

Adapting (and adopting) the experimental paradigm for Computing

David Budgen & Barbara Kitchenham

Definition

Experimental studies have provided an important exploration and ‘validation’ element for most branches of science and engineering, allowing new ideas to be developed from observations and the resulting hypotheses and theories to be tested. One of our difficulties in computing (and especially, but not uniquely, in Software Engineering) is the absence of widely accepted experimental research practices akin to (say) the epidemiological studies used in medicine, the laboratory procedures used in psychology, and of course, the practices used in the more physical areas. As a Grand Challenge we therefore propose the task of developing (and using) an **experimental framework** for computing, that will provide the vocabulary, standards, and procedures that are needed for us to persuade others of the validity of our work and its outcomes.

Background and motivation

The assessment of the success or failure of computing research is often performed by advocacy or by assertion. As objective evidence of this use of advocacy/assertion, the 1995 study of Tichy *et al.* and the 1998 work of Zelkowitz and Wallace, based on analysing the research literature in conference proceedings and journals, found both a low proportion of work using ‘independent’ experimental studies and also a wide variety in the forms of such studies as were used. More subjectively, our experience with running an annual conference on *Evaluation & Assessment* (EASE) since 1997, is that this remains a niche area of research, has limited influence on mainstream activities, and employs few agreed practices or criteria. Frameworks do exist, along with definitions, but have neither been accepted or employed very widely. Indeed, and in contrast to the other disciplines mentioned above, there is probably considerable commercial and industrial pressure in favour of having no widely accepted frameworks or definitions, and (to date) no legislation that might encourage change.

This is not to say that there is no experimental work or research into experimentation. However, one of the paradoxes of Software Engineering is that, although it extensively employs concepts and practices that stem from experience and observation (for example, *information hiding, design patterns*), there have been few studies that seek to validate their use through experimentation. On the other side of that coin though, over the last decade, we have observed that the expectation that a PhD thesis will contain a chapter on some form of reflective ‘evaluation’ has become almost the norm. So even though these evaluations might not always be conducted very systematically, the need for such an element has become quite widely recognised and accepted!

One of the problems facing the would-be experimenter (where such experimentation can involve a range of forms, from surveys to formal laboratory studies), is that there is little guidance available. There are relatively few papers, currently only one textbook (of limited usefulness), and little in the way of ‘methodological research’ to help identify when to apply particular techniques, and how to assess both the scope of their applicability and of the results. There are even fewer sources of practical ‘how to do it’ guidelines. (We both have had the experience of being asked for advice by researchers and teachers, and are particularly aware of the limited range of resources that we can identify in response to such requests.)

Our proposed *Grand Challenge* is therefore one with several elements:

- ❑ Establishing a ‘methodological framework’, backed by suitable research, to provide a vocabulary and guidance on the use of a well-defined set of techniques.
- ❑ Developing the procedures and the tools that will enable the use of the framework by computing researchers.
- ❑ Acquiring a proper understanding of the roles for experimental practices that may be most appropriate for the different branches of computing.
- ❑ Encouraging the adoption of such practices.

The criteria

Here we examine our proposed Grand Challenge against the criteria identified by the UKCRC.

- ❑ *Does it arise from scientific curiosity about the foundations, the nature, or the limits of a scientific discipline?* All other scientific and engineering disciplines have such frameworks, forming an important baseline for professional practice, and underpinned by clear philosophical principles. Admittedly, many build upon experimental paradigms where the control and repeatability of a study is much more attainable, but that is not the issue. We see our proposal as fully addressing this criterion.
- ❑ *Does it give scope for engineering ambition to build something that has never been seen before?* Since our concern is with the underpinning methodological aspects of computing, this element is not one that we address.
- ❑ *Will it be obvious how far and when the challenge has been met (or not)?* It is tempting to suggest that the criteria for ‘obvious’ will itself be an outcome of this work! We can certainly identify social measures such as the expectations of curriculum designers, programme committees and journal referees, as well as the acceptance of a ‘standard’ vocabulary. More objective measures through literature analysis can also be used.
- ❑ *Will it have enthusiastic support from (almost) the entire research community, even those who do not participate and do not benefit from it.* Our experience is that the *idea* of conducting experimental studies is generally well received, even if

the actual expectations are not always realistic, particularly from those who have not attempted such work. We also note that the draft IEEE/ACM Curriculum for undergraduate software engineering programs has included this topic in the *foundations* section with general acceptance.

- *Does it have international scope, will participation increase the research profile of a nation?* The answer here is definitely yes. Workshops on empirical studies have been appearing at a number of international conferences, including ICSE. One of us (DB) has recently led such a workshop at STEP'02 in Montreal. Indeed, this does appear to be an area of computing where the UK does have some degree of leadership in the field.
- *Is it generally comprehensible, does it capture the imagination of the general public, as well as the esteem of scientists in other disciplines?* We would argue with regard to the last part of this question that it is the *lack* of such a framework that reduces the esteem with which our work is viewed by other scientists and engineers. Indeed, there is an expectation from scientists and public that an engineering discipline of any form (which includes the various branches of computing) should be able to substantiate their claims empirically, using standards that are accepted by that discipline as a whole. Returning to the first part of the question, it is certainly comprehensible, although not particularly glamorous. However, even there, including widely used commercial products as objects of study may well attract public interest, given the pervasive nature of such software.
- *Was it formulated long ago, and still stands?* The ideas and philosophy underpinning experimental study have a long history, and we would suggest that extending and adapting these to computing is an important and critical challenge. (Does computing need a Rudolph Carnap?)
- *Does it promise to go beyond what is initially possible, and require development of understanding, techniques and tools unknown at the start of the project?* There is certainly ample scope for such developments, especially concerning tools (the STEP'02 workshop identified the lack of these as a key inhibitor).
- *Does it call for planned cooperation among identified research teams and communities?* In terms of communities, yes this is essential. There is also an important interdisciplinary element.
- *Will it encourage and benefit from competition among individuals and teams, with clear criteria on who is winning, or who has won?* Definitely not, although it might provide the criteria and measurements for such activities in other spheres. This is challenge that calls for extensive cooperation within the research community, rather than one benefiting from competition.
- *Does it decompose into identified intermediate research goals, whose achievement brings scientific or economic benefit, even if the project as a whole fails?* There are certainly beneficial intermediate goals, since the issues of vocabulary,

techniques and ‘standards’ need to be at least partially resolved before the development of significant degrees of tool support and are useful goals in themselves. Overall, there are clear scientific benefits from being able to conduct better empirical evaluations, and commercial benefit from evaluation and comparative studies.

- ❑ *Will it lead to a radical paradigm shift, breaking free from the dead hand of legacy?* Our concern is more one of breaking free from the dead hand of advocacy!
- ❑ *Is it likely to be met simply from commercially motivated evolutionary advance?* This is very unlikely. Software producers tend to be wary of unbiased and rigorous assessment practices and hence are not motivated to develop such techniques and tools. Those who might benefit (the ‘consumers’) are less likely to be in a position to conduct such studies.

Summary

We believe that the above analysis, based upon the questions posed, do indicate that our proposition has the potential to be a very important Grand Challenge, not least because it is likely to provide the means of underpinning the claims and ambitions of other Grand Challenges! As a challenge it is likely to draw upon a range of other disciplines for inspiration and validation, including philosophy, statistics, psychology and measurement theory. We would also argue strongly that it is the lack of such a framework (and its use) that continually undermines our position when we work with other engineering or scientific disciplines, or when we compete with them for resources. As a challenge it is perhaps less clearly targeted than some, which again reflects its supportive role, but we suggest that it is a very necessary part of the foundations of our discipline, and one that we should be seeking to address more comprehensively than at present.