Lessons from a Cross-domain Investigation of Empirical Practices

David Budgen & John Bailey
Department of Computer Science
Durham University, U.K.
{david.budgen,j.o.bailey}@durham.ac.uk

Mark Turner, Barbara Kitchenham & Pearl Brereton
School of Computing & Maths
Keele University, U.K.
{m.turner,b.a.kitchenham,o.p.brereton}@cs.keele.ac.uk

Stuart Charters
Applied Computing Group
Lincoln University, N.Z.
charters@lincoln.ac.nz

Abstract

Context: We are seeking the best ways to employ evidence-based practices in software engineering research and practice.

Objectives: To help assess our guidelines for conducting systematic literature reviews we have investigated how other academic disciplines use evidence-based practices.

Method: This involved performing two studies, one using a questionnaire with a set of experts, and a second using semi-structured interviews.

Results: We have identified how disciplines with similar empirical constraints to software engineering place weight upon different forms of empirical data.

Conclusions: We describe both the resulting changes to our systematic literature review guidelines and some issues this raises for empirical software engineering.

Keywords: Evidence-based, empirical studies

1. INTRODUCTION

Over the past two decades, there has been a dramatic change in medical research practices, at least for clinical studies, with the adoption of an evidence-based paradigm. In 2004, the question was posed as to whether such practices might usefully be employed in software engineering too (Kitchenham et al. 2004)? That doing so *is* practical is demonstrated by the publication of some 20 papers reporting the outcomes of conducting systematic literature reviews in software engineering over the period between 2004 and June 2007 (Kitchenham et al. 2007). Many of these referenced the initial set of guidelines that this study set out to update (Kitchenham 2004).

Indeed, the *evidence-based paradigm* has now been adopted (and adapted) by a number of other disciplines that, like software engineering, involve human-centric activities—including education, non-clinical branches of healthcare, and the social sciences. In this paper we describe how we have investigated some of the experiences and practices of these academic disciplines, and consider how these can be applied to software engineering¹.

Software engineering has a relatively weak track record of adopting empirical practices from other disciplines (Glass et al. 2004). There may be many reasons for this, including the difficulty of performing empirical studies, the limited scope that often accompanies the results, and the lack of

¹To clarify the terminology used in this paper: we use the term *domain* for the wider subject area, such as chemistry or computing; and the term *discipline* for a specific academic area, such as organic chemistry and software engineering.

overall agreement by the community on suitable empirical practices. Sometimes too, researchers may not be familiar with techniques that other disciplines regard as well-established, such as protocol analysis (Owen et al. 2006).

In this paper we describe how we have been seeking to identify and learn from those disciplines that are both adopting evidence-based practice and that also have similar problems with accumulating empirical evidence. The main motivation for this has been to see how the original guidelines for conducting systematic literature reviews in software engineering (Kitchenham 2004), based upon the standards from clinical medicine (that were all that was available when they were formulated), should be revised. Within this, a further question was whether the advice for assessing primary study quality needed to be refined.

2. EVIDENCE ACROSS DOMAINS

To understand the role of evidence, we need to recognise that, across a wide spectrum of disciplines of study, there is a common requirement to find objective practices for 'secondary' studies that can be used to aggregate the outcomes of different (primary) empirical studies in an objective, unbiased and consistent manner. This need arises in part because using conclusions derived from multiple studies gives greater confidence, and also because individual studies can be influenced by so many aspects of their context.

Aggregation, in turn, is very dependent upon the form (and quality) of the primary studies. Aggregating (say) experimental studies measuring the mass of the electron is largely a matter of using mathematically-based transformations to adjust for variations in experimental conditions; whereas drawing together the results from a set of surveys, that may have employed different sets of questions and been administered to rather different populations, presents a much less mathematically tractable problem. A key issue underlying this difference is the role of the human in the process of data collection: in the former the only involvement is as an external *observer*, while in the latter, the human is a *participant* in the treatment itself.

