glass[42] and Harris

[1] The author has recently become aware of an unpublished manuscript by Pickard *et al.*[32], which discusses meta-analysis and its potential use in Software Engineering case-studies.

and Rosenthal[16] for two differing viewpoints on the validity of various types of meta-analysis involving multiple variables and analyses.

Therefore, the study shall only focus on those variables that relate to our specific question or hypothesis. Even given these restrictions sometimes it is necessary to reformulate the answers of some experiments to ensure that we are measuring the same effect — Porter and Johnson[33] provide a detailed account of a possible procedure, and hence this paper will not repeat that discussion.

Once the researcher has completed these processes, they are ready to start the meta-analysis process, if the analyst has more than two studies, as is the case here, six options are open to them:

- *Comparing Studies - Diffuse Tests:* used to determine if two or more studies produce significantly different results, but does not reveal if the difference is based upon any systemic cause.
  - **Significance Testing:** based upon the recorded p values from the studies. Only used when information is not available for evaluation of effect sizes.
  - **Effect-Size Estimation:** based upon the recorded values of the inferential statistics (F or t values, for example) along with the associated degrees of freedom[2]. Effect sizes are estimated from these statistics. Strongly preferred over significance testing.
- *Comparing Studies - Focused Tests:* used to determine if two or more studies produce significantly different results based upon a theoretically predictable or systemically meaningful way.
  - **Significance Testing:** similar to above.
  - **Effect-Size Estimation:** similar to above.
- *Combining Studies:* used when you want to determine the potency of a variable across studies.
  - **Significance Testing:** can be used after comparing studies to arrive at an overall estimate of the probability of obtaining the p values under the null hypothesis (i.e. no casual relationship).
  - **Effect-Size Estimation:** can be used after comparing studies to evaluate the average impact across studies of an independent variable on the dependent variable.

Once having decided upon their approach, (normally at least one comparative study to check the homogeneity of the quantities under review, followed by a combinational calculation) the meta-analyst examines their data for:

- Errors - As pointed out by Rosenthal and Rubin[37, 38], the meta-analyst should expect errors in any large set of data. These range from the simple arithmetic miscalculation to the misrecording of the raw data. For some

[2] This assumes that raw data is not available.

errors no correction is possible (often because they are undetectable), but often they are easily corrected, by recalculation, or the data can be reformulated to account for the problem. What is important, is for the meta-analyst to undertake this duty seriously and diligently.

- Quality - One of the major criticisms of meta-analysis is that poor studies are summarised as well as good studies. Hence, once all the studies have been found the meta-analyst must make a decision about the quality of each study, and hence, assign them a 'weight' for the meta-analysis. Obviously this process can exhibit considerable bias, and the recommended practice is to organise for an independent panel of experts to adjudicate on the quality of the studies under review. See Walker and Lev[43] for a discussion on the reliability of judgements of quality by independent experts.

Finally the meta-analyst is ready to attempt the series of calculations that they have decided upon. Unfortunately multiple sets of alternative calculations are available to the meta-analyst, the chosen approach is dependent upon a number of factors, but is most heavily influenced by the choice of effect size metric. Effect size metrics can be derived by two different mechanisms: differences between means or proportions, and product moment correlations and derived functions. Rosenthal[36] provides a description of most common effect size measures and supplies many useful formulae for converting between them. It has been shown[19] that the performance of most effect size measures is equivalent for experiments involving more than ten subjects. Note: although effect size measures tend to be parametric, nonparametric forms do exist, see Wolf[44] for some examples.

A full discussion of all the possible permutations of the approaches to calculating the final meta-analytic result is outside the scope of this paper, hence only one approach will be illustrated for the chosen case study. Again see Rosenthal[36] (or Hedges[19]) for alternative approaches. Common wisdom is that the choice of technique is relatively unimportant, as most techniques produce similar results; in fact the meta-analyst is encouraged to use several techniques to check the results' sensitivity to the choice of the analysis technique. Sensitivity analysis of the results against the various options open to the meta-analyst is an important part of the process.

3 Case Study Overview: Defect Detection Experiments

This topic has seen a large number of experiments based upon similar hypothesis. The basic premise investigated by these experiments is "Which (if any) defect detection technique is most effective at discovering faults?". On reviewing the literature five independent studies were found, which have investigated this concept:

- The first experiment was performed by Hetzel[20] in 1976. He compared code reading, functional testing and a combination of functional and struc-

Author(s)	Code Reading (R)	Functional (F)	Structural (S)
Hetzel	37.3	47.7	46.7
Myers	38.0	30.0	36.0
Basili ...	54.1	54.6	41.2
Kamsties ...	43.9	44.2	35.0
Roper ...	43.4	55.1	57.9

Table 1: Average Percentage of defects found in each experiment

tural testing. His results implied that the two testing techniques were equally effective, with code reading seeming to be inferior.

