Empir Software Eng (2008) 13:97–121
DOI 10.1007/s10664-007-9053-5

Evaluating guidelines for reporting empirical software engineering studies

Barbara Kitchenham · Hiyam Al-Khilidar · Muhammed Ali Babar · Mike Berry · Karl Cox · Jacky Keung · Felicia Kurniawati · Mark Staples · He Zhang · Liming Zhu

Published online: 17 October 2007

Editor: Jose Carlos Maldonado

Abstract
Background Several researchers have criticized the standards of performing and reporting empirical studies in software engineering. In order to address this problem, Jedlitschka and Pfahl have produced reporting guidelines for controlled experiments in software engineering. They pointed out that their guidelines needed evaluation. We agree that guidelines need to be evaluated before they can be widely adopted.

M. Ali Babar was working with the National ICT Australia when the reported work was performed.

H. Al-Khilidar · M. Berry · K. Cox · J. Keung · F. Kurniawati · M. Staples · H. Zhang · L. Zhu
National ICT Australia Ltd, Sydney 1466 NSW, Australia

K. Cox
e-mail: karl.cox@nicta.com.au

J. Keung
e-mail: jkeung@cse.unsw.edu.au

F. Kurniawati
e-mail: felicia.kurniawata@nicta.com.au

M. Staples
e-mail: mark.staples@nicta.com.au

L. Zhu
e-mail: liming.zhu@nicta.com.au

M. A. Babar
Lero, The Irish Software Engineering Research Centre, University of Limerick, Limerick, Ireland
e-mail: malibaba@lero.ie

H. Al-Khilidar · M. Berry · J. Keung · H. Zhang
School of Computer Science & Engineering, University of New South Wales,
Sydney 2052 NSW, Australia

B. Kitchenham (✉)
Keele University, School of Computing and Mathematics, Keele, Staffordshire, ST5 5BG, UK
e-mail: Barbara.Kitchenham@cs.keele.ac.uk

Aim The aim of this paper is to present the method we used to evaluate the guidelines and report the results of our evaluation exercise. We suggest our evaluation process may be of more general use if reporting guidelines for other types of empirical study are developed.
Method We used a reading method inspired by perspective-based and checklist-based reviews to perform a theoretical evaluation of the guidelines. The perspectives used were: Researcher, Practitioner/Consultant, Meta-analyst, Replicator, Reviewer and Author. Apart from the Author perspective, the reviews were based on a set of questions derived by brainstorming. A separate review was performed for each perspective. The review using the Author perspective considered each section of the guidelines sequentially.
Results The reviews detected 44 issues where the guidelines would benefit from amendment or clarification and 8 defects.
Conclusions Reporting guidelines need to specify what information goes into what section and avoid excessive duplication. The current guidelines need to be revised and then subjected to further theoretical and empirical validation. Perspective-based checklists are a useful validation method but the practitioner/consultant perspective presents difficulties.
Categories and Subject Descriptors K.6.3 [*Software Engineering*]: Software Management—Software process.
General Terms Management, Experimentation.

Keywords Controlled experiments · Software engineering · Guidelines · Perspective-based reading · Checklist-based reviews

1 Introduction

This paper reports an exercise undertaken by staff and students in the Empirical Software Engineering (ESE) group at NICTA (National ICT Australia) to evaluate the reporting guidelines for controlled experiments proposed by Jedlitschka and Pfahl (2005). In spite of the existence of a specialist book to help software engineers conduct experiments (Wohlin et al. 2000), software engineering experiments are still subject to criticism. The guidelines were developed in response to general criticisms of current standards of performing and reporting empirical studies (Kitchenham et al. 2002), and more specific criticisms that the lack of reporting standards is causing problems when researchers attempt to aggregate empirical evidence because important information is not reported or is reported in an inconsistent fashion (e.g. Pickard et al. 1998; Wohlin et al. 2003).

In fact, controlled experiments are performed infrequently in software engineering. In a recent survey of 5,453 software engineering articles from 12 leading conferences and journals, Sjøberg et al. (2005) found only 103 articles that could be categorized as experiments. However, there is evidence that current reporting practice is inadequate. Dybå et al. (2006) had to exclude 21 experiments from their analysis of power because the authors did not report enough information for a power analysis. Authors did not report any statistical analysis for 14 experiments and in seven cases the experiments were so badly documented that Dybå et al. "did not manage to track which tests answered which hypothesis or research question". This result confirms the need for reporting guidelines for software engineering experiments.

Jedlitschka and Pfahl recognised that their guidelines need to be evaluated, saying:

> "Our proposal has not yet been evaluated e.g. through peer review by stakeholders, or by applying it to a significant number of controlled experiments to check its usability. We are

aware that this proposal can only be the first step towards a standardized reporting guideline." (Jedlitschka and Pfahl 2005)

We agree with the need for guidelines to be evaluated. If the guidelines are themselves flawed, they could make the problem of poor quality reporting worse than it is currently.

Our evaluation exercise took place between 5th October 2005 and December 14th 2005. It was organized as a series of eight working meetings each taking between 1 and 2.5 h. In this paper, we report the evaluation method we used and the results of our evaluation. We have already reported our results to Jedlitschka and Pfahl, so the main purpose of this paper is to report our evaluation method, since it might prove useful to other groups wanting to evaluate the next version of the reporting guidelines or future reporting guidelines for other forms of empirical study such as case studies, surveys, or systematic reviews.