The area of medicine occupies an intermediate position between the two examples above. Clinical medicine (at least) is able to make extensive use of *randomised controlled trials* (RCTs) as its experimental paradigm, and in these, the role of the human is as a *recipient* of the experimental treatment. This makes it possible to adopt the use of statistical techniques for aggregation of the outcomes, and this, along with the nature of some of the outcomes, has helped to cement the success of evidence-based medicine. Software engineering also has elements of both forms, but as its characteristics make it a 'design' discipline, the human is most frequently a participant.

In (Kitchenham et al. 2004), the authors identify the following key steps in evidence-based practice for performing a secondary study in both medicine and software engineering.

1. Converting the need for information into an answerable question.
2. Tracking down the best evidence for answering that question.
3. Critically appraising that evidence for validity, impact and applicability.
4. Integrating the critical appraisal with domain expertise and the stakeholder's values.
5. Evaluating the effectiveness and efficiency in performing steps 1–4 and seeking ways to improve these.

Steps 2 and 3 are evidently major factors in the success (or otherwise) of the approach. In particular, evidence-based practice makes extensive use of *Systematic Literature Reviews* for the first three steps (Kitchenham et al. 2004, Petticrew & Roberts 2006).

Cumulative experience suggests that the first three steps are possible for software engineering, with step 4 being more problematical, although some of the support mechanisms (such as electronic databases) and their searching facilities are not as advanced as those available to many other disciplines (Brereton et al. 2007). Quality assessment (step 3) is also made more difficult by the poor reporting standards often demonstrated in empirical software engineering papers. In this study we therefore set out to answer the following questions:

1. *Which domains using evidence-based practices have empirical characteristics that are generally similar to software engineering?*
2. *What lessons can we learn from them to help revise the guidelines for software engineering?*

3. AN INITIAL COMPARISON

Clinical medicine is usually viewed as being the 'classical' model of evidence-based practice. However, medical standards make the assumption that most primary studies will be RCTs that provide comparative trials of a proposed treatment (e.g. a new drug) against either a placebo or a current standard treatment under extremely strict conditions. In particular: subjects are real patients with real diseases recruited by medical practitioners to take part in the experiment; and neither experimenters nor subjects know which treatment a subject has received (double blinding).

Such studies are virtually impossible in software engineering. We therefore sought to make an initial study of the domains that are now using the evidence-based paradigm in some way, in order to identify the ones that are most 'similar' to software engineering and hence that would merit a more detailed examination of how the academic disciplines within these are using evidence-based practices. So, while our goal for the initial study has been to be both systematic and objective, it was not intended to provide an in-depth study of all domains using evidence-based practices.

3.1. Domain Similarity Assessment

In order to develop guidelines that are suited to the type of primary study we are most likely to encounter in software engineering (i.e. 'toy' laboratory experiments—often using students, uncontrolled observational studies and non-probabilistic surveys), we need to consider the procedures for aggregating evidence that have been adopted in disciplines that exhibit experimental limitations similar to those of software engineering. To help identify such disciplines, we developed the questionnaire shown in Appendix A. (The terms used in this are defined in Appendix B, and we would note that there appears to be no 'agreed' vocabulary that can be employed for describing the different forms of empirical practice used in software engineering. In addition, we discovered that other disciplines sometimes used the vocabulary with somewhat different meanings.) Some initial findings from this phase were reported in (Budgen et al. 2006).

The resulting comparison of software engineering with other disciplines has been based on the following assumptions about the characteristics of our own domain.

- Software engineering primarily makes use of laboratory experiments, observational field studies and convenience samples. There is little use of qualitative methods and formal field experiments.
- Software engineering has major difficulties with blinding either the experimenter or the subjects as most software engineering tasks require human expertise.
- Because using a technology requires expertise, subjects need to be trained in any techniques being evaluated. This can cause bias, since it may be difficult to train people to the same level of competence in different techniques.
- In real software projects, the difficulty of tasks and the quality of materials used as inputs to tasks can have a major effect on performance. Thus, software experiments ought to randomise with respect to tasks and materials as well as participants.

3.2. Data Collection

We used our questionnaire with a selection of experienced researchers in those disciplines where we could identify that evidence-based practices were being employed. In addition, we asked a researcher from a more 'classical' science (chemistry) to also complete our questionnaire, in order to give us a baseline. We ourselves made the assessment for both software engineering and also clinical medicine.