- The next study was undertaken by Myers[28]; he compared 'informal' versions of code reading, functional testing and structural testing; this time within a team-oriented setting. He reported that all the techniques were of similar effectiveness, but that effectiveness seems to be highly dependent on defect 'type'. He further demonstrated that the techniques were complementary with regard to finding different sets of defects.

- The third experiment was by Basili and Selby[2]; they performed three experiments comparing functional testing using equivalence partitioning and boundary value analysis; structural testing using 100% statement coverage as its stopping criteria and code reading by stepwise abstraction. They reported 'weak' evidence that code reading was more effective. Again they suggested that a combinational approach of defect finding strategies seemed to offer advantages.

- More recently Kamsties and Lott[24] replicated the Basili and Selby experiment twice. They reported no significant differences between the three techniques.

- Finally, Roper *et al.*[35] replicated the Kamsties and Lott experiment. Again, they reported no significant differences, but again, reported that the different techniques were complimentary.

Table 1 summarises the results from the various experiments. Examining this table informally yields little understanding of the relative merits of the various techniques, hopefully meta-analysis will present a clearer picture.

4 Analysis of Defect Detection Experiments

The previous section describes all the evaluations comparing these three techniques, and hence completes the review of the literature.

4.1 Evaluation of the Quality of the Studies

At this point in time, it is unlikely that the field is sufficiently mature and sufficiently populated with relevant experts to accommodate the independent assessment, from a meta-analytical viewpoint, of these studies. Hence, regrettably, the author has undertaken this process. In an attempt to limit the biases this will undoubtedly introduce, the analysis was restricted to points, which are (hopefully) non-controversial:

- In the study undertaken by Hetzel, the structured testing component has been replaced by a functional and structural testing component, hence we have given this study a weighting of zero, i.e. excluded it from the meta-analysis.
- Kamsties and Lott state in their technical report[23] that the design of the initial experiment (but not the subsequent replication) was 'flawed'. Hence, again, we have given the initial study a weighting of zero, but we have given the subsequent replication a full weighting of one.

Other points could have been considered (such as the limited guidance provided by Myers to the subjects on applying the techniques; other experiments show that there is an interaction between the techniques and the programs used), but there is no sound basis for deriving a quantitative weighting, and hence a conservative approach is adopted - when in doubt, do nothing! Meta-analysis is an aid to thought, not a substitute, it can never produce exact numerical statements, which cannot be questioned or debated.

Alternatively, if we are still concerned after the analysis, we could run a sensitivity analysis to estimate the impact of the data from the Myers experiment, for example, on the overall results. Petitti[31] discusses the use of sensitivity analysis in meta-analysis on epidemiological data.

4.2 Statistical Methods and Decisions

In line with the 'standard' conservative approach to meta-analysis, it was decided to analyse the case study, as three independent meta-analyses. This protects against inflating the error terms, causing conceptual confusion and, on average, lowers the probability of failing (due to increased cohesiveness) the test of homogeneity (comparative studies) between the studies.

The decision about which effect size metric to choose, is relatively unimportant. The normal choice is to select the one which fits the data best. Since our data either supplies standard deviations and means, or supplies the raw data allowing their calculation, we will use Cohen's d[9]. To illustrate the unimportance of the point, Rosenthal[36] provides simple equations to transform d into r (Pearson's product moment correlation) and *vice versa*. The effect size estimates for the four remaining studies are given in Table 2. The Basili and Selby effect size estimate is derived from the combination of the three estimates from sub-studies, all the sub-studies have independent subject bases[3], but dependent designs, materials, etc. When combining, or averaging, effect sizes a researcher has three options: a traditional average; an average weighted by size (number of subjects) — larger studies are more likely to be more reliable; an average

[3] This also implies that an average sample size measure has also been constructed.

	Effect Size		
Author(s)	R Vs F	F Vs S	R Vs S
Myers	0.31	-0.40	0.15
Basili ...	0.02	0.67	0.60
Kamsties ...	-0.07	0.33	0.30
Roper ...	-1.25	-0.13	-1.22

Table 2: Effect Sizes in each analysis — a negative effect size indicates that the second technique in the pair possessed a greater mean than the first technique

	Corrected for bias		
Author(s)	R Vs F	F Vs S	R Vs S
Myers	0.30	-0.39	0.15
Basili ...	0.02	0.66	0.59
Kamsties ...	-0.07	0.32	0.29
Roper ...	-1.24	-0.13	-1.21

Table 3: Effect Sizes in each analysis — a negative effect size indicates that the second technique in the pair possessed a greater mean than the first technique

weighted by variance: studies with smaller variance are likely to be more precise. Since this study is composed of three internal replications, and most of the experimental design parameters except the sample size are constant, weighting the estimates by size has been chosen as the most appropriate option. This follows the procedure recommended by Kulik[25] and others.