In Section 2 we give a brief overview of the proposed guidelines. In Section 3 we discuss the various options available for evaluating experimental guidelines and provide a rational for our choice of perspective-based reviews. In Section 4 we report our evaluation process. In Section 5 we report our evaluation results. In section 6 we discuss our results.

An earlier version of this paper was presented at ISESE06 (Kitchenham et al. 2006). In this paper, we have extended the report of our evaluation exercise to include:

- A more detailed discussion of our evaluation making it clear that we have adopted a method based on perspective-based checklists.
- Consideration of the advantages of the guidelines. We identify the questions in each perspective that were addressed by the guidelines.
- The full list of amendments classified according to amendment type.
- A list of questions that are applicable to all (or most) perspectives. This will enable other users of this evaluation method to separate general questions from perspective specific questions.

2 Proposed Reporting Guidelines

Jedlitschka and Pfahl (2005) propose the reporting structure for experiments shown in Table 1. Table 1 identifies the recommended section and subsection headings in a report of an experiment together with a brief description of the information required in each section and a cross reference to the subsection in the guidelines that discusses the information that authors should supply in each section.

3 Evaluation Options

At our first working meeting, we discussed various theoretical and empirical evaluation methods and considered the viability of each type. Theoretical evaluation can be based on several different approaches:

- T1. An assessment of each element in the guidelines from the viewpoint of why the element is included in the guidelines; what it is intended to accomplish in terms of supporting readers to find the information they are looking for; and what evidence there is to support the view that the element is important.

Table 1 Proposed reporting structure

Section heading	Subsection heading	Contents	Cross-reference
Title			
Authorship			
Structured Abstract		Summarises the paper under headings of Background or Context, Objectives or Aims, Method, Results, and Conclusions	3.1
Motivation		Sets the scope of the work and encourages readers to read the rest of the paper	3.2
	Problem Statement	Reports what the problem is; where it occurs, and who observes it	3.2.1
	Research Objectives	Defines the experiment using the formalized style used in GQM	3.2.2
	Context	Reports environmental factors such as settings and locations	3.2.3
Related work		How current study relates to other research	3.3
Experimental design		Describes the outcome of the experimental planning stage	3.4
	Goals, Hypotheses and Variables	Presents the refined research objectives	3.4.1
	Design	Define the type of experimental design	3.4.2
	Subjects	Defines the methods used for subject population sampling and group allocation	3.4.3
	Objects	Defines what experimental objects were used	3.4.4
	Instrumentation	Defines any guidelines and measurement instruments used in the experiment	3.4.5
	Data Collection Procedure	Defines the experimental schedule, timing and data collection procedures	3.4.6
	Analysis Procedure	Specifies the mathematical analysis model to be used	3.4.7
	Evaluation of Validity	Describes the validity of materials, procedures to ensure participants keep to the experimental method, and methods to ensure the reliability and validity of data collection methods and tools	3.4.8
Execution		Describes how the experimental plan was implemented	3.5
	Sample	Description of the sample characteristics	3.5.1
	Preparation	How the experimental groups were formed and trained	3.5.2
	Data Collection Performed	How data collection took place and any deviations from plan	3.5.3
	Validity Procedure	How the validity process was followed and any deviation from plan	3.5.4
Analysis		Summarizes the collected data and describes how the data was analyzed	3.6
	Descriptive statistics	Presentation of the data using descriptive statistics	3.6.1
	Data set reduction	Describes any reduction of the data set e.g. removal of outliers	3.6.2
	Hypothesis testing	Describes how the data was evaluated and how the analysis model was validated	3.6.3

Table 1 (continued)

Section heading	Subsection heading	Contents	Cross-reference
Interpretation		Interprets the findings from the Analysis section	3.7
	Evaluation of results and implications	Explains the results	3.7.1
	Limitations of Study	Discusses threats to validity	3.7.2
	Inferences	How the results generalize given the findings and limitations	3.7.3
	Lesson learnt	Descriptions of what went well and what did not during the course of the experiment	3.7.4
Conclusions and Future work		Presents a summary of the study	3.8
	Relation to Existing Evidence	Describes the contribution of the study in the context of earlier experiments	3.81
	Impact	Identifies the most important findings with respect to cost, time and quality	3.8.2
	Limitations	Identifies main limitations of approach i.e. circumstances when the expected benefits will not be delivered	3.8.3
	Future work	Suggestions for other experiments to further investigate the research question	3.8.4
Acknowledgements		Identifies any contributors who do not fulfill authorship criteria	3.9
References		Lists all cited literature	3.10
Appendices		Includes raw data and/or detailed analyses which might help others to use the results	3.11

- T2. A review of the process by which the guidelines were constructed identifying the validity of the source material, the aggregation of the source material, and the evaluation process.
- T3. Reading the guidelines in order to detect defects and areas for improvement taking the viewpoint of different roles that might want to read a report of a software experiment (i.e. a form of perspective-based reading).
- T4. Mapping any established experimental methodology guidelines to the reporting guidelines.

