Empirical Software Engineering, 2, 333–338 (1997)

Meta Analysis—A Silver Bullet—for Meta-Analysts

ANDY BROOKS
Department of Computer Science, University of Strathclyde, Glasgow G1 1XH, Scotland

Many of those who recognize that software engineering problems are people problems, and who engage in subject-based empirical software engineering, will have formed the view by now that their research paradigm is broadly similar to that employed by psychologists. Null hypotheses are framed, experiments are designed, subjects are found (from somewhere), and the resulting data are analyzed by statistical means, interpretations of the data are bound up very much in the null hypothesis and the statistical test of significance. The null hypothesis states that there is no difference between treatment and control group means i.e., the data are all drawn from the same population. The intention, of course, is to reject the null hypothesis and make a claim that something (the treatment) has been found that makes a difference e.g., that software engineers using tool or technique A can deliver in less time and with fewer errors than those using tool or technique B. But what lessons have psychologists learnt in conducting research along these lines for a good few decades now?

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 even 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. There are reasons, of course, for this hellish scenario. It takes time to get good at doing subject-based experiments (ten years?) but, during the apprenticeship, careers depend on publication output. What traps can the unsuspecting experimenter fall into?

Both major styles of experimental design, between- and within-subjects, can trap the unsuspecting experimenter into making false conclusions. In between-subjects designs, separate groups are exposed to one of the control and treatment cases. The assumption is that the groups are broadly comparable and often a simple randomization technique is employed to allocate subjects to groups. The wary experimenter knows that there is, in fact, no guarantee that everything you might desire to be randomized actually is. Wary experimenters really try to know their subjects through pre- and post-questionnaires and even simply talking to them after the experimenter is long over. Some outlier data in a single group might be attributable to exceptional background or experience: and this outlier data may be enough to shift a mean through either side of significance. In within-subjects design, care is taken to expose subjects to all conditions in a randomized order to cater for possible learning effects as subjects make their way through each stage of the experiment. Wary experimenters know learning effects may be asymmetrical, that their attempts to harmonize task difficulty between the stages may not have been good enough, and that subjects may have even grown weary by the time the last condition is presented to them. All these influences need to be checked for.

Unsuspecting experimenters need to know that correlations can arise artificially through restricted sampling and that correlations do not necessarily imply cause and effect. Experiments still get run where time or accuracy, the two major product measures of performance, can avoid being measured or controlled in a precise way. Such errors of omission are really inexcusable. The wary experimenter knows that the time/accuracy trade-off must be fully accounted for.

Another major lesson from psychology has been that researchers still need to take lessons on statistical methods. The relationship between the power of the study, the effect size, the sample size, and alpha is rarely appreciated. Given any three, the fourth can be determined. The power of a study is the probability that statistical significance will be attained in the presence of a real treatment effect. Effect size is often measured as the difference between the treatment and control group means expressed in standard deviation units. Sample size is the number of subjects in each group. The alpha level is the threshold probability, typically 0.05, below which it is believed a difference between means is real and not simply due to chance samplings from the same population. For a given effect size and alpha, power can be increased by increasing the sample size. At a single site, a research team usually makes full use of the available subject pool, so if the null hypothesis has to be accepted, there is little opportunity for the team to run the experiment with more subjects to increase power.

It is, of course, possible to reject the null hypothesis when it is in fact true (alpha or Type I error) and to accept the null hypothesis when it is in fact false (beta or Type II error). There is almost an obsessiveness with controlling Type I error (the smaller the probability value the better, with researchers quoting a significance level of 0.001, if they can, rather than simply say the value is less than 0.05 or 0.01) and often little or no appreciation of the need to control Type II error which is just as crucial an issue in considered experimental design. Failing to reject a null hypothesis has little or no interpretive value, especially if the power of the study has been poor. In this case, the individual experiment has largely been a waste of time and effort unless the data is used to firm up power calculations for the next experiment (or if the results can be contributed via a meta-analysis where an aggregation of experimental results is attempted). So researchers often try and avoid accepting the null hypothesis at all costs. In the twinkle of an eye the obsessiveness with Type I error evaporates and the new focus of discussion are the trends in the data which support the model. The Type I error threshold of 0.05 is subjective anyway, isn't it? The problem, of course, is all the other problems that can go wrong in an experimental study. Even the Type I error rate may not be correct if assumptions behind the statistical test are not met. The wary researcher knows it is better to seek out explanations of the data set and not hold on to the notion of generalization.