Hedges[18] has derived the sampling distribution for d and has showed that it is a biased estimate of effect size. In a further paper[17] he showed that a weighted estimator of effect size estimate can be constructed, which is asymptotically efficient and accurate when the harmonic mean of the sample size is greater than 10 and the effect sizes are less than about 1.5. Fortunately our data conforms to these criteria and we can convert our data via a simple procedure, which is detailed in Rosenthal[39]. Table 3 shows our corrected effect size estimates. This adjustment can be thought of as counteracting the finite sample size of the experiments; further adjustments exist for countering the restriction (or enhancement) of the range of variables involved[22] or the precision of the statistical tests applied in the individual studies[13]; it was decided that neither of these additional adjustments was required for the above studies. Further there is not universal acceptance that these further 'corrections' always lead to superior estimates[36].

Comparing Studies: Diffuse Tests

Here we want to assess the statistical heterogeneity of our unbiased effect size estimates. The ability to demonstrate the homogeneity, i.e. estimates only differ due to subject variability, of the set of effect size measures is important as it allows us to assume that we are measuring only one effect and that the studies are representative samples of the general population. In meta-analysis this is often known as a fixed-effects model or system. If we are unable to demonstrate this then the system has become a random-effects model and inference is based upon the assumption that the studies are a random sample of the population. Although random-effects models are tolerant to variation, they are still not effective in highly heterogeneous situations. Obviously the statistical methods used to combine studies when fixed-effects are assumed differ from the methods used when random effects are used.

Unfortunately no empirical basis for preferring either model exists. But many experts agree that the choice of model is secondary to the examination of the reasons for the lack of homogeneity. If the studies are homogeneous then it has been shown that both models produce identical results[3]. Unfortunately Software Engineering is currently in a weak position to 'prove' that its studies are homogeneous, even if our studies pass the test, can we trust this result? Given the small number of studies we can only produce a homogeneity test with relatively low statistical power, and hence it has an increased risk of inappropriately predicting that the studies are homogeneous.

Given the uncertainty of the above position, it is proposed to proceed by using both models. If heterogeneity is discovered — stop, and investigate the causes, if possible. If the studies seem relatively homogeneous again use both models to derive estimates for the combination of the results from the studies. Random-effects models are generally conservative compared to the fixed-effects models, and its recommended use is to provide conservative estimates rather than tackle heterogeneous situations. It is believed that it is important not to use meta-analysis as a statistical method or test, but to consider it as a multicomponent approach to understanding information for independent sources, hence the seemingly insoluble problems of choosing between the models can to a certain extent be ignored.

For the fixed-effect and random-effect models we are going to use models described in [19][4]. Firstly, the test for homogeneity is equivalent to testing if the variance of the effect sizes is zero. We can test for the homogeneity of effect sizes in an analogous manner to testing for the homogeneity of statistical tests, this generates a χ^2 distribution (Q) with $K - 1$ $d.f.$, where K is the number of studies. Additionally the random-effect model partitions the variance estimate into components representing the sampling error $(S.E.)$ and variations in the population parameters. This further illustrates the degree of homogeneity within the data set. Obviously if one hundred percent of the observed variance is explained by the sampling error, then the data is completely homogeneous; but for figures less than one hundred, then the residual variance is due to systematic factors, which should be explored before combining results from the various studies. Table 4 describes the results of the homogeneity test for our three sets of experiments.

As can be clearly seen, all three meta-analyses have serious homogeneity

[4] details of the fixed-effect model can be found in Chapter 6; and details for the random-effect model in Chapter 9.

Type of test	Statistic	R Vs F	F Vs S	R Vs S
Fixed	Q	18.72	7.30	30.13
	p	<<0.01	0.06	<<0.01
Random	Q	18.60	7.26	29.90
	p	<<0.01	0.06	<<0.01
	$S.E.(\%)$	23.18	41.29	16.67

Table 4: Homogeneity tests — all tests have 3 d.f.

problems, and further exploration is required: comparative studies - focussed.

Comparative Studies - Focussed

Unfortunately the distribution of effect sizes in the case studies are rather heterogeneous. The purpose of focussed testing is to attempt to derive an alternative hypothesis where one (or more) characteristics of the study significantly contribute to the observed variation. If we can establish such a hypothesis we can subdivide the studies, based upon this variable, and run a number of independent meta-analyses based upon our new hypothesis. If homogeneous subsets with different effect sizes emerge then the hypothesised study characteristics are established as moderators, and the new hypothesis has been established.

Often the researcher may have little idea as to which characteristic may be responsible for the remaining variation. Here an inductive approach may help, and an informal inspection of the rank-ordered effect sizes to obtain an initial impression of possible groupings is often a good starting place. In order to make a sound judgement, the adoption of a multivariate statistics approach is recommended: such as factor analysis[40], multidimensional scaling[11] or cluster analysis[34]. Alternatively artificial intelligence techniques, such as inductive analysis[5], can be used, especially in situations where nominal or ordinal scale data is present.