Empirical evaluation can be based on a variety of possible approaches, for example:

- E1. Take a sample of published articles reporting experiments constructed without support of the guidelines and identify whether important information has been omitted from the articles that would have been included if the guidelines had been followed. This is similar to the approach taken by Moher et al. (2001) who compared papers in journals that used the CONSORT guidelines with those that did not. The objection to this approach is that the guidelines being evaluated are the basis for their own evaluation.
- E2. Take a sample of published articles reporting experiments and re-structure them to conform to the guidelines. Then use the duplicate versions as the experimental

material in an experiment aimed at evaluating whether the guidelines make it easier a) to understand the papers and/or b) to extract standard information from the papers.

When deciding which evaluation process to undertake, we considered:

1. Whether the evaluation approach itself was valid i.e. likely to lead to a trustworthy assessment of the strengths and weaknesses of the guidelines.
2. Whether the evaluation approach was feasible given our resources (effort, time and people).
3. Whether the approach was cost effective given the value of the proposed guidelines. We noted that formal experiments are currently not often used in software engineering research. It is possible that industry case studies and surveys might be more relevant.
4. Whether the approach provided a good learning opportunity for our research group. This was an important issue because the group included PhD students who were learning about empirical software engineering.

After evaluating each approach, as summarized in Table 2, we concluded that an evaluation based on reading the guidelines in order to detect defects and areas for improvement (T3) would be the most appropriate evaluation method for us to undertake. We felt that empirical evaluation was extremely problematic. Experiments based on re-writing existing papers would be too difficult for a group including novice researchers. It would also be biased if information required by the guidelines was not available in the original papers. Of the theoretical evaluation methods, we felt the perspective-based reading approach would provide the best learning opportunity for the PhD students and junior researchers, giving them an opportunity to consider the needs of different readers and discuss, with more experienced researchers, how to meet those needs. We chose this evaluation approach to suit our own pedagogical purposes. It is not our intention to claim that it is inherently better than the other theoretical approaches nor to suggest that the other evaluation methods should not be used. All the theoretical evaluation methods are valuable and could be used together as part of a comprehensive evaluation program.

4 Applying Perspective-based Reading to Evaluating the Experimental Guidelines

In this section we discuss the process we used to evaluate the experimental guidelines. Our evaluation process was organized as a series of eight meetings each of which took between 1 and 2.5 h and took place between 5th October and 14th December 2005 with a maximum of one meeting a week. The results of each meeting were documented after each meeting to provide feedback to participants. The meeting schedule is shown in Table 17 in the Appendix.

Table 2 Assessment of evaluation methods

Evaluation method	Type	Valid	Feasible for us	Cost effective	Learning potential
T1. Evaluation of each element	T	Yes	Yes	Yes	No
T2. Evaluation of process	T	Yes	Yes	Yes	No
T3. Perspective-based Reading	T	Yes	Yes	Yes	Yes
T4. Mapping to existing guidelines	T	Yes	Yes	Yes	No
E1. Review of existing papers	E	No	Yes	No	Yes
E2. Formal experiment	E	Yes	No	No	No

The type of approach: Theoretical (T) or Empirical (E)

4.1 Evaluation Process

The first issue we considered was how to apply perspective-based reading to the goal of evaluating the guidelines. Conventional perspective-based reading is intended to assist reviewing software artifacts from the viewpoint of stakeholders such as the customer, the designer, or the tester who will use the artifact, see for example Shull et al. (2000). Reviewers taking a particular perspective consider a scenario describing how they will use the artifact and ask questions derived from that scenario. For example, Shull et al. described a tester reviewing a requirements document. The tester is required to generate a test or set of test cases that allow him/her to ensure that the system implementation satisfies the requirements. The tester then answers a number of questions related to the test case generation task.

For our evaluation, it was clear that there were different perspectives related to reading a report of a software experiment and that different perspectives would require different information from the report. However, it was not clear that we could develop appropriate operational scenarios to match perspectives, because we were not intending to review a specific experimental report, we were reviewing guidelines intended to assist writing a report. For this reason we decided to base our review of the guidelines on a checklist of questions related to the information required by each perspective. Thus we ended up applying a hybrid reading method using perspective-based checklists.

We also departed significantly from the standard review process. Instead of having a single review meeting with each reviewer taking a different perspective, we decided to undertake a series of reviews where each review addressed a single perspective. We chose this approach because of the learning opportunities implicit in this process. Assigning individual perspectives to each reviewer would have been more efficient, but it may not have ensured that the same level of scrutiny was given to each perspective.

4.2 Identification of the Relevant Perspectives

Our first step was to identify which perspectives we would incorporate into our evaluation process. We identified the following perspectives of interest:

- *Researcher* who reads a paper to discover whether it offers important new information on a topic area that concerns him or her.
- *Practitioner/consultant* who provides summary information for use in industry and wants to know whether the results in the paper are likely to be of value to his/her company or clients.
- *Meta-analyst* who reads a paper in order to extract quantitative information that can be integrated with results of other equivalent experiments.
- *Replicator* who reads a paper with the aim of repeating the experiment.
- *Reviewer* who reads a paper on behalf of a journal or conference to ensure that it is suitable for publication.
- *Author* who would be expected to use the guidelines directly to report his/her experiment.