Students not majoring in psychology or statistics are fortunate indeed if the relationship between the power of the study, the effect size, the sample size, and alpha are given proper treatment in their undergraduate curricula. Even those armed with the right knowledge know that considering Type II errors takes up time and effort and that it is far easier to play the part of the glory hunter dreaming of discovering a large effect size, so large, that any purely statistical inadequacies can be forgiven (and are). When effect sizes become large enough, visualizations provide convincing enough an argument, and the glory hunters feel

fully justified in their approach. And why not, if something valuable has been discovered, and the results are replicable by an independent research team.

When the null hypothesis is rejected, there is often no standing back from the data to ask if alternative interpretations are possible. Confirmation bias can take a strong hold of not only the glory hunters but even the most objective amongst us when checks should be made that significance doesn't hinge on a few values being dependent on, for example, our most able subjects or subjects who happened to have some additional knowledge or expertise or subjects who employed an unanticipated strategy that the randomization process simply was never designed to take account of. In software development, careful design does not mean errors are not to be found in the code. In experimental work, careful design does not mean that alternative explanations are not to be found in the data.

During analyses, it is important to maintain a constant vigil for these alternatives. The need to check for several influences has already been mentioned. The tactic of knowing your subjects through questionnaires and interviewing should always been employed. Subjects can be asked what they felt contributed to their performance and what they felt to be easy or difficult. Insights can be obtained that would be otherwise impossible to obtain if interpretations were based solely on the outcome of statistical tests. Our subjects are not rats. They can talk back. They should be asked. Wary researchers know to do this and that sometimes what they discover will actually reinforce the purely statistical interpretation, giving real confidence to the results.

Subject-based experimentation involves the most complex organisms we know of (ourselves) and it really is impossible to control every psychometric variable. The wary experimenter knows that there are many variables not being controlled but which may ultimately reveal themselves as explanations of the data set.

Another common mistake is to datamine with statistical tests until a significant result is found whilst failing to recognize that the Type I error rate is no longer at the pre-set level. If you do 100 tests at the 5% level, should it be a surprise that five of them reveal significance? Of course not. A simple Bonferroni adjustment to make is to divide the alpha level through by the number of tests applied, but note that such adjustment procedures have spawned a whole literature for themselves and researchers determined to datamine with statistical tests should really acquaint themselves with this work.

Parametric statistics, where the assumption is that the data are distributed normally, may be erroneously applied to data that is highly non-normal. Unsuspecting experimenters choose to believe the robustness of the statistical procedures that are welded to their designs i.e., they choose to believe that the Type I error computed under conditions of non-normality is reasonably representative of the actual Type I error. Wary experimenters know to test their data for normality, that sources of non-normality are of interest, and that understanding the non-normality may be the key to the explanation of the results. If explanations are not forthcoming, then prior to applying parametric statistics, the wary experimenter knows to check the literature for robustness studies relevant to their experimental designs and analyses.

The discussion here is by no means comprehensive: the list of what can go wrong in a single study is frightening. It is perhaps no surprise that it is proving difficult to replicate empirical software engineering experiments, but try we must if software engineering stu-

dents are ever to enjoy the rich learning experience of repeating classic experiments in the university laboratory.

In a possible future world of empirical software engineering, confronted with dozens, possibly hundreds, of studies on the same topic which together do not paint a coherent picture, what is to be done? The classical narrative review can provide a measure of understanding but will usually fail at determining what knowledge really has accumulated. Such reviews often fail to provide deep criticism (positive or negative) because it can take a lot of time and effort to fully appreciate even a single piece of experimental work: it is almost easier to obtain such appreciation by replicating or partially replicating the experiment. Besides, shortcomings in the statistical analyses cannot be corrected in a purely narrative approach. We all exhibit bias (an undeniable human trait) and it is difficult to maintain complete objectivity when constructing a narrative review. Sometimes a narrative review degenerates into referencing a list of inconclusive empirical work as justification for the reported empirical study (which more often than not turns out, unsurprisingly, to be inconclusive as well). The answer (?), of course, lies in meta-analyses, where effect size estimates from individual studies are used to estimate an overall effect size in a statistically rigorous and objective way. Doubt is swept aside and a quantitative understanding is obtained of how the treatment makes a difference. Such analyses are relatively commonplace in medical and psychological research.

What are the drawbacks? The researchers who are generalists who want to perform a meta-analysis prior to model building and hypotheses testing must learn how to do meta-analyses. Given that many generalists are typically not strong in their ability to perform rigorous statistical analyses of their own experiments, this learning may present a significant barrier. And during the apprenticeship period, publications must still flow from the research laboratory. Meta-analyses are just as susceptible to misapplication and meta-analyses of broadly the same literature can yield different outcomes. (Would meta-meta-analyses be capable of sorting these problems out?) Of course, there is also no guarantee that the meta-analysis will reveal an effect size that the software project manager or developer need concern themselves with. A saving of a few seconds or the detecting of one more software defect on a project lasting years with a cumulative error profile running into thousands is of no consequence. At least a line can then be drawn under that particular avenue of research.