Our initial selection of experts was drawn from Keele and Durham Universities on the basis that we were seeking general domain-related knowledge rather than specific research expertise. The

questionnaire was administered by one of the team, so that in collecting the data we would become aware of any issues that the experts felt to be relevant.

3.3. Data Analysis

To assess overall similarity we applied a simple algorithm to help us assess 'nearness'. This was based upon the values obtained for the following six major characteristics:

- use of field experiments or quasi-random experiments
- use of laboratory experiments
- use of other types of empirical study
- whether studies involve human expertise
- whether the participants can be blinded as to treatment
- whether the experimenters can be blinded to treatment

For each of these we extracted a *yes/no* value from the responses to the questionnaires. (For example, software engineering was scored as: no, yes, yes, yes, no, no.) To calculate similarity we adopted the following relatively simple linear rule-set, whereby for each contribution:

score 0 if the value is different to that for software engineering; and score 1 if they are the same, multiplying by the weight assigned for that characteristic. Then sum the values and divide by the sum of the weights.

We performed the calculation twice. The first time all characteristics were weighted as 1. The second time, the last three (relating to the 'human' element) were weighted as 2 to see what effect a greater emphasis upon these would have. This produced the results shown in Table 1.

TABLE 1: Nearness calculations by discipline

Discipline	Number of Interviewees	Weights 1	Weights 1 + 2
Clinical Medicine	—	0.17	0.11
Criminology	2	0.83	0.9
Education	1	0.83	0.9
Empirical Psychology	1	0.66	0.55
Nursing & Midwifery	1	0.83	0.9
Organic Chemistry	1	0.83	0.9
Primary Care	1	0.33	0.33
Software Engineering	—	1	1

In all cases, the effect of the weighting made relatively little difference to the outcomes beyond widening the 'gap' between those that are similar and those that are different.

Perhaps the most surprising element was the inclusion of Organic Chemistry in the set of four disciplines that could be considered as being 'near'. Closer inspection sounds a note of caution though. Like software engineering, Organic Chemistry relies heavily upon laboratory experiments. However, while this is sensible for the overall domain of chemistry, where the properties of the chemicals do not change outside of the laboratory, it is less so for software engineering, where there is a real question as to how representative laboratory experiments are of real-world software engineering activities. So the similarity in experimental practice does not reflect a meaningful similarity between the disciplines—and, while we did not consider this to have significantly affected the validity of the outcomes, this does indicate that care is needed when interpreting the results of Table 1.

4. THE DEEPER EXPLANATIONS

To gain a clearer understanding of the values in Table 1, we set out to probe more deeply into the ways that other disciplines employed evidence and performed empirical studies.

4.1. Method

To elicit expert knowledge about research practices, we decided to employ *semi-structured interviews* as being well suited for this purpose, since “the interviewees are able to speak with more detail on the issues you raise, and introduce issues of their own that they think relevant to your themes” (Oates 2006).

One of the lessons from the first study was the problem of vocabulary. Different domains (and disciplines within these) have their own vocabularies, and it may well be that what is basically the same empirical procedure is described using different terms in two disciplines. Because we administered the questionnaires ourselves, we were able to recognise where this occurred, and it did also cause us to produce the list of definitions in Appendix B. For this study we therefore agreed a detailed (six-page) specification of how the interviews were to be conducted, the issues to be probed, and the terminology to be used. A summary of this is given in Appendix C.

Interviews were usually conducted by two members of the team, one acting as interviewer, and the second as ‘recorder’. In practice, as we made recordings of all of the interviews (with permission of the interviewees), the second team member usually confined their role to making notes and checking that all of the issues had been addressed. On a few occasions, interviews were conducted by one person. We aimed to keep interviews to less than an hour overall, and they were scheduled on an *ad hoc* basis, depending on availability of interviewees.

Our original plan was to concentrate on the ‘near’ disciplines, but this was extended to include some of the health-related areas that had scored lower marks, on the basis that understanding the reasons for differences could be as important as understanding those for similarity. We also decided that the discipline of *Social Science* was close to *Criminology* and that its use was preferable as being more representative of the general domain of sociology. Table 2 lists the set of disciplines that have so far been included in this study.