We also identified the perspective of the editorial board of journals (or the program committee of conferences) who might choose to adopt reporting guidelines. The adoption or not of a set of international guidelines could have both good and bad impacts:

- It might suggest to authors that there is a fast track to publication or acceptance by using the guidelines irrespective of the quality of the paper.

- It might discourage authors of non-experimental studies from submitting to the journal.
- It might improve the quality of papers.
- It might improve the quality of reviews.

However, although we believe the perspective of an editorial board is important, we did not think it was one that we could realistically adopt.

For each perspective, we used brainstorming to assess what an individual with each perspective would require from a paper and converted these issues into a number of questions that summarize the issues of importance to each perspective. The checklists we developed for Researcher, Practitioner/Consultant, Meta-Analyst, Replicator and Reviewer

Table 3 Researcher checklist

Number	Question	Rationale
Res-1	Is the paper *easy to find*?	Researchers need to find potentially relevant research results
Res-2	Is it a *relevant* paper?	Researchers need to identify quickly whether an article is relevant to his/her research
Res-3	Is the *overall structure* of the paper appropriate?	Researchers need to find easily specific pieces of information within a paper
Res-4	Is the *research problem hypothesis easy to identify*?	Researchers need to be sure what hypothesis is being tested
Res-5	Is there an *underlying causal model*? If so, what is it?	It is important to know whether the research was derived from an underlying model and what it is
Res-6	Is the *terminology defined* and explained?	All specialized terminology needs to be defined
Res-7	Is the level of *assumed knowledge excessive*?	Junior researchers and researchers from other fields need sufficient explanation to follow the paper, or at least need to
Res-8	Is required *background knowledge referenced*?	be directed to text books or reference articles where they can obtain background information
Res-9	Is the research related to other *relevant research*?	Researchers need to know what the state of knowledge was prior to the experiment and how the current experiment contributes to new knowledge
Res-10	Is the *experimental design appropriate*?	Researchers need to know whether the experiment was capable of properly testing the hypothesis
Res-11	Is the *statistical analysis correct*?	Researchers need to be sure that the analysis was performed correctly
Res-12	Is the *raw data available*?	Researchers should be able to replicate the analysis or investigate alternative analysis methods. In order to do that the raw data should either be published in the paper or stated to be available on request
Res-13	Is it easy to identify the findings / *results* of the experiment?	Researchers need to know what the results of the experiment were
Res-14	Do the *conclusions arise from the results*?	Researchers need to be sure that the conclusions arise from the reported research results
Res-15	Is the *argumentation clear*?	Researchers need to be sure that any claims made in the paper (such as generalizations) are clearly linked to evidence which supports those claims
Res-16	Are *limitations* of the experiment made clear?	Researchers need to know the limitations, risks and constraints that apply to the experiment and the conclusions
Res-17	Is there any discussion of required *further research*?	Researchers need to know what still needs to be investigated

Table 4 Practitioner/consultant checklist

Number	Question	Rationale
P-1	Is the paper *easy to find*?	Consultants need to be able to find relevant research results
P-2	Is it a *relevant* paper?	Consultants should be able to identify quickly whether or not an article is relevant to their requirements
P-3	What does the paper *claim*?	Consultants need to identify exactly what claims the paper makes about the technology of interest
P-4	Are the *conclusions/results useful*?	Consultants need to know whether the conclusions/results have practical relevance
P-5	Is the *claim supported by believable evidence*?	Consultants need to be sure that any claims are supported by evidence
P-6	Is it clear how the current research relates to *existing research* topics and trends?	Consultants need to know how the current work relates to existing research trends
P-7	How can the *results be used in practice*?	Consultants need guidance on how the results would be used in industry
P-8	In what *context* is the result/claim useful/ relevant?	Consultants needs to know the context in which the results are expected to be useful
P-9	Is the *application type* specified?	Consultants need to know what type of applications the results apply to. In particular whether they are specific to particular types of application (e.g. finance, or command and control etc.)
P-10	Is the availability of required *support environment* clear?	Consultants need to know whether any required tool support is available and under what conditions
P-11	Are any *technology pre-requisites* specified?	Consultants need to know whether there are any technological prerequisites that might limit the applicability of the results
P-12	Are the *experience* or *training costs* required by development staff defined?	Consultants need to know the training/ experience requirements implicit in the approach
P-13	Is the *expense* involved in adopting the approach defined?	Consultants need some idea of the cost of adopting the approach, in order to perform return on investment (ROI) analyses
P-14	Are any *risks* associated with adoption defined?	Consultants need to know whether there are any risks associated with adoption of the technique
P-15	Do the results *scale to real life*?	Consultants need to be sure that the results scale to real life
P-16	Is the experiment based on concrete *examples of use/application* or only theoretical models?	Consultants need to be sure that the results have a clear practical application
P-17	Does the paper discuss existing technologies, in particular the *technologies it supersedes* and the *technologies it builds on*?	Consultants need to be sure that the experiment involves comparisons of appropriate technologies. They need to know that a new approach is better than other equivalent approaches not a “straw man”
P-18	Is the new approach, technique, or technology *well described*?	Consultants must be sure that they understand the new approach/technique/technology well enough to be able to adopt it