Of course, given the number of studies at hand the use of a formal technique is not required, we can simply look at the data to try to discover obvious patterns:

- Roper *et al*: results when using reading techniques are comparatively lower (against both testing techniques) than within the other studies.

- Basili and Selby: results when using structural testing techniques are comparatively lower (against both of the other techniques) than within the other studies.

Can we find evidence in the studies to explain these potential differences? Considering the first potential partition, in their article Roper *et al.* statistically show that the performance of the detection techniques is dependent on the 'type' of faults encountered. This type of interaction has also been noted by several of the other studies, which report that the techniques tend to find different types of defects, and hence benefit can be derived by combining the techniques. Although some evidence exists to an important effect, further exploration would be pointless, as we are not in a position to transform the outcome into a moderator variable, due to the lack of definition of defect type. This inability to transform moderator explanations into variables is a major problem within Software Engineering, as we will be left in a position of being unable to partition the population into homogeneous sub-populations. This deficiency is a major obstacle to the path of the field producing reliable generalisable results, and should be seen as a top priority for the discipline.

One point worth noting before leaving this topic is the relationship between two of the studies. Roper *et al.* re-uses the experimental materials from Kamsisics and Lott; although from a simple replication point of view, this seems attractive, from a meta-analytical point the practice is undesirable, as it creates strong correlations between the two studies (with regard to variables involving the material) and hence a potentially serious threat to the independence of the studies. This point also creates a problem with respect to an earlier decision. Are the Kamastics and Lott, and the Roper *et al.* experiments significantly more independent than the three sub-experiments of Basili and Selby. The correct viewpoint at this juncture, is to admit that we don't know and to produce a sensitivity analysis by calculating all possible permutations. With homogeneous data, the results should converge and the matter can be discarded; with heterogeneous data the results will tend to be unstable, and hence should decrease our confidence in the final combined estimate.

Alternatively we could decide to discard or reduce the reliability rating of Roper *et al.* to reflect our concerns with this issue. The 'contribution' of a study to the calculations is attenuated by the square of its reliability estimation in most meta-analytic models. But do we have enough proof to justify this decision? Certainly Roper *et al.* show a significant interaction between defect detection technique and program. Here program can be considered as the list of defects and their context. But since Roper *et al.* re-used the materials from Kamastics and Lott we would except the same situation to occur. Unfortunately Kamastics and Lott fail to provide any results about the potential interaction. A retrospective examination of their data was unable to produce a reliable statement about the interaction. The Kamatics and Lott experiment suffered from a large subject attrition rate, leaving the blocked design highly skewed, in fact some of the blocks have no subjects, this means that no reliable estimate can be found. What about the other two experiments: Basili and Selby show that one of their three sub-studies had a significant interaction, but unfortunately don't provide any details about the other two phases, both of which use a subset of the materials from the phase showing the interaction; the Myers experiment only has one program, but even here he shows that significant differences, from the average behaviour, exist at the individual defect level between the techniques. Hence it is believed that it is unsafe to alter the reliability rating of Roper *et al.*. Recommended practice would be to run a sensitivity analysis to estimate the impact of this decision.

Considering the second potential partitioning a case can be made that the experiments are measuring a very different cognitive processes. In the latter

Type of test	Statistic	R Vs F	F Vs S	R Vs S
Fixed	$d+$	-0.29	0.09	-0.49
	$d+^U$	0.00	0.36	-0.21
	$d+^L$	-0.58	-0.18	-0.78
	σ^2	0.02	0.02	0.02
	p	0.02	0.25	<<0.01
Random	Δ	-0.07	0.11	-0.30
	Δ^U	0.71	0.57	0.35
	Δ^L	-0.85	-0.33	-0.96
	σ^2	0.64	0.22	0.45
	p	0.42	0.31	0.17

Table 5: Combining effect sizes from all the experiments

experiments, structural testing is undertaken using a tool to assist with the calculation and monitoring of the coverage criteria. In the Basili and Selby experiment, coverage calculation and monitoring was manual, this greatly alters the intellectual 'puzzle' presented to the subjects[5]. Hence we could propose the binary-valued variable of *tool support?* as a moderator variable[6]. Although we should now proceed by testing the new proposed partitions via separate meta-analyses, given that the largest partition has two[7] studies it was decided to stop the process, for the sake of brevity.

Combining Studies

Unfortunately since we have not being able to resolve the heterogeneous nature of the data set, we cannot safely continue and combine them. Hence the following figures and discussion are principally for illustration rather than further serious analysis. To finally produce a combined effect size estimate we return to the fixed-effect and random-effect models of the previous section. Again applying the procedures from Hedges[19], the meta-analyst can compute a combined effect size estimation as illustrated in Table 6. The table uses the notation from Hedges' book; for both models: the combined effect size estimate, the 95% confidence levels, the variance and the probability of the combined effect size estimate being significant are given. As we can see the two models are, as expected, divergent. The fixed-effect model is extremely overly optimistic about the combined effect size of the disparate studies; the random-effect model, as expected in this situation, provides results which seem much more reasonable, and as expected was more able to accommodate the large variance within the studies. The confidence intervals provide another sensitivity check of the results; these can easily be converted into probabilistic statements, and for a

[5] The experiment by Myers had no tool support and no coverage criteria.