TABLE 2: Pattern of semi-structured interviews

Discipline	Number of interviewees	Number of interviews
Education	2	2
Nursing	3	2
Physiotherapy	1	1
Primary Care	2	2
Public Health	2	1
Social Science	1	2

In two cases (one of the Nursing interviews and that for Public Health), we interviewed two people at once and they provided us with a joint opinion. The two interviewees classified as Primary Care were also drawn from different areas, one being concerned with practice, the other with research. We also visited two major centres involved in evidence-based practices. The EPPI-Centre (Evidence for Policy and Practice Information and Coordinating Centre), is funded by the U.K. government to conduct research into education in order to support policy-making. The CRD (Centre for Reviews and Dissemination) at the University of York, again funded by the U.K. government, has as part of its remit the task of communicating the outcomes of health-related studies to various forms of health practitioner. Both organisations provided us with further (reinforcing) information about the role of evidence-based studies in their respective domains.

4.2. Outcomes

Our description of the outcomes is organised around the major factors relating to empirical studies, rather than on a per-discipline basis.

4.2.1. *Experimental Practices*

All of the disciplines we studied used a wide range of experimental forms and selected from these on the basis of the needs of the given problem. However, there were two distinct cultural approaches that could be identified.

The first is the balance between the use of qualitative and quantitative forms. For both Education and Social Science there seemed to be a distinct bias towards the use of qualitative forms of study, reflecting their emphasis upon such 'internal' issues as perception. The more health-related disciplines, while using a mix of forms too, clearly saw the RCT as a 'gold standard' to aim for, even if it was rarely a practical option. In contrast, it was suggested that many Social Scientists were actively opposed to the use of quantitative forms (we do not have space to address the *positivist* versus *interpretivist* debate here, beyond noting that our approach in this paper is essentially positivist (Oates 2006)).

The second approach was mentioned explicitly in two of the healthcare interviews, and embraced the idea of making sequential use of different experimental forms, eventually building up to a full RCT. One example given was to begin with field observation, via (say) a case study, followed by qualitative studies (to give fuller understanding), then a quantitative pilot study and finally an RCT. It was suggested that this might typically take 15 years or more—a point that perhaps reflects the continuity of the subject matter when contrasted with such disciplines as social science, education and software engineering.

Two aspects of RCTs that few of the disciplines could emulate were randomisation and blinding. While allocation of treatments could sometimes be performed randomly, the selection of subjects rarely could. Equally, the blinding of subjects (recipients of the treatment) was rarely practical (or necessarily ethical), but blinding of the data collection process could sometimes be achieved. (One interviewee referred to "quasi-randomised controlled trials" for this type of context.)

Overall, there was a strong sense that the healthcare-related disciplines were generally seeking to get as near to the 'ideal' of an RCT as possible, at least for intervention studies. All put emphasis upon the importance of 'field studies' as opposed to laboratory experiments.

4.2.2. *Experimental Quality*

There was very little support for (and some opposition to) the idea of there being a hierarchy of experimental forms—a concept originally formulated for clinical medicine. The common view was that it was more important to match the type of study to the needs of the problem. For some of the health-related disciplines (Physiotherapy and Primary Care) RCTs were seen as the 'gold standard', at least for intervention studies, largely because of the potential for rigour and the quantitative outcomes, but the view was that the use of these was still unusual.

Within the healthcare domain, the Cochrane guidelines² for assessing the quality of studies were specifically mentioned by several researchers, although there were clearly guidelines available from many different organisations.

We asked about the use of meta-analysis for aggregating the data from primary studies, but there appeared to be very little opportunity to use this. The variations between study forms on the one hand, and study context on the other, seemed to make this impractical for all but a few studies involving interventions. Narrative reviews were therefore quite common. However, it was suggested by one researcher that as the 'evidence base' grows, and with repositories such as that maintained by the Cochrane researchers making data more widely available, there may be more scope to perform meta-analyses in the future, at least for some forms of study.

²The *Cochrane Collaboration* is a not-for-profit organisation that co-ordinates research and dissemination related to evidence-based studies in medicine on an international basis. Web site is at www.cochrane.org

4.2.3. Undertaking Systematic Literature Reviews

Almost all of our interviewees had first-hand experience of conducting a systematic literature review.

Two issues that arose in several of the interviews were the quality of reporting of primary studies, and the need for the researcher to acquire searching skills when accessing electronic databases. Relating to the former, it was observed that the Cochrane Collaboration expect reviewers to track down the source data, and so performing a review may well involve contacting the original authors of a paper! Most of the interviewees performed their own searching of electronic databases, but almost everyone noted this as a skill that had to be learned, and it was clearly felt that many researchers did not realise this. One interviewee observed that it was necessary to formulate different versions of the search terms for use with specific electronic databases.

The development and reviewing of research questions was only mentioned for Nursing. Here there was a Cochrane group that would help with scrutinising research questions and provide peer review of a research protocol.

4.2.4. Dissemination

In all domains, the publication of reviews in academic journals was seen as an important mechanism for quality control, with an alternative (especially for fully detailed reports) being repositories such as that maintained by the Cochrane Collaboration. Beyond this, there are clearly two distinct 'target' audiences: *Policy-makers*, concerned with *strategic* issues; and *Practitioners*, making decisions about individual cases.

As might be expected, disciplines such as Education, Social Science and (partly) Public Health were largely concerned with policy questions and therefore journal publication was seen as their main focus. One caveat noted was to the effect that policy-makers were more concerned to seek *ideas* rather than evidence. It was also observed that policy-makers were more receptive to results that agreed with their plans! One other problem identified that relates to providing input to policy-making was that many reviews in areas such as Education do not produce "robust conclusions" because of a shortage of adequate evidence.

In contrast, Nursing, Primary Care, Physiotherapy and (in part) Public Health saw the needs of the practitioner as their main dissemination goal. This second group made use of a wider variety of forms, including professional journals/magazines, web sites (again though, with quality caveats about exactly which sites), published abstracts, and the repositories provided by Cochrane and Campbell³. Word of mouth was also seen as important. A note of caution was added by one of the primary care interviewees, who pointed out that the better practice came at a price, and that reading, discussion and dissemination did impose an overhead on practitioners (a figure of up to 30% of extra time not facing patients was considered as a reasonable estimate).

4.2.5. Other issues identified

The main issue that we identified, over and above those concerned with experimental practices in general, was that of who funded evidence-based research. The main cost of a systematic literature review is time, so for small studies at least it is possible to conduct such a study without external funding. Beyond that, funding tended to be either from government bodies with policy-making responsibilities (especially in Education and Social Science, where contract-driven research provided the main stimulus) or from charities. Research councils were mentioned, but not seen as a major funder.

Several of the healthcare-related disciplines get funding from NICE (National Institute for Clinical Excellence), a U.K. government body concerned with determining (among other things) the forms of treatment that can be funded by the health service. Also, concern about possible bias arising

³www.campbellcollaboration.org

from the influence of funders was raised—although the view seemed to be more that where results did not suit, they might be ignored. One interviewee also observed that policy was more likely to be evidence-*informed* rather than evidence-*based*.

5. DISCUSSION

For this section we concentrate on what was learned, and assessing any threats to validity.

5.1. What lessons can we learn?

The initial analysis, as presented in Section 3, assessed 'nearness' in terms of the experimental context alone. While this obviously affects primary studies, secondary studies such as systematic literature reviews do introduce some other factors, as is evident from the analysis presented in Section 4. Perhaps the most important of these factors, in terms of its influence upon evidence-based practice within a discipline, is that of the target audience.

When we consider the *users* of evidence then we see a clear divide between those disciplines that chiefly provide evidence to inform policy-makers (most notably Education and Social Science) and those that aim to inform both policy-makers and practitioners—or even practitioners alone. (As an aside, we only interviewed one 'practitioner', whose responsibilities included tutoring others. Some of the insights we obtained about the eventual use of evidence, how to access and assess it, suggests that the practitioner viewpoint could be a fruitful topic for study in its own right.)

The other important factor, again not considered in our first study, is the nature of the objects studied. Disciplines such as Social Science and Education study *artefacts* (often largely of their own creation) as does software engineering. In contrast, the health-related disciplines largely study phenomena related to such objects as human physiology and disease that are less likely to evolve or even disappear (although treatments may change), and hence they can often take a relatively long-term view of empirical studies.

So, to answer our first question, in terms of the audience for our studies, we are probably closer to the health-related disciplines, while in terms of the volatility of our subject matter, we are closer to Education and the Social Sciences.

For the second question, there are a number of useful lessons in Section 4 as a whole. Indeed, we have already used the raw material from this study to assist with updating the guidelines for performing systematic literature reviews in software engineering (Kitchenham & Charters 2007). Some specific issues include:

1. Our interviewees were unaware of any domain-specific standards available for use (so in that sense, our own guidelines have a pioneering role). All domains referenced medical standards (Cochrane standards) while at the same time acknowledging that these were oriented to RCTs which they usually did not use. The recent textbook taking a sociological viewpoint (Petticrew & Roberts 2006), was drawn to our attention by the EPPI Centre.
2. Study quality should not be based simply on the study type—the Australian medical standards and the CRD standards used for clinical studies introduced the concept of a hierarchy of evidence based upon the type of study, with RCTs as the highest quality study, then various types of quasi-experiment, observational (correlational) studies, finally expert opinion and laboratory studies. This concept was emphatically not supported in this study, and in our Guidelines, this was replaced by the concept of quality assessment being relative to the type of study and the need.
3. In the general absence of RCTs, there was much less emphasis upon using meta-analysis to aggregate data, with most disciplines tabulating the results and using meta-analysis if possible. Only one group had tackled the problem of trying integrate qualitative and quantitative studies in a single systematic review and we incorporated into our guidelines their recommendation to aggregate the different forms separately and then look at whether the qualitative results could be used to help explain the quantitative results(Thomas et al. 2004).

4. The concept of employing a sequence of study forms to 'open up' a topic is a useful one that should be considered further. The idea of a sequence of study forms has been previously suggested in software engineering (Linkman & Rombach 1997), although that paper proposed a progression from formal laboratory studies towards *less* formal industry studies, rather than towards *more* formal experiments conducted in industry settings.

Collectively, these certainly give greater confidence about the usefulness of the evidence-based paradigm for software engineering—notwithstanding practical issues about availability of primary study data and the like.

5.2. Threats to Validity

Both questionnaires and semi-structured interviews are well-established tools for qualitative research. We therefore argue that our construct validity is soundly based upon established practices and therefore concentrate the discussion in this section on the way that we used these techniques.

5.2.1. Internal Validity

Our main concern here is whether our conduct of the two studies was appropriate for the task. We have identified the following possible threats.

1. The number of interviewees was small (one or two per discipline) and selection was on a convenience basis, which may not have given a valid representation of the discipline. However, when using more than one interviewee (always from separate institutions) we found little that contradicted in their views. We also conducted 'sanity checks' through our contacts with various experts and with organisations such as EPPI-Centre and CRD.
2. The variations in the vocabulary used by each discipline may have led to misinterpretation on our part. However, we did take care to define our own terms (Appendix B) and to check these with the interviewees.
3. Our skills as interviewers may have affected the outcomes. Since semi-structured interviews are designed to probe into a subject, the main risk from this is that we have failed to explore an issue or have failed to identify one. However, given that interviewer and interviewee had a shared interest, and interviewees were experienced researchers, this seems a low risk.

5.2.2. External Validity

When considering the wider applicability of our study, one question must be how much it was influenced by being conducted within a single country (the U.K.). However, all of these disciplines have strong international interactions, and it seems unlikely that issues such as empirical practices are significantly affected by the overall funding context.

Many of the disciplines we examined depend largely on public funding—whereas software engineering, rather like clinical medicine, has a mix of proprietary and public development (industry and open source respectively). However, while this may affect the availability of data there seems no reason to consider that it alters the relevance of research practices.

6. CONCLUSIONS

The main motivation for this study was to update the initial guidelines for performing systematic literature reviews in software engineering (Kitchenham 2004). Such guidelines have been important in clinical medicine and the interviewees did widely refer to them. Indeed, many of the reviews identified in (Kitchenham et al. 2007) used the original software engineering guidelines.

The guidelines (Kitchenham & Charters 2007) have been quite extensively updated and extended, partly as a result of the findings from this, and other studies, and also by consulting the most recent

sociological textbooks on this topic⁴. As examples, there are new references from domains other than medicine, the concept of a hierarchy of experimental forms has been removed, and there is better guidance on how to employ quality criteria that are tailored to different empirical forms, including qualitative ones.

We also identified some ideas that will merit further exploration—including the question of how to persuade others of the value of random and quasi-random field experiments when convincing policy-makers and practitioners of the merits of particular practices or technologies. While the proprietary nature of software development poses a challenge here, it is one that collectively we do need to address.

7. ACKNOWLEDGEMENTS

The contributions of all the experts who allowed us to interview them are gratefully acknowledged. We would also like to recognise the support of the U.K.'s EPSRC for funding the project conducting this study as part of the Project EP/C51839X/1 (Evidence-Based Software Engineering).

REFERENCES

- Brereton, O., Kitchenham, B., Budgen, D., Turner, M. & Khalil, M. (2007), 'Lessons from applying the Systematic Literature Review process within the Software Engineering domain', *Journal of Systems & Software* **80**(4), 571–583.
- Budgen, D., Charters, S., Turner, M., Brereton, P., Kitchenham, B. & Linkman, S. (2006), Transferring the evidence-based paradigm to software engineering, in 'Proceedings of WISER Workshop, ICSE 2006', ACM Press, pp. 7–13.
- Glass, R., Ramesh, V. & Vessey, I. (2004), 'An Analysis of Research in Computing Disciplines', *Communications of the ACM* **47**, 89–94.
- Kitchenham, B. (2004), Procedures for undertaking systematic reviews, Technical Report TR/SE-0401, Department of Computer Science, Keele University and National ICT, Australia Ltd. Joint Technical Report.
- Kitchenham, B., Brereton, P., Budgen, D., Turner, M., Bailey, J. & Linkman, S. (2007), A systematic literature review of evidence-based software engineering. Accepted for publication in *Information & Software Technology*.
- Kitchenham, B. & Charters, S. (2007), Guidelines for performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.
- Kitchenham, B., Dybå, T. & Jørgensen, M. (2004), Evidence-based software engineering, in 'Proceedings of ICSE 2004', IEEE Computer Society Press, pp. 273–281.
- Linkman, S. & Rombach, D. (1997), 'Experimentation as a vehicle for software technology transfer—a family of software reading techniques', *Information & Software Technology* **39**(11), 777–780.
- Oates, B. (2006), *Researching Information Systems and Computing*, SAGE.
- Owen, S., Budgen, D. & Brereton, P. (2006), 'Protocol Analysis: A Neglected Practice', *Communications of the ACM* **49**, 117–122.
- Petticrew, M. & Roberts, H. (2006), *Systematic Reviews in the Social Sciences: A Practical Guide*, Blackwell Publishing.
- Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., Brunton, G. & Kavanagh, J. (2004), 'Integrating qualitative research with trials in systematic reviews', *British Medical Journal* **328**, 1010–1012.

A. QUESTIONNAIRE TO DETERMINE SIMILARITY TO SOFTWARE ENGINEERING

1. Does the domain undertake the following types of empirical study:

- Field experiments (with random allocation)
Frequently (4)/ Occasionally (3)/ Rarely (2)/ Never (1)

⁴A copy of the guidelines can be obtained from www.ebse.org.uk.

- Quasi-random field experiments
Frequently (4)/ Occasionally (3)/ Rarely (2)/ Never (1)
 - Field observational studies (including case studies)
Frequently (1)/ Occasionally (2)/ Rarely (3)/ Never (4)
 - Laboratory experiments
Frequently (1) /Occasionally (2)/ Rarely (3)/ Never (4)
 - Surveys based on random sampling of a population
Frequently (4)/ Occasionally (3)/ Rarely (2)/ Never (1)
 - Surveys based on convenience samples
Frequently (1)/ Occasionally (2)/ Rarely (3)/ Never (4)
 - Document analysis (Content analysis)
Frequently (1)/ Occasionally (2)/ Rarely (3)/ Never (4)
 - Discourse analysis
Frequently (1)/ Occasionally (2)/ Rarely (3)/ Never (4)
2. Is human expertise a factor in techniques, tools or methods that are evaluated in empirical studies?
Always (1) /Often (2)/Sometimes (3)/Rarely (4)/Never (5)
 3. Is it possible to blind subjects as to the techniques, tools, methods that are evaluated in empirical studies?
Always (5)/Often (4)/Sometimes (3)/Rarely (2)/Never (1)
 4. Do experimenters use procedures to blind experimenters as to the allocation of subjects to the techniques, tools, methods that are evaluated in empirical studies?
Always (5)/Often (4)/Sometimes (3)/Rarely (2)/Never (1)
 5. Is subject training a significant cause of potential bias?
Always (1)/Often (2)/Sometimes (3)/Rarely (4)/Never (5)
 6. Is randomisation required for materials, tasks and settings as well as subjects for valid generalisation?
Always (1)/Often (2)/Sometimes (3)/Rarely (4)/Never (5)

B. DEFINITIONS

The terms used in the questionnaire are defined below.

- **Field experiments with random allocation** are formal experiments undertaken in a software engineering company by software engineers performing standard project activities (i.e. excluding experiments taking place as part of training exercises). Participants are allocated to the software engineering method at random. This would be equivalent to a **Randomised Controlled Trial** (RCT) in medicine.
- **Quasi-random field experiments** are studies where randomisation is not possible for practical or ethical reasons. They are used when different groups of interest occur naturally and no random assignments can be performed, such as successful or unsuccessful projects.
- **Field observational studies** are observational studies undertaken in a software engineering company. They are neither experiments nor quasi-experiments and may be qualitative rather than quantitative. These include correlation and regression studies such as cost estimation studies.
- **Laboratory experiments** are randomised experiments performed by subjects (whether students or practitioners) that are not part of a real software engineering project.
- **Surveys** are questionnaire based or interview based exercises where opinions or data are extracted from a relatively large number of participants. Surveys based on random sampling are based on having a defined population and ensuring that each member of the population has an equal chance of being requested to participate. Convenience sampling is based on non-random sampling. Participants are those people who choose to take part.
- **Document analysis**, and in particular, **content analysis**, is a semi-qualitative method that codes words or phrases in a document and counts the number of incidences of the coded phases/words.
- **Discourse analysis** is a range of methods relating to a theoretical position that are applied to different forms of discourse, including interviews.

C. SEMI-STRUCTURED INTERVIEWS

As the basis for conducting these we developed a six-page set of guidelines. These described the process to be followed by the interviewer; provided a pro-forma note ('Pre-Interview Information') to be supplied to the interviewee in advance that spelt out our goals; a guide for the interviewer that included the main headings to be covered and some sample questions for these; and some notes on empirical forms to help with clarifying terminology.

The main headings for the interview guide were as follows (the main topics for the questions are included in brackets).

- Introduction (purpose of interview)
- Subject Context (the discipline; forms of phenomena examined; role of experimental studies)
- Experimental Practices (effect of human expertise; forms of experimental practice; how a typical experimental study would be conducted; recruitment of subjects; roles of blinding or randomisation)
- Experimental Quality (how different experimental forms are ranked; what the ranking is; highest and most common forms; how quality is assessed in reports; techniques for meta-analysis used)
- Systematic Literature Reviews (how used in the discipline; personal experience of the interviewee; infrastructure support; who uses the outcomes; problems with undertaking or using outcomes; how co-ordinated; existence of accepted guidelines; forms of publication; sources of funding for reviews)
- Dissemination (how published; format of publications; how end users are made aware of reviews)
- Closing issues (recap of discussion; snowball questions to follow up issues identified)