Table 4 (continued)

Number	Question	Rationale
P-19	Does the paper make it clear who is funding the experiment and whether they have any *vested interests*?	Consultants need to be sure that the experiment is as objective as possible
P-20	Does the paper make it clear what *commitment* is required to adopt the technology?	A consultant needs to know whether adoption of an approach/technology requires a complete and radical process change or can be introduced incrementally
P-21	Are *Technology Transfer issues* discussed?	Consultants need to know what the objections to a new technology are likely to be, and whether there are any clear motivators or de-motivators
P-22	Is there any discussion of *required further research*?	Consultants need to know whether the research is complete or the approach needs further development

are shown in Tables 3, 4, 5, 6 and 7 respectively. Since the tables are rather long, the main keywords for the questions are shown in italics to assist readability. For the Researcher and Practitioner/Consultant perspective we did not attempt to remove duplicate questions thinking that it was important to fully represent each perspective. After applying both of these perspectives, we developed the Meta-analyst, Replicator and Reviewer perspectives. For these perspectives, we concentrated on the main differences between each perspective and the Researcher and Practitioner/Consultant perspectives. After our experience with the first two perspectives, we realized that there would be too much redundancy in the questions if we produced a complete checklist for each perspective.

We also decided not to attempt to construct a checklist for the Author perspective since it would be too close to the Researcher perspective. Instead we decided to undertake a separate review of the guidelines where we considered each element in turn discussing whether:

- Including the information would be difficult for authors.
- The guideline element was necessary.
- Including the information would improve the paper.
- Including the information would make the paper more difficult to read or write.

Using a different approach for reviewing from the author perspective gave us the chance to address issues not raised explicitly by the perspective-based questions.

4.3 Validity of Checklist Approach

The validity of the checklist approach depends on the validity of the checklists and that, in turn, depends on the experience of the participants. Table 8 confirms that we included participants with extensive experience either in industry or academia (or both). All participants had experience of performing and reporting empirical studies of various types. Furthermore, all of the participants, except the research associate had some experience acting as reviewers and some of the participants had extensive experience. Only one researcher had experience of acting as a replicator, and only the senior researcher had experience of acting as a meta-analyst, although all the participants had some exposure to the principles of systematic literature reviews (Kitchenham 2004), which are a necessary prerequisite to performing a quantitative meta-analysis. Thus, we have some confidence in the validity of the researcher,

Table 5 Meta-analyst perspective

Number	Question	Rationale
M-1	How many *experimental units per treatment*?	The number of experimental units (subjects) is critical for meta-analysis
M-2	What was *effect size* (or mean effect for each treatment and the variance)?	The effect size is the basic datum required for meta analysis
M-3	Are *treatments/technologies* clearly *defined*?	The meta-analyst must ensure that information from different studies pertains to the same treatments so that it can be aggregated
M-4	Are the *measures* properly *defined*?	It is important to be sure that the measures used in different papers are equivalent
M-5	Is the *data collection* process *reliable*?	It is important to be sure that the measurement collection follows a rigorous process
M-6	Is the *experimental procedure* well defined?	It is important to ensure that experimental procedures are equivalent in different papers
M-7	Does the *data analysis method* match the stated experimental design?	It is important that the analysis results are correct
M-8	Are any *data transformation* or *reduction* processes reported?	A meta-analyst needs to know if and how the data has been manipulated before analysis
M-9	How are *drop outs* analyzed?	Differential drop-outs can seriously bias experimental results. The analysis protocol needs to address how drops out were handled
M-10	Were experimental units *allocated at random* to treatment conditions?	Random allocation is a basic requirement for a randomized controlled experiment (as opposed to a quasi-random experiment)
M-11	Was the *random allocation process* defined?	Unless the randomization process is reported it cannot be assumed that random allocation (as opposed to haphazard allocation) has taken place
M-12	Was *sensitivity analysis* performed?	The meta-analyst needs to know that the results are robust (i.e. not the result of one or two atypical values)
M-13	Was any form of *blinding* used?	Blinding is an essential means of reducing experimenter expectation bias. Opportunities are limited in software engineering experiments but it is sometimes possible to perform blind marking, and/or blind allocation to treatment. It is also possible to perform blind analysis (treatments are coded before data are given to the analyst)
M-14	Are any *side-effects*, or *risks* associated with the treatments defined?	It is important to be sure that any risks associated with new treatments are reported

practitioner, and reviewers checklists, but less confidence in the validity of the meta-analyst and replicator checklists. We also have a fair degree of confidence that we appreciated the issues associated with reporting empirical studies.

Another practical problem associated with our approach is that because the checklist questions were developed without direct reference to the guidelines, it is difficult to cross-reference the checklist questions to specific guidelines items. However, we think it is more important to have some degree of independence between the evaluation criteria (i.e. checklist questions) and the item being evaluated (i.e. the guidelines) than to have simple traceability between one and the other, so this problem is inherent in the basic approach.