[6] To be more accurate a second variable must be introduce of *coverage criteria?* to cover the Myers experiment

[7] tool support: Roper & Kamsties; Manual + Coverage: Basili; Manual + No Coverage: Myers

significant result provides a statement about the proportion of the combined effect size estimation's distribution that would pass any significance test. Producing sensitivity analyses from various viewpoints is essential to establish the reliability of a significant result derived from meta-analytic procedures.

Where does this leave us? After analysing our case study we are unable to provide a consistent picture of which detection technique is best. Further, and of greater concern, we are unable to say that the empirical work undertaken to date, on this topic, is additive and that we are on the 'correct path' to providing a definitive answer. So what is the problem? In a word — variation! Software Engineering experiments are subject to massive amounts of variation from a large number of different sources; if we want to build a solid empirical discipline for the subject the field must invest in techniques, existing and new, which seek to reduce some of these sources.

5 Discussion

The defect detection case study has shown that the heterogeneity of current empirical results is a major limitation in our ability to apply meta-analytic procedures. Further, there is no reason to believe that our case study is unrepresentative of the current 'state of the art'. The good news is that we are not alone in this battle, other disciplines also struggle with this issue, they have just been struggling longer, and hence have progressed beyond the point where we currently find ourselves. Other disciplines traditionally attempt to reduce the heterogeneity within a set of experiments by having:

- Common criteria for the inclusion or exclusion of studies, see[31].

- Standardised effect size measures and simple presentations, — this reduces the error term and the conceptual confusion in comparing similar terms. The binomial effect size display, displaying odds ratios, is a good example of 'state of the art' in other disciplines[36].

- Well-defined 'moderator' variables, that transform our heterogeneous data set into a group of homogeneous data sets, this area was touched upon in the last section with the inability to define 'defect type'.

Other disciplines have introduced various mechanisms in an attempt to advance these objectives:

- The Cochrane Collaboration — In the seventies, Archie Cochrane criticised the medical profession for not having an established system for producing 'up to date' summaries of research results[8]. The Cochrane Collaboration was founded in 1993 to respond to this challenge.

  The Cochrane Collaboration is an international organisation that aims to help individuals make well-informed decisions about healthcare by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare treatments. The Collaboration's work is based upon eight key principles[8]:

[8] http://www.cochrane.co.uk

  - Collaboration, by internally and externally fostering good communications, open decision-making and teamwork.
  - Building on the enthusiasm of individuals, by involving and supporting people of different skills and backgrounds.
  - Avoiding duplication, by good management and co-ordination to maximise economy of effort.
  - Minimising bias, through a variety of approaches such as scientific rigour, ensuring broad participation, and avoiding conflicts of interest.
  - Keeping up to date, by a commitment to ensure that Cochrane Reviews[9] are maintained through identification and incorporation of new evidence.
  - Striving for relevance, by promoting the assessment of healthcare interventions using outcomes that matter to people making choices in healthcare.
  - Promoting access, by wide dissemination of the outputs of the Collaboration, taking advantage of strategic alliances, and by promoting appropriate prices, content and media to meet the needs of users world-wide.
  - Ensuring quality, by being open and responsive to criticism, applying advances in methodology, and developing systems for quality improvement.

  It is unlikely that the current amount of activity within empirical Software Engineering is sufficient to currently require the formation of such an organisation. Hence this should be seen as a long-term aim of the discipline.

- Electronic recording of empirical results: e.g. MEDLINE is a database for retrieving published studies of epidemiological data[31]. Unfortunately this is still not sufficient as most ‘mature’ empirical disciplines suffer from ‘publication bias’. Publication bias is described as the greater likelihood of research with statistically significant results being published compared with research with nonsignificant and null results. Other disciplines have developed quasi-statistical methods to attempt to estimate the possible impact of publication bias[29]. Software Engineering has an advantage here, if we start recording **all** the studies now, then the field can effectively ignore publication bias. Recording this information in a database, with say WWW access is trivial, and only requires willingness on behalf of the funding bodies and the practitioners to realise this objective and avoid this problem. Publication bias will also effect heterogeneity tests, as since the published articles have significant results, the heterogeneity tests will only see a skewed proportion of the population, and hence the test, in general, will underestimate the potential problem.

[9] This is the main output of the collaboration, and the principal mechanism used in achieving their goal.

- Eligibility Criteria: the presented case study only implemented a very loose idea of eligibility (measuring similar effect), given the number of studies available any tighter definition of eligibility would have probably stopped the meta-analysis due to lack of studies. Other disciplines with more studies at their disposal can enforce more exact eligibility criteria, For example
  - Study Design — In epidemiological trials, the average effect of a new treatment has generally been found to be larger in nonrandomised than in randomised studies[27]. Hence the meta-analysis should only use either data set.
  - Sample Size — Some of the statistical methods for meta-analysis are asymptotic methods. Asymptotic methods will tend to overestimate the precision of small studies[14]. When this occurs, the study will over contribute to the meta-analytic results.
- Study Quality — As for eligibility criteria, other disciplines are developing detailed criteria for accessing the quality of a study[7]. Information about quality rating can be incorporated into the statistical procedures.
- Quality Standards for Meta-Analysis: other disciplines recommend the adoption of a study protocol before starting your meta-analysis. This protocol derived from a common framework ensures that each study is undertaking the same tasks and should ensure repeatability and reliability of the individual studies[41].

If we are to progress with the application of meta-analysis to Software Engineering data, the above must be adapted and adopted. Unfortunately Software Engineering is likely to require further methods and procedures to counteract variation as the lack of well-defined terms means that the problems are probably more acute than in many other disciplines.

6 Limitations of Meta-Analysis

Many of the concerns about meta-analysis are based upon philosophical arguments about the nature of the undertaking. Further formal analysis of the technique has not always been favourable. Finally despite its initial conception, meta-analysis is now being applied to non-experimental studies, these studies have additional requirements, which meta-analysis finds difficult to accommodate.

6.1 Philosophical Issues

These can be broadly classified as "Garbage in, Garbage out" and "meta-analysis tries too compare Apples with Oranges?". The first point is with regard to the quality of the studies under investigation, and certainly holds weight if all the meta-analysis has is poor quality studies. But if we don't have confidence that an 'additive process' can be defined from a series of independent high-quality studies, then what is the point in doing empirical research in Software Engineering, this point strikes at the heart of the topic — if we believe in empirical research in Software Engineering, then we should believe that the results of independent studies can be combined. The question of distinguishing the quality of studies is a key concept within meta-analysis, and certainly requires further attention, especially within Software Engineering. The second issue refers to the diversity ideas, subjects, study design, etc., that will be pooled together in a single meta-analysis. Meta-analysis and statistical testing have direct parallels here. Subjects behave differently within studies, but we are happy to generalise with regard to them, then why not studies? Further, if subjects behave very differently, when we blocked them on study characteristics to help our understanding, again why not studies? Finally if we want to make statements about "Fruit", what else can one do except talk about "Apples" and "Oranges" together.

6.2 Formal Evaluations of Meta-Analysis

Sacks *et al.*[41] and Chambers *et al.*[6] have evaluated the quality of a large number of meta-analyses of epidemiological data. Sacks *et al.* are critical that the standard of the application of meta-analysis was insufficient to produce definitive results, he recommended that further work was required on the specification of both the frameworks for replication and meta-analysis reporting. Chalmers *et al.* studied the replicability of meta-analysis. His results suggest some concern as the results of meta-analysis estimation are regularly disparity with the single 'definitive' study, which represents the current understanding within medical circles. Additionally he investigated the replicability between meta-analyses and found that for a large number of (medical) studies the replicability was good. The discrepancies between the meta-analyses and the 'definitive' studies is obviously concerning, and obviously requires more work to present a clear picture. The replicability of meta-analysis of non-experimental studies remains unstudied.

6.3 Using meta-analysis with non-experimental studies

In experimental studies, randomisation (in theory) eliminates bias and confounding within the experimental results. Non-experimental studies will not provide any safeguards with regard to these issues, and hence results from them cannot be considered either unbiased or unconfounded. Currently no work exists, which attempts to validate the use of meta-analysis for non-experimental results; unfortunately many disciplines have started to use meta-analysis in this fashion, this paper strongly recommends that Software Engineering does not follow this example until some research on the question of validity has been carried out.

Further meta-analysis should not be employed to resolve differences between conflicting results. Meta-analysis was designed to combine results from a series of experiments, each of which had insufficient statistical power to reliably accept or reject the null hypothesis. It was not designed to pool disparate answers, meta-analysis is at its weakest when dealing with heterogeneous studies, and again this paper strongly recommends that Software Engineering avoids using the technique in this manner.

7 Conclusions

The ability to reliably combine empirical results from independent experiments is an essential building block in any discipline attempting to build a solid empirical foundation. The standard technique used in most disciplines to achieve this goal is meta-analysis. This paper attempts to introduce meta-analysis into the Software Engineering context. Unfortunately the results are disappointing, and if the case study is representative, indicates that the discipline must embark upon a period of improvement to reduce the variability between replicated experiments or experiments examining the same hypothesis.

The paper outlines various approaches both within the meta-analytic procedure and within its surrounding infrastructure, which have helped other disciplines advance their practice. In addition to these large-scale advances, there are a number of alterations that we, the individual experimenters, can make to our current practice to ready the field for the application of meta-analytical techniques:

- Although the significance test is obviously an important result from the experimental procedure, it is by no means the full story. The effect size is equally important, without it other researchers are in a poor position to estimate the importance of the results, even if they are significant. Unfortunately few, if any, Software Engineering experiments report effect size estimates, their incorporation into the results of empirical studies would greatly aid other researchers.

- The inclusion of raw data is obviously the ideal scenario when reporting an experiment. If this is not possible, because of space restrictions, it is important to report the mean[10], variance (or standard deviation), and details about the normality of the data, quite often a simple histogram is sufficient.

- As described above, randomisation is important to meta-analysis (and experimentation, in general), it is important that these details are fully described.

- Meta-analysis calculations require the number of subjects to be defined, this is not always derivable from the degrees of freedom associated from a test. The inclusion of both[11] is required.

[10] This assumes that the experiment is comparative in nature and normally distributed

[11] In fact, given the number of subjects, deriving the degrees of freedom is trivial, and hence it could be omitted.

- When reporting non-significant results it is important to give an estimate of the statistical power of the experiment[26]. Without it subsequent experimenters have no basis for accepting the null result at any effect level.

- Drawing reliable conclusions from reading an article is a difficult task, this can be made considerably easier if the author describes their search for interactions between the variables, both significant and null interactions. This information is also of great use to the meta-analyst, as it often provides a starting point for undertaking focussed comparative studies upon encountering heterogeneous data.

- For meta-analysis, the independence of the studies is important, hence although attractive from a timesaving point of view, the re-use of materials, etc. sets up potentially dangerous correlations. Hence it is better for the author to describe the essential characteristics of the experiment and use the current version of the experiment as an example rather than simply describing the current version.

- 'Recipe improving' and changes to the experimental design, procedures, materials, etc are seen in a positive light as they help to ensure independence.

- Definitions of measurements — what are required here is not abstract mathematical models describing the theory, but precise details in the actual implementation.

- Simple experiments — Often Software Engineering experiments have asked subjects to undertake a number of different intellectual activities, this is believed to be a high-risk strategy. The Subject's motivation, enthusiasm, ability to comprehend the task, ability to learn (if the activity is new) the task, etc. are all undermined every time they are asked to undertake another activity. It is suggested that the minimum 'two activities or concepts' subject-based experiments should be the norm. This obviously has cost and time implications for the topic in the short-term, but it is believed that the gain in producing more reliable results will outweigh this restriction. Additionally 'two concept' experiments lead to easier formulations of effect sizes and run, on average, less risk of creating conceptual confusion amongst subsequent reviewers.

This may all sound like a very negative view of the field, but remember — "Rome was not built in a day!". The field is still very young, and requires time and effort to mature, this maturity will bring enormous benefits currently enjoyed by mature empirical disciplines such as the medical sciences and social psychology.

References

[1] R.D. Banker and C.F. Kemerer. Scale economies in new software development. *IEEE Transactions on Software Engineering*, 15(10):1199–1205, 1989.

[2] V.R. Basili and R.W. Selby. Comparing the effectiveness of software testing techniques. *IEEE Transactions on Software Engineering*, 13(12):1278–1296, 1987.

[3] J.A. Berlin, N.M. Laird, H.S. Sacks, and T.C. Chalmers. A comparison of statistical methods for combining event rates from clinical trials. *Statistical Medicine*, 8:225–230, 1989.

[4] A. Brooks. Meta analysis — a silver bullet — for meta-analysts. *Journal of Empirical Software Engineering*, 2:333–338, 1997.

[5] A Brooks, J Daly, J Miller, M Roper, and M Wood. Replication's role in experimental computer science. Research report EFoCS-3-94, Department of Computer Science, University of Strathclyde, Glasgow, 1994.

[6] T.C. Chalmers, H.S. Sacks, H. Levin, D. Reitman, and R. Nagalingam. Meta-analysis of clinical trials as a scientific discipline: II. replicate variability and comparison of studies that agree and disagree. *Statistical Medicine*, 6:733–744, 1987.

[7] T.C. Chalmers, H. Smith, B. Blackburn, B. Silverman, B. Schroeder, D. Reitman, and A. Ambroz. A method for assessing the quality of a randomised control trial. *Controlled Clinical Trails*, 2:31–49, 1981.

[8] A.L. Cochrane. *Effectiveness and Efficiency. Random reflections on health services.* London: Nuffield Provincial Hospitals Trust, 1972.

[9] J Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Academic Press, first edition, 1969.

[10] R. Davidson and J.G. MacKinnon. Several tests for model specification in the presense of alternative hypothsese. *Econometrica*, 49:781–793, 1981.

[11] M.L. Davison. *Multidimensional Scaling.* Krieger Publishing, 1992.

