Research Synthesis in Software Engineering: A Case for Meta-Analysis

Will Hayes

Software Engineering Institute: Carnegie Mellon

Abstract: The use of meta-analytic techniques to summarize empirical software engineering research results is illustrated using a set of 5 published experiments from the literature. The intent of the analysis is to guide future work in this area through objective summarization of the literature to date. A focus on effect magnitude, in addition to statistical significance is championed, and the reader is provided with an illustration of simple methods for computing effect magnitudes.

Background

The field of empirical software engineering research can benefit from research synthesis techniques that help summarize and assess the body of empirical results accumulating in the literature. Research synthesis techniques that go beyond the subjective review of literature found in nearly every dissertation could help researchers to build on a framework of existing results, rather than merely commenting on it. In order for the benefits of research synthesis techniques to be realized however, basic standards of research conduct and reporting must be present. Failing some level of comparability among subject populations, sampling standards, and research designs, summaries of existing research tend to focus on the shortcomings or strengths of the research methods used – rather than the results and data reported. Meta-analytic methods allow us to summarize the outcomes of previous research, and form empirically derived expectations for future research focus.

In coining the term "Meta-Analysis," Glass (1976) defined it in the following way:

" I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature."

Though Glass is credited with the popularization of meta-analysis, statistical procedures for combining the results of multiple research studies were also explored in the 1930s during the development of statistical methods for the analysis of agricultural experiments (Tippet 1931, Fisher 1932, Karl Pearson 1933, Cochran 1937). This paper provides an illustration of a minimal set of meta-analytic methods, using a series of 5 published experiments on requirements inspection techniques.

Meta-Analysis in Software Engineering

In their illustration of the use of meta-analysis in software engineering, Pickard, Kitchenham and Jones (1998) conclude that:

"*Meta-analysis is appropriate for homogeneous studies when raw data or quantitative summary information, e.g., correlation coefficient, are available. It can also be used for heterogeneous studies where the cause of the heterogeneity is due to well-understood partitions in the subject population."*

The problems associated with estimating comparable effect magnitudes across a series of heterogeneous experiments are immediately obvious to anyone who has ever written a literature review. Comparing apples and oranges without resorting to a discussion of fruit juice is very difficult. Advocates of meta-analysis recognize this as a grey area in the use of these methods. For instance, Cook and Leviton (1980) emphasize that there will always be subjective elements to the practice of meta-analysis, as in most research methods, when they say:

"…all literature reviews share both qualitative judgments and quantitative techniques. Meta-analysis is rife with qualitative judgments – about the population of studies that are relevant, the breadth of the constructs to be investigated, the criteria by which studies are to be grouped into those of high and low methodological quality, etc. "

Reconciling the implications for differences in sampling frames, operational definitions, and study designs (to name only a few challenges) is rarely a straightforward exercise. However, the ability to introduce new independent variables in the analysis and to explore novel combinations of experimental conditions makes meta-analysis

a desirable addition to the subjective narrative found in most literature reviews.

A little closer to home, Brooks (1997) commenting on the lessons from psychological research for the field of software engineering noted:

"A major lesson has been that a single study is extremely unlikely to be definitive. Replication variants of the study can often fail to be confirming. Dozens and eventually hundreds of studies can follow on the same topic. Knowledge singularly fails to accumulate, however, and it even becomes attractive to dismiss an entire research literature."

Brooks goes on to conclude that meta-analysis can be useful "if there is indeed a single effect size out there waiting to be estimated" but that the field of software engineering simply lacks a sufficient body of replicated empirical studies to support this investigation. The answer to the first challenge is faced by every researcher who performs an experiment using their operational definitions of the important variables and relies on statistical analyses to reject the null hypothesis. The second problem, shortage of empirical results, is being addressed through efforts of researchers like the ones described here. In addition, the journal *Empirical Software Engineering* strives to provide its audience with detailed data from the studies it publishes in order to facilitate building on the findings from the single study which initiates a line of inquiry.

Software Requirements Inspection Techniques – A Series of Replicated Experiments

Porter, Votta and Basili (1995) initiated a program of experimental research to study inspection method in software engineering. The authors set forth a logical basis for the desirability of a new inspection technique, where each member of the inspection team is assigned specific responsibilities based on scenarios associated with particular types of defects. The experimental results reported provide statistical support to the logical argument set forth. These researchers also made their experimental materials available to other researchers, so that replications of their original experiment could be performed independently.