Table 6 Replicator perspective

Number	Question	Rationale
Rep-1	Can I *contact* the authors if there are ambiguities in the description of the experiment?	Replicators need to be able to contact the experimenters if details are missing. This question is also important for meta analysts
Rep-2	Are the *hypothesis* fully defined?	Replicators may (and perhaps should) change the details of the experimental protocol. However, they must keep the same hypotheses (or they are not performing a replication)
Rep-3	Are *subject groups* clearly defined?	Whether the replicator wants to use different subjects or replicate with the same type of subject, he/she needs to know what sort of subjects were used in the first experiment
Rep-4	Is it clear how the *method/technology* works including all necessary assumptions?	The replicator need to understand the technologies/ methods being evaluated in order to construct test materials and devise test tasks
Rep-5	Is the *conduct* of the experiment clearly defined?	The replicator must know how the experiment was performed in order to replicate it
Rep-6	Are any *problems* or *difficulties* associated with the *experimental protocol* identified?	The replicator needs to know if there are any issues with the experimental protocol that need to be improved in a replication
Rep-7	Is the *effect size* reported for power analysis?	A replicator should be able to perform a power analysis to determine the required number of experimental units
Rep-8	Are the *training requirements* for subjects clear?	A replicator needs to provide appropriate training for subjects for all treatment conditions
Rep-9	Are *experimental materials* available for consultation?	A replicator may need to consult the experimental materials used by the original experimenters

Table 7 Reviewer's perspective

Number	Question	Rationale
Rev-1	Is the paper *original*?	The first priority for a reviewer is to establish that the paper is neither plagiarized nor a copy of a previously published paper
Rev-2	What is the *contribution* of the paper?	A reviewer needs to assess whether the contribution of the paper is sufficient to warrant publication
Rev-3	Are the *references appropriate*?	A reviewer needs to assess whether the author has an appropriate knowledge of the field
Rev-4	Is *background work* cited?	Related to the issue of references, reviewers need to assess whether all relevant background material is properly cited
Rev-5	Is the *design* correct?	Reviewers need to assess whether the design is appropriate to test the stated hypotheses
Rev-6	Is the *analysis* correct?	Reviewers need to confirm that the analysis is consistent with the specified design
Rev-7	Is it *readable* to the intended audience?	Given the audience of the journal or the expected background of conference participants, the reviewer must assess whether the language used in the paper appropriate

4.4 Performing the Reviews

For the first two reviews, in order to assist us to understand each perspective, we agreed to read a paper reporting an experiment from the International Symposium on Empirical Software Engineering (ISESE 04) at the same time as we read the guidelines. (Note. This initial reading activity took place before the group review meeting.) Four of the 26 papers in the ISESE 04 conference proceedings reported experiments (Abdelnabi et al. 2004; Abrahao et al. 2004; Schroeder et al. 2004; Verelst 2004) and each member of the group chose one of the papers to help with the review process. The choice of paper was not mandated and most people chose to read Verelst's paper, while no one opted for Abdelnabi et al.'s paper (see Table 8). This preliminary reading was intended simply to set the scene for reviewing the reporting guidelines. For this reason, we thought it was preferable to read an article that interested us rather than mandate the same article for everyone. We note that the relatively small number of experiments reported in a conference specializing in empirical methods confirms that experiments are currently not a major part of empirical software engineering.

While reading their chosen paper, each person in the group took one of the perspectives (self-chosen while ensuring both perspectives are covered). The allocation to paper and perspective is shown in Table 8. Everyone who took the practitioner viewpoint had worked for some time in industry (see Table 8), however, some participants with extensive industry experiment were studying for PhDs. In addition, one of the review team only took part in the later review meetings. He was a PhD student with 8 years industrial experience and two year's research experience. Participation in the workshops was not mandatory, and some NICTA staff attended only attended one or two meetings. These staff contributed to the discussion of the meetings they attended but are not included in Table 8 and did not coauthor this paper. The senior researcher attended all the meetings and kept a record of the discussions. Minutes were circulated after each meeting.

Although each person reviewed his/her chosen ISESE paper from a particular perspective, in the review meetings (first the research perspective and next the practitioner perspective), they were encouraged to contribute to the discussion of the other perspective. We had originally planned for each person to provide a written list of issues/defects from their allocated perspective. This was done for the first two reviews but not done for the last three reviews. In practice, we worked through each of the questions, discussed any issues arising and agreed whether the question raised any problems or identified defects with the guidelines. After the first two reviews, we did not attempt to allocate individuals to specific perspectives.

Table 8 Review perspective, paper selection and experience of reviewers

Perspective	Paper	Post graduate research experience (years)	Industrial experience (years)	NICTA position
Researcher	Verelst 2004	>3	>2	PhD student
Practitioner	Verelst 2004	>4	>6	PhD student
Practitioner	Abrahao et al. 2004	8	25	PhD student
Practitioner	Verelst 2004	>5	1.5	Researcher
Researcher	Abrahao et al. 2004	>2	0	PhD Student
Researcher	Schroeder et al. 2004	10	18	Senior researcher
Researcher	Verelst 2004	3	0	Research assistant
Practitioner	Schroeder et al. 2004	>3	5	Researcher
Researcher	Verelst 2004	>3	2 (+2 part-time)	PhD student

Table 9 Questions addressed and not addressed by the guidelines

Perspective	Questions addressed	Questions not addressed
Researcher	2, 3, 4, 8, 9, 10, 11, 12, 13, 16, 17	1, 6, 7, 14, 15
Practitioner	2, 3, 4, 5, 6, 8, 9, 13, 17, 19, 22	1, 7, 10, 11, 12, 14, 15, 16, 18, 20, 21
Meta-analyst	2, 4, 5, 6, 7, 10, 11, 12, 13, 14	1, 3, 8, 9
Replicator	2, 3, 5, 6, 7, 8, 9	1, 4
Reviewer	1, 2, 3, 4, 5, 6, 7	

The final review taking the author perspective proceeded differently. Again we used the ISESE papers to assist our understanding of the author perspective by re-reading our chosen article before taking part in the group review meeting. However, instead of using perspective-based questions at the meeting, we discussed each section of the guidelines sequentially.