[12] J. Gilbert, B. McPeek, and F. Mosteller. Statistics and ethics in surgery and anesthesia. *Science*, 198:684–689, 1977.

[13] G.V. Glass, B. McGaw, and M.L. Smith. *Meta-analysis in social research.* Beverly Hills, CA:Sage, 1981.

[14] S. Greenland. Quantitative methods in the review of epidemiologic literature. *Epidemiological Review*, 9:1–30, 1987.

[15] J.A. Hall. Skill in nonverbal communication: Individual differences. In R. Rosenthal, editor, *Gender, gender roles, and nonverbal communivcation skills*, pages 32 – 67. Cambridge MA: Oelgeschlager, Gunn and Hain, 1979.

[16] M.J. Harris and R. Rosenthal. Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, 97:363–386, 1985.

[17] L.V. Hedges. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6:107–128, 1981.

[18] L.V. Hedges. Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, 92:490–499, 1982.

[19] L.V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis.* Academic Press, 1985.

[20] W.C. Hetzel. *An experimental analysis of program verification methods.* PhD thesis, Department of Computer Science, University of North Carolina at Chpael Hill, 1976.

[21] Q. Hu. Evaluating alternative software production functions. *IEEE Transactions on Software Engineering*, 23(6):379–387, 1997.

[22] J.E. Hunter, F.L. Schmidt, and G.B. Jackson. *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA:Sage, 1982.

[23] E. Kamsties and C.M. Lott. An empirical evaluation of three defect detection techniques. Technical report, ISERN-95-02, International Software Engineering Network, 1995.

[24] E. Kamsties and C.M. Lott. An empirical evaluation of three defect-detection techniques. In *Proceedings of the Fifth European Software Engineering Conference*, pages 362–383, 1996.

[25] J. Kulik. Book review: Review of G.V. Glass et al., meta-analysis in social research. *Evaluation News*, 4:101–105, 1983.

[26] J. Miller, J. Daly, M. Wood, M. Roper, and A. Brooks. Statistical power and its subcomponents – missing and misunderstood concepts in emprical Software Engineering research. *Journal of Information and Software Technology*, 39:285–295, 1997.

[27] J.N. Miller, G.A. Colditz, and F. Mosteller. How study design affects outcomes in comparisons of therapy: II. surgical. *Statistical Medicine*, 8:455–466, 1989.

[28] G.J. Myers. A controlled experiment in program testing and code walkthoughs/inspections. *Communications of the ACM*, 21(9):760–768, 1978.

[29] R.G. Orwin. A fail safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8:157–159, 1983.

[30] K. Pearson. Report on certain enteric fever inoculation statistics. *British Medical Journal*, 2:1243–1246, 1904.

[31] D.B. Petitti. *Meta-analysis, decision analysis and cost-effectiveness analysis: methods for quantitative synthesis in medicine.* Oxford University Press, 1994.

[32] L.M. Pickard, B.A. Kitchenham, and P.W. Jones. Combining empirical results in software engineering. Technical report, University of Keele, Dept. Computer Science, 1998.

[33] A.A. Porter and P.M. Johnson. Assessing software review meetings: Results of a comparative analysis of two experimental studies. *IEEE Transactions on Software Engineering*, 23(3):129–145, 1997.

[34] H.C. Romesburg. *Cluster analysis for researchers.* Lifetime Learning Publications, 1984.

[35] M. Roper, M. Wood, and J. Miller. An empirical evaluation of defect detection techniques. *Information and Software Technology*, 39:763–775, 1997.

[36] R. Rosenthal. *Meta-Analytic Procedures for Social Research.* Beverly Hills, CA: Sage, 1984.

[37] R. Rosenthal and D.B. Rubin. Interpersonal expectancy effects: The first 345 studies. *The Behavioural and Brain Sciences*, 3:377–386, 1978.

[38] R. Rosenthal and D.B. Rubin. Issues in summarizing the first 345 studies in intersonal expectancy effects. *The Behavioural and Brain Sciences*, 3:410–415, 1978.

[39] R. Rosenthal and D.B. Rubin. Further meta-analytic procedures for assessing cognitive gender differences. *The Journal of Educational Psychology*, 74:708–712, 1982.

[40] R.J. Rummel. *Applied factor analysis.* Evanston IL: Northwestern University Press, 1984.

[41] H.S. Sacks, J. Berrier, D. Reitman, V.A. Ancona-Berk, and T.C. Chalmers. Meta-analysis of randomized controlled trials. *N. England Journal of Medicine*, 317:450–455, 1987.

[42] M. Smith and G. Glass. Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal*, 17:419–433, 1980.

[43] H.M. Walker and J. Lev. *Statistical Inference.* New York: Holt, Rinehart and Winston, 1953.

[44] F.M. Wolf. *Meta-Analysis: Quantitative methods for research synthesis.* Sage Publications, 1986.