Since the publication of the first experiment, 4 replications have been reported in the same journal (Empirical Software Engineering) with a variety of results. Taken as a set, these 5 experiments address a variety of important research questions, including analysis of 'meeting gain' to investigate the hypothesis that holding the inspection meeting does not significantly increase the defect detection rate. As well, there are differences in the statistical models employed for analysis in some of the articles. Most authors performed a series of one-way ANOVAs to examine each independent variable, then went on to examine other effects like the interaction between inspection technique and specification, or the performance of individual inspectors. However the basic underlying designs are quite similar. For the sake of brevity, the focus of this paper is restricted to the effect of inspection technique, and the effect of specification on the defect detection rate of inspection teams. The following 5 sections provide a brief summary of each experiment.

The Porter, Votta and Basili Experiment

Porter, Votta and Basili (1995) conducted an experiment using 48 graduate students enrolled in a course on formal methods. The design of this experiment resembles a Greco-Latin Square arrangement, where multiple layers of a Latin-Square design are used to test conditions arising from the manipulation of a number of independent variables (Winer 1971). An internal replication within the design required 2 sets of 24 students. In addition, the experimenters exercise some control with regard to subjects' exposure to the different inspection techniques. They avoided situations where use of a systematic method of inspection is followed by use of a less systematic method. The concern is that the previous experience could add structure to the subsequent inspection round, resulting in a different process than that of the subjects who did not have that previous experience.

Given the importance of distinguishing the effects of related treatments (inspection methods that vary along a single dimension – how systematic they are) and to make the most of a limited subject pool, these design choices seem to reflect a good tradeoff. However, some researchers will prefer a saturated factorial model where no effects are confounded, and all interactions can be estimated. Indeed we see such a full factorial model used in one of the replications described later in this paper.
The chief results of this first experiment (for the purposes of the present paper) included a 35%

improvement in detection rate for inspection teams using the scenario method, and statistically significant differences in inspection rate associated with the inspection method as well as the specification inspected (for teams of inspectors).

The First Replication

Fusaro, Lanubile and Visaggio (1997), using the experimental kit provided by the authors above, replicated the experiment using undergraduates in an advanced software engineering course. All participants had "experience from a previous software engineering course in SRS reading but only in the information systems domain." The experimental materials were translated into Italian.

The experimental design is presented in an experimental plan along with the plan from the original experiment. An additional constraint was placed on the set of permissible sequences for teams' exposure to inspection techniques for this replication. These authors also exercised some control over inspection team formation by using a matching procedure as "an adjunct to randomization." Finally, these investigators found additional defects in the requirements specifications that were not considered during the initial experiment. Results pertaining to the original set of defects as well as the expanded list are reported.

These authors reported statistically significant differences between the two requirements specifications, but no statistically significant differences among the inspection techniques were found, using a series of one-way ANOVAs.

The Second Replication

Miller, Wood and Roper (1998) provide a third set of empirical results for the experiment. The subjects in this replication were third-year undergraduates in a formal university course. These authors dropped the Ad Hoc approach from the design, leaving only two detection methods to compare – checklist and scenario. In addition, a new constraint on the order of presentation for the requirements specifications was used, to minimize the chance of the experimental data being contaminated by subjects' discussions with one another regarding the specification they had just reviewed. Finally, the subjects in this replication each used only one inspection technique – either checklist or scenario – for both experimental rounds.

The article describing this replication, like the other articles, provides a thorough discussion of threats to internal and external validity. These authors in particular discuss several key issues associated with the design and statistical power of the experiment. They make explicit their intention to boost statistical power by increasing the sample size – while simplifying the design. This replication contains the largest sample size of all the replications, with a total of 50 subjects organized into 16 teams.

The unit of analysis underlying the majority of results reported in this article is the individual inspector, rather than the inspection team. Because of this focus, the data available from the published article alone cannot support a detailed meta-analysis in conjunction with the published figures from the other replications. However group means and standard deviations are estimated from tables in the article (1), and used here with a minor concern for the distorting effects of rounding error. The results of F-tests performed using individual level data are hardly comparable with the group level analyses reported in the other replications.

These researchers found no statistically significant difference in the detection rate of individual inspectors associated with the inspection technique used. The authors comment that the p-value of .10 associated with this test suggests that their findings are more similar to the outcome of the original experiment than the first replication. In the context of our meta-analysis, the similarities between the mean detection rates (to be illustrated later) provide a more compelling case for claiming this similarity. Indeed, the author's own analyses confirm that results for individual data are not comparable to the results for team inspection performance. There are methodological and statistical grounds for this difference, and the actual ANOVA results reported by these authors differ depending on the unit of measurement.

Analysis of the data at the group level (combining inspectors into teams) reveals no statistically significant effect for inspection method. However, the authors comment that the higher p-values associated with the group level analysis suggest that their results support the non-significant method effect uncovered in the first replication, rather than the result reported in the original experiment.

The Third Replication
Sandahl, Blomkvist, Karlsson, Krysander, Lindvall, and Ohlsson (1998) published the third replication of the experiment. The subjects in this replication consisted of a sample of 24 undergraduate students in Sweden. These researchers focused on only three of the independent variables in the original model (method, specification, and order). They also elected to focus on only two inspection techniques (scenario and checklist).

The randomized factorial design was balanced in this replication, with each team inspecting both specifications and using both techniques. The statistical techniques used to analyze the experimental data took advantage of the full factorial design, and differed substantially from the analysis strategies in the other replications. However, these authors published a complete table of detection rates for all teams in the experiment. This table and the clear explanation of the research design provided by the authors permit us to re-analyze the raw detection rates using a model of our choosing.

The original experimental materials were used, supplemented with some instructions written in Swedish and subjects were given access to a dictionary during the experiment. The authors characterize their subjects as novice inspectors, reporting that this experiment represents the first participation in a requirements inspection for most of the students. Lack of familiarity with and a dislike for the notation used in the experimental materials was also reported by these subjects. Finally, very few subjects had driven automobiles with cruise control systems, which is the focus of one of the requirements specifications used in these experiments.

The ANOVA table reported for the three main effects and four interaction effects in the model shows that none of the effects accounts for a statistically significant percentage of variability in the dependent variable. The authors also use a pair of normal probability plots of these effects along with a set of pre-specified contrasts, to interpret trends in the data. They conclude, “the specification is the most significant explaining factor of the variance amongst the independent variables.”

The Fourth Replication
Porter and Votta (1998) performed a replication of their own experiment using 18 professional subjects. Their confirmation of the original experimental results regarding inspection method is used to bolster the credibility of initial research results based on graduate student-subjects. The researchers used a design much like that of the original experiment.

Because of the restrictions on the ordering of inspection techniques used and the smaller sample size of this replication, the number of inspection teams for each level of each independent variable is not always equal. The unbalanced Greco-Latin square design used in this replication presents some challenges in assuring that the effect of maturation does not impact the results.

Given the similarities in design, materials used and experimental procedures, the authors treat this replication as a third in the series of experiments they have conducted in this area. Recall that the first experiment contained an internal replication, where each set of 24 subjects was studied in a separate experimental run.

The findings of this replication with respect to the inspection method variable, are (as mentioned above) consistent with the results of the first experiment. Namely, these researchers found statistically significant differences in the defect detection rates of teams using different inspection methods. However, there were no statistically significant differences in detection rates associated with the specification being inspected. This finding differs from the results of the original experiment, and all replications described above.

Combining the Five Experiments
How will the authors of the next replication reconcile these seemingly contradictory results? The narrative review of literature in the introduction will likely focus on the design adequacy of the preceding replications. If contradicting results are not totally ignored, subjective criticisms of study designs may be offered in order to encourage readers to discount previous results or modifying their interpretation.

In the case of this series of experiments, the authors of the 4 replications listed above have provided thoughtful explanations to reconcile differences. It is interesting to note however, that

in the final replication, Porter and Votta make no mention of the other replications whatsoever. By making the experimental materials available to other researchers, these authors were obviously inviting the replications. Given their approach to sharing experimental materials, it would not be surprising to hear that the authors of the final replication were aware of, and perhaps even helpful to, the authors of the preceding replications. Yet they do not comment on the implications associated with lessons learned from the replications. Why?

The Problem

Where conditions that warrant differential interpretation of individual replications are readily obvious, authors typically lack objective methods for driving that interpretation. The challenge addressed in this paper is the USE, rather than dismissmissal of seemingly contradictory findings. By implication, the approach in common use today leads each succeeding author to sort-out the wheat from the chaff as they attempt to sum-up the state of the knowledge in the domain under study. This process of picking out the profound knowledge among all the scholastic calisthenics is hardly a scientific process in most applications. If we are to truly grow the body of knowledge, we must consider a more scientific (empirically driven) basis for summing-up (Light and Pillamer 1984) our collective knowledge.

Combined Tests of Significance

In the development of statistical methods to support agricultural experiments, a number of procedures for combining multiple statistical tests can be found. Among these the work done by Tippet (1931) is perhaps the earliest instance of a method for testing the statistical significance of combined results. Soon after, R.A. Fisher (1932) and Karl Pearson (1933) independently derived the same procedure for combing statistical significance across multiple studies. This latter method (alternatively called Fisher's method or Pearson's method) uses the equation below to test the null hypothesis that the treatment effect in every study is equal to zero.

$$\chi^2 = -2 \sum \log_e p \qquad [1]$$

When the null hypothesis is true the sampling distribution of the χ^2 calculated above is approximately distributed as a chi-square distribution with 2n degrees of freedom, where n is the number of p-values being combined (Wolf 1988).

Applying Fisher's method

Table 1. below gives the p-values deriving from one-way ANOVAs carried out for each of the replications of the requirements inspection experiments. The first row contains p-values associated with differences between inspection techniques, and the second row contains the p-values for differences between the specifications. Each column in the table represents a replication, with the first column containing the p-values for the original experiment. (The p-values in the third column derive from a different unit of analysis, as described above, and are included solely for the purpose of illustration).

Table 1. P-values for each experiment

	1	2	3	4	5
Method	.0016	.77	.10	.8934	.0103
Specification	.0014	.003	.01	.0248	.6997

Multiplying the sum of the natural logarithms of each of the p-values in the first row by –2, we obtain a test statistic of 27.380 for the Method effect. The corresponding test statistic for the Specification effect is 42.079. Comparing these values with the tabled chi-square value for 2n (or 10) degrees of freedom we find sufficient reason to reject the null hypothesis at the .05 level of significance.

The tests carried out above, in rejecting the null hypothesis for each effect, found that not all effects are equal to zero. This conclusion however is not very satisfactory as it tells us very little about the specific effects. Our omnibus test does not support a ranking of the 5 sets of experimental results, and in fact says nothing about the magnitudes of the effects measured in each experiment.

Effect Magnitudes

The preferred approach to meta-analyses relies on quantifying the effect magnitudes for the results to be combined. This more methodologically challenging approach is the primary focus of leading authors in meta-analytic methodology (e.g., Hedges and Olkin 1985, Hunter, Schmidt and Jackson 1982). The focus on effect magnitude, rather than statistical significance alone, has some intuitive appeal to many researchers. After all, if some as-yet undiscovered holy Grail of research methods could be used to test our hypotheses with 100%

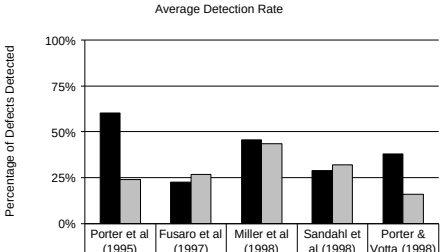
certainty (rather than the probabilistic treatment we give them now), our next question would surely be "by how much?" "By how much does method A improve performance over method B?" This is a question of effect magnitude, and by implication a question of cost and benefit. A million-dollar idea that costs 3 million dollars to implement is hardly 'significant.' The field of empirical software engineering research must address the issue of effect magnitude in order to assure that the results of our research are of consequence to the field of software engineering.

As Pickard et al (1998) conclude, "combining study results is not likely to solve all of the problems encountered in empirical software engineering studies." However, the explicit focus on effect magnitude that meta-analytic methods bring to a research literature would surely be a step in the right direction for our field.

Estimating Effect Magnitudes in the Inspection Experiments

In this section of the paper, a brief and simplistic meta-analysis is performed using data from the 5 experiments summarized above. The results presented in this section provide an objective summary with respect to two of the outcomes studied in each experiment – the effects of inspection method and specification on team detection rate.

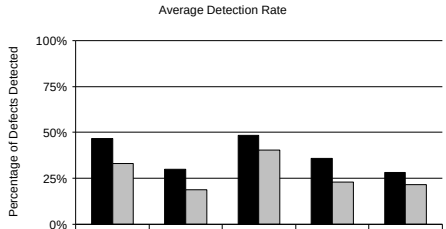
One of the basic ways of understanding the effect of independent variables is to examine dependent variable means for each level of the independent variables. The two figures below provide bar charts for the means of each level of each of the two independent variables selected for meta-analysis.



	Porter et al (1995)	Fusaro et al (1997)	Miller et al (1998)	Sandahl et al (1998)	Porter & Votta (1998)
Scenario	0.60	0.23	0.46	0.29	0.38
Checklist	0.24	0.27	0.44	0.32	0.16

Figure 1. Detection Rates for Inspection Methods

Notice that the bar chart shown in Figure 1 shows very clearly that the 5 experiments differ substantially in the way the two inspection methods compare. The scenario method (shown in black) is obviously associated with a higher average detection rate in the first and last experiment. While the third experiment also shows a higher average for the scenario method than the checklist method, the difference is very small. We know from the reports provided by the authors, that the results of statistical analysis follow the same pattern.



	Porter et al (1995)	Fusaro et al (1997)	Miller et al (1998)	Sandahl et al (1998)	Porter & Votta (1998)
WLMS	0.47	0.30	0.49	0.36	0.28
Cruise	0.33	0.19	0.41	0.23	0.21

Figure 2. Detection Rates for Specifications

In the case of Figure 2, we see a pattern of consistent differences in average detection rate associated with the two requirements specifications used in the experiments. The pattern of results from the statistical tests conducted by the authors does not necessarily lead the reader of these articles to this image.

Based on graphical examination of these means, we might begin to formulate some general conclusions about the 'agreement' or 'disagreement' among the reported results. However, a reliance on central tendency alone, without an understanding of the impact of differences in variability may well lead us astray. Fortunately a method for standardizing the mean differences is provided, using pooled sample standard deviations.

The method given in Hedges and Olkin (1985) for pooling sample standard deviations is shown in [2] below. Note that the reference to experimental and control groups can be replaced with the two conditions of the independent variables we are examining in this case.

$$s_{pooled} = \sqrt{\frac{(n^E - 1)(s^E)^2 + (n^C - 1)(s^C)^2}{n^E + n^C - 2}} \quad [2]$$

Where:
n^E is the sample size of the experimental group

n^C is the sample size of the control group
s^E is the sample standard deviation of the experimental group, and
s^C is the sample standard deviation of the control group

Table 2 (below) lists the means plotted in Figures 1 and 2, the pooled standard deviations for each independent variable, and the effect magnitudes for each variable.

Table 2. Means, Standard Deviations, and Effect Magnitudes for the 5 experiments.

	Porter et al (1995)	Fusaro et al (1997)	Miller et al (1998)	Sandahl et al (1998)	Porter & Votta (1998)
Means					
Scenario	0.60	0.23	0.46	0.29	0.38
Checklist	0.24	0.27	0.44	0.32	0.16
WLMS	0.47	0.30	0.49	0.36	0.28
Cruise	0.33	0.19	0.41	0.23	0.21
Pooled Standard Deviations					
Method	0.10	0.10	0.09	0.13	0.09
Specification	0.14	0.07	0.09	0.11	0.13
Effect Sizes					
Method	3.541	-0.391	0.225	-0.235	2.420
Specification	0.954	1.522	0.901	1.178	0.525

Relying on the characterization of effect magnitude conveyed by the standardized mean differences in the last two rows of Table 2, we can clearly see the differences among the outcomes of the 5 experiments. The largest effect size found in the table, 3.541 indicates that the difference in detection rate associated with the method of inspection in the first experiment is the most pronounced of all differences reported.

The value reported suggests an astonishing difference between scenario and checklist methods that represents an advantage for the scenario method of more than 3.5 standard deviations. When we look at the means (shown in the table, as well as in Figure 1) the pattern of difference is also obvious. Some researchers provide guidelines for interpreting effect magnitudes, based on standards deriving from the field of inquiry (Wolf 1988). While the field of empirical software engineering research lacks established standards for effect magnitudes, the value of 3.5 substantially exceeds any standard known to this author (a typical standard for a "large" effect is 0.8).

Based on the average and standard deviation of the effect sizes, it is possible to construct a confidence interval for the 'population effect size' if our theory suggests such a parameter exists. The average and standard deviations for the effect sizes are given in Table 3, below.

Table 3. Average and Standard Deviations for Effect Sizes in the 5 experiments.

	Average	SD
Method	1.11	1.77
Specification	1.02	0.37

The large value of the standard deviation for the effect sizes associated with inspection method suggests that the effect varies a great deal among the replications. The effect for specification however, seems reasonably stable – though no explicit criteria is offered to guide this judgment.

The 95 percent confidence interval for the method effect size ranges from –0.44 to 2.66. The 95 percent confidence interval for the specification effect size ranges form 0.69 to 1.34. The fact that the first confidence interval spans zero, and that the effect size reported in the first experiment falls outside this interval gives us reason to question the interpretability of the effect sizes for method, as a set. There may be reasons why combining results for the method effect could be misleading.

At this point, methods for formally testing homogeneity can be employed, and a host of other investigative avenues are available – given a sufficient sample of experimental results.

Homogeneity of Effect Magnitudes

We have reason to suspect that the effect magnitudes in this series of experiments do not all derive from a single sampling distribution of the 'population effect magnitude.' Tests of homogeneity can be applied to both sets of effect magnitudes estimated above (one test for the method effect and a second for the specification effect).

Wolf (1988) presents a simple chi-square test to test the homogeneity of effect sizes. The formula for the test statistic with K-1 degrees of freedom

(where K is the number of effect sizes being tested) is given in [3] below.

$$c^2 = \sum \left(w\left(d - \bar{d}\right)^2 \right) \qquad [3]$$

Where;

$$\bar{d} = \frac{\sum wd}{\sum w}$$

and,

$$w = \frac{2N}{8 + d^2}$$

and N is the total sample size associated with each effect size (summing across both levels of the independent variable).

For the method effect in our meta-analysis, the equation in [3] yields a test statistic of 58.341, which is statistically significant ($p < .000001$). For the specification effect, the test statistic equals 3.856, which is not statistically significant.

We therefore conclude, with some confidence that the effect sizes for the method effect are heterogeneous across the set of 5 experiments. The test conducted above does not suggest that a similar statement can be made for the specification effect. This presents us with a rather interesting situation. Is it reasonable for one effect to be homogeneous, while the other is not – in the same set of experiments?

No authoritative answer is offered here, but further analysis (using a larger sample of empirical results) could help identify mediating/moderating variables that operate across the set of studies. Such variables may impact the method effect without a corresponding impact on the specification effect, or vise versa. A brief exploration of some hypotheses along these lines is provided in the next section.

Interpreting the Results

Recall that the samples used in the experiments differed across the set. The initial experiment was conducted with graduate students enrolled in a course on formal methods. The final replication (by two of the original authors) was conducted using a sample of professional engineers who were participating in a training seminar. The three remaining replications were all conducted using undergraduates.

One hypothesis that could explain the differences in the results reported above centers around the subjects' familiarity with the notation used in the requirements specifications. One might argue that a group of graduate students enrolled in a formal methods course would have the greatest degree of facility with the specification language – such formalisms are likely to be a major theme in the course. Furthermore, the authors of the final replication report that the specification language is in use in portions of the organization from which their subjects were drawn. In contrast, several mentions are made in the other articles regarding the students dislike for the specification language. Facility with the specification language might be a mediating variable worth investigating.

Another observation to be made about the different experiments lies in the differences in familiarity with the software domains used in the experiment, as reported by the authors. College students in Italy and Sweden who have not driven an automobile equipped with cruise control may well have a different experience from graduate students and software professionals in the United States. Familiarity with the product domain may be a moderating variable worth investigating.

Each of the replications reported in this paper provides thorough discussions of validity and generalizability issues. Several of the authors used questionnaires to assess subjects' reactions to the experiment to assess hypotheses like the ones mentioned above. Further data collection and replication of the experiment will benefit from these insights as well.

Using these Results

This paper makes no attempt to provide the final word on the interpretation of the set of results discussed above. Rather, the analyses and discussions above tend to open the door for additional, focused, investigation of the effects under study. Using revised versions of the questionnaires offered by some of the researchers, the next replication could focus on collecting data that might explain the differences highlighted here. The estimates for effect magnitudes provided here can also help to set quantitative expectations for the effects to be observed in future research in this area.

Conclusions

Meta-analysis, or the analysis of analyses, can support investigations of important hypotheses and provide researchers with a powerful way to quantitatively summarize a field of inquiry. While the techniques described in this paper can be used to aggregate a set of existing empirical results, these methods also provide ways of building beyond the scope of individual research designs. By incorporating new variables into the analysis, based on conditions reported in the original studies, the meta-analyst can look beyond the set of independent variables included in the previous research. Rather than a technique for uncovering the ultimate verdict on a popular research hypothesis, this paper strives to illustrate how meta-analysis can be used to build upon existing results and focus future inquiry.

Notes

1. Table 11, page 54 of Miller, Wood and Roper (1998) contains a listing of sample sizes, means and standard deviations for the crosstabulation of technique by specification. The effect size estimated here is based on means and standard deviations of each level of each variable, computed across (rather than within) each level of the other variable. Means are simply estimated by multiplying each tabled mean (in the article) by it's associated sample size, then deriving new means using the totals and sample sizes in combination. The pooled standard deviation is computed using formula [2] above, pooling two sample standard deviations at a time.

References

Brooks, A. 1997. Meta analysis – a silver bullet – for meta-analysts. Empirical Software Engineering 2: 333-338.

Cook, T.D., and Leviton, L.C., 1980. Reviewing the literature: A comparison of traditional methods with meta-analysis. Journal of Personality, 48, 449-472.

Fisher, R. A. 1948. Combining independent tests of significance. American Statistician 2(5): 30.

Fusaro, P., Lanubile, F., and Visaggio, G. 1997. A replicated experiment to assess requirements inspection techniques. Empirical Software Engineering 2: 39-57.

Glass, G.V. (1976) Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, Vol. 5, No 10, pp3-8.

Hedges, L. V., and Olkin, I. 1985. Statistical Methods for Meta-Analysis. Orlando, FL: Academic Press.

Hunter, J.E., Schmidt, F.L., and Jackson, G.B. 1982. Meta-Analysis: Cumulating Research Findings Across Studies. Beverly Hills, CA: Sage Publications.

Hwang, M.I., 1997. The use of meta-analysis in MIS research: promises and problems. Database for Advances in Information Systems, vol. 27, no 3, p35-48.

Light, R.J., and Pillemer, D.B. 1984. Summing Up: The Science of Reviewing Research. Cambridge, MA: Harvard University Press.

Miller, J., Wood, M., and Roper, M. 1998. Further experiences with scenarios and checklists. Empirical Software Engineering 3: 37-64.

Pickard, L.M., Kitchenham, B.A., and Jones, P.W. 1998. Combining Empirical Results in Software Engineering. Information and Software Technology 40:811-821.

Porter, A. A., Votta, L. G., and Basili, V. R. 1995. Comparing detection methods for software requirement inspections: A replicated experiment. Technical report, University of Maryland, Dept. of Computer Science.

Porter, A. A., and Votta, L. G. 1998. Comparing detection methods for software requirements inspections: a replication using professional subjects. Empirical Software Engineering 3: 355-379.

Sandahl, K., Blomkvist, O., Karlesson, J., Krysander, C., Lindvall, M., and Ohlsson, N. 1998. An extended replication of an experiment for assessing methods for software requirements inspections. Empirical Software Engineering 3: 327-354.

Tippett, L. H. C. 1931. The Methods of Statistics. London: Williams and Norgate.

Winer, B.J, 1971, Statistical Principles In Experimental Design, McGraw-Hill, New York.

Wolf, F. M. 1988. Meta-Analysis: Quantitative Methods for Research Synthesis. Beverly Hills, CA: Sage Publications.