5 Results

We found that the guidelines addressed many of the questions in each perspective (see Table 9). Overall they addressed 11 of the 17 Researcher perspective questions (65%), 12 of the 22 Practitioner perspective questions (55%), 10 of the 14 Meta-analyst questions (71%), 7 of the 9 Replicator perspective questions (78%) and all the 7 Reviewer perspective questions (100%). The percentages for the Researcher and Practitioner are not directly comparable to the percentages for the Meta-analyst, Replicator and Reviewer perspectives because we omitted general questions specified in the Researcher and Practitioner perspective from these perspectives. However, these results imply that specialist viewpoints are quite well-addressed by the guidelines but more general perspectives are less well addressed. In particular, the practitioner perspective is not very well addressed.

Although the guidelines addressed many questions, in some cases the guidelines were not specific enough about what needed to be reported and in other cases too precise. Overall, the perspective-based reviews using the Researcher, Practitioner, Meta-analyst, Replicator and Reviewer found 44 unique issues that we believed suggested the guidelines should be amended or clarified (see Tables 10, 11, 12, 13 and 14 respectively). The Researcher perspective identified 13 possible amendments, the Practitioner / Consultant perspective identified 21 possible amendments, the Meta-analyst perspective identified six possible amendments, the Replicator identified three possible amendments and the Review perspective identified one possible amendment. Of these amendments, most (i.e. 32) requested more detailed clarification of the information required in a guideline section. Four amendments requested the guidelines be less prescriptive, three requested more background information; two identified possible additional sections. The remaining three proposed amendments suggested (a) standardizing the contents of each section in the guideline document; (b) moving information from one section to another and (c) avoiding possible repetition.

We also identified eight items we classified as defects (see Table 15). The most significant defects are D2, D3, D4 and D8. D2 arises because the guidelines are inconsistent with reporting standards used by other experimental disciplines. It is a very significant step to disassociate our discipline from the standards used by all other scientific disciplines. We need to be sure that this step is necessary. At the very least we need to articulate the reasons for this divergence, so software engineering researchers and practitioners understand why it

Table 10 Proposed amendments arising from researcher perspective questions

Question Number	Id	Suggested amendments	Type
Res-1	1	The abstract should mention all relevant interventions, or conditions (i.e. independent variables) and dependent variables	More detail
	2	The title needs to be informative. Specify the interventions (i.e. independent variables) and dependent variables avoiding unnecessary redundancy	More detail
	3	The keywords should define the interventions, dependent variables and study type	More detail
Res-3	4	Suggest Introduction as alternative to Motivation for the section heading	Less prescription
	5	Justify deviations from the standard research paper structure	More background
	6	Remove unnecessary duplication and, where overlaps remain, clarify exactly what must be specified in each related section	More detail
Res-4	7	Authors should specify which are the main hypotheses and which are ancillary hypotheses and exploratory analyses	More detail
Res-5	8	The guidelines should require any causal model to be specified either in the Related work or the Motivation section	More detail
Res-6	9	Specify that the interventions must be fully described	More detail
Res-7	10	Identify the need for more general references as well as specific targeted references in the related work section	More detail
Res-9	11	The description of scope of Relation to Existing Work in the Conclusions and Future work section ought to say "relation of the results to earlier research" rather than "relation of the results to earlier experiments"	Less prescription
Res-12	12	If the raw data is not reported, the guidelines should require authors to specify under what conditions the raw data will be made available to other researchers	More detail
Res-14	13	The guidelines should advise authors to make it clear how they arrive at their interpretation given the specific results	More detail

is necessary. D3 is an important issue because it is an area that, if not addressed, may result in guidelines that make reporting experiments worse than it is currently. D4 is a general problem but a significant one. If we cannot write so that practitioners can understand and use our results, empirical software engineering is not very useful. D8 concerns the general principles of guidelines and standards—it should be clear what is mandated and what is optional.