What are the benefits? Enormous. A properly executed meta-analysis can deliver a meaningful determination of effect size if the theory is good and there is indeed a single effect size out there waiting to be estimated. And it is not only large effect sizes that can be of interest. Small effect sizes can be of importance too. In medicine, saving a few more lives has value beyond measurement. In software engineering, saving a little time on a task that is repeated dozens of times a day by hundreds of employees can bring substantial cost-benefit. So it is important to have process models in place to help guide our investigations into worthwhile areas and to judge the real-world impact of any treatment that has been deemed to have a measurable effect. If the distribution of effect sizes is spread out (i.e., there is no single true explanation) then meta-analyses can highlight other variables of concern which can lead to theory revision or refinement and lead to experiments in more meaningful directions. Meta-analysis is like a silver bullet for meta-analysists: the sweat and labor of the dozens of experimenters and hundreds of subjects contributes to their productivity. Should

you become one? At the moment, perhaps not. There are simply not enough replications around in the empirical software engineering literature. But if the research community could organize itself to run multi-site experiments, then we could begin on the journey the fields of medicine and psychology started some time ago. A coordinated approach would create an impetus to learn and employ meta-analytic techniques. A coordinated approach would allow raw data to be aggregated and analyzed for comparison with the meta-analytic procedures: this would provide a rich learning experience. The effort, however, would be considerable: designing and documenting an experimental package for replication is not a trivial task, never mind actually performing the study itself. It is by no means clear that such an approach is justified currently when there is so little theory to guide us and when we are perpetually concerned over the question of the validity of subject-based experimentation: we tread a tightrope and it is fairly easy to fall off into a stylized, controlled world that has no bearing whatsoever on the real world of software development and maintenance. Advocates of longitudinal studies can also make convincing arguments against short-term studies no matter how realistic the experimental situation. Leaning curves matter. CASE tools are not learnt in a day, a week, or a month.

Wait. What about all those deficiencies that single studies are susceptible too? Meta-analyses cannot possibly correct all of them. An alternative approach is possible.

Greater emphasis should be attached to exploring alternative explanations of the data by the primary investigators. These explanations are never claimed to be generalizable: they are simply explanations of the data set which may be competing or complementary. (Datamining or rule-induction tools can help find these explanations but sometimes just creating a database of all known facts can aid the experimenter in uncovering patterns of explanation.) When replications are attempted, similar care should be taken to realize alternative explanations. If replications fail to be confirming, these two sets of alternative explanations can be examined and interpretations derived explaining both sets of results. This is rather like maintaining a little breadth in the state-space of experimental exploration so that there is a built-in possibility of backtracking a little. And it is important the original researchers do this: they are closest to the scene of investigation and can even do some additional information gathering whilst they have contact with their subjects.

Any interpretations that arise, obviously have to be tested for their ability to generalize i.e., they become hypotheses for the next experiments. By taking a little extra care, the glory hunters need not be productivity fodder for the statistical meta-analysts of the future. Deny them their silver bullet.

For those actually wanting to become statistical meta-analyzers, the following two articles are recommended:

Hwang, M. 1996. The use of meta-analysis in MIS research: promise and problems. *The DATA BASE for Advances in Information Systems* 27: 35–48.

Schmidt, F. L. 1992. What do data really mean?, research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist* 47: 1173–1181.

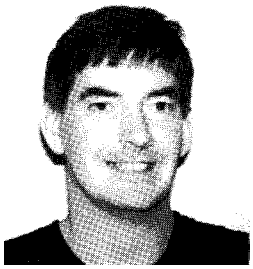
For those wishing some background on statistical developments in psychology, see:

Judd, C. M., et al. 1995. Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Reviews of Psychology* 46: 433–465.

Shaffer, J. P. 1995. Multiple hypothesis testing. *Annual Reviews in Psychology* 46: 561–584.

A good textbook discussion of statistical power can be found in:

Lipsey, M. W. 1990. *Design Sensitivity Statistical Power for Experimental Research*. SAGE Publications.



Andrew Brooks recieved the Bsc degree in Astrophysics from the University of Edinburgh in 1978, the Mphil degree in Astronomy from the University of Edinburgh in 1983, and the PhD degree in Computer Science from the University of Strathclyde in 1990. Since 1985 he has been a Lecturer in Computer Science at the Unviersity of Strathclyde, Glasgow, Scotland. In 1992, he co-founded (with Miller, Roper and Wood) the Empirical Foundations of Computer Science research group. His research interests include experimental computer science, inductive data analysis, and object-oriented programming. He is a member of the ACM, IEEE, and the British Computer Society.