Whether D1 is a defect or a design decision depends on whether the guidelines aim to address every section or only the most important sections of a research paper. If the guidelines are aiming for completeness, we suggest the need for appropriate relevant keywords be mentioned since well-chosen keywords will help readers find the paper. Defects D5, D6 and D7 could easily have been classified as possible amendments. D5 and D6 are both related to the reporting of the technology or technologies being evaluated. If such technologies are not properly described it is difficult for practitioners to use them. D7 was a specific example of an issue that arose for several of the suggested report section headings where the guidelines were too specific and should have used more general terms. Another example is the use of the term "subjects" rather than "experimental units". This raises another general issue that the guidelines may be too people/team centric. They do not address well the large number of technical tool "experiments" that get done in the Software Engineering

Table 11 proposed amendments arising from practitioner perspective questions

Question Number	Id	Suggested amendments	Type
P-1	14	The scope of the guidelines should make it clear that the title, keyword and abstract should contain commonly used industry terms	More detail
P-2	15	The scope should advice authors to identify cost, benefits, risks and transition issues in the abstract	More detail
	16	Abstracts will be of limited size so the guidelines should suggest priorities for what should appear (or indicate how authors ought to prioritize)	More detail
P-5	17	The scope information associated with the Related Work section (3.3) should request authors to comment on levels of industrial use of the techniques being evaluated (including the control)	More detail
	18	The scope information associated with the Inferences Section in Interpretation (3.7.3) should warn authors against making claims they cannot support	More detail
P-7	19	Guidelines should advise authors to clarify where results occur in the R&D lifecycle or the maturity of the technology	More detail
	20	If the technique is mature, the guidelines should advise authors to include an "implementation consequences" section	Missing section
	21	The description of any control treatment should be sufficient for readers to determine whether the control is realistic	More detail
P-9	22	The scope of the context section should be more specific about the information required	More detail
P-10	23	The guidelines should advise authors to report this issue in the Context Section (3.2.3) or the Impact section (3.8.2)	More detail
	24	The guidelines should ensure that information related to the practical use of the technology is reported	More detail
P-14	25	The scope of the Impact section (3.8.2) should require reporting any risks associated with the technology	More detail
P-15	26	Authors should be advised to mention in Context section (4.3) if the technology has been applied to real software projects	More detail
	27	For technologies not in use, authors should be advised to discuss scale-up issues in the Limitations of the Study section (3.7.2)	More detail
P-16	29	The Experimental design should discuss the task the subjects are asked to perform	More detail
P-18	30	If the authors report their data collection plan in the Data Collection Procedure section (3.4.6), they should only report deviations from that plan in the Data Collection Performed section (3.5.3)	More detail
	31	The Data collection performed section (3.5.3) would be better called "Deviations from experimental plan" and specify all deviations from the experimental protocol	Less prescription
	32	Activities involving marking the outcomes of the experimental tasks and training provided for markers should be in the Data Collection Procedure section (3.4.6) not the Goals, Hypotheses and Variables section (3.4.1)	Change allocated section
	33	The guidelines should have a section for specifying the treatments that are being compared	Missing section
P-19	34	Authors should be advised to mention personal vested interests in the Related work (3.3) or Limitations of the Study (3.7.2) sections	More detail

Table 12 Proposed amendments arising from meta-analyst perspective questions

Question Number	Id	Suggested amendments	Type
M-1	35	The information accompanying the guidelines should advice authors to report the number of experimental units	More detail
M-2	36	The guidelines should explain why meta-analysts need access to raw data (i.e. to standardize analyses) and make it clear that reporting raw data (or making clear the conditions under which raw data will be made available) is extremely important. Furthermore if raw data is not reported authors have an obligation to present effect size information	More background
	37	The guidelines should clarify what information needs to be reported in which part of section 3.6	More detail
M-5	38	Supporting information should be in a standard format identifying the meaning of the element, why an element is required, and what should be reported (scope and format)	Standardising guidelines
M-10	39	The guidelines should spell out what types of empirical study they apply to	More background
M-15	40	Use the term experimental unit rather than subjects	Less prescription

discipline (of which Schroeder et al. 2004 is an example). Are these considered different types of studies? If so, it would be useful to clarify this in the scope of the guidelines; if not, the guidelines should be amended to make them more relevant to technology-intensive experiments.

The final review based on the Author's perspective re-iterated many issues noted previously. In particular, we were concerned about suggestions to impose reporting structures that were incompatible with those used in other disciplines, such as the template structure for reporting research objectives and the section headings (see Harris 2002 and Moher et al. 2001 for more conventional section headings). The problem of possible duplication was also reiterated. The main issues that were not raised previously were that:

- The relationship between the "Experimental Design" and the "Execution" section needed to be clarified. If the first section was really the "Experimental Plan" and was fully reported, then the "Execution" section should be restricted to reporting deviations from the plan.
- The ordering of sections was not always appropriate, for example sometimes it is necessary to introduce the measurement concepts before specifying the hypotheses.

Table 13 Proposed amendments arising from replicator perspective questions

Question Number	Id	Suggested amendments	Type
Rep-1	41	Guidelines should advise authors to ensure e-mail contact information is available for all authors	More detail
Rep-8	42	There should be more advice on what is reported in Lessons learnt as opposed to other places for reporting deviations from plan	More detail
Rep-9	43	The guidelines should advise authors to make a statement about how experimental materials can be obtained and for how long they will be maintained	More detail

Table 14 Proposed amendments arising from reviewer perspective questions

Question Number	Id	Suggested amendments	Type
Rev-3	44	Authors should be advised to include all related work whether supportive or contradictory in Section 3.3	More detail

6 Discussion and Conclusions

The guidelines addressed many of the questions raised by each perspective, but we found many instances where the guidelines might benefit from amendment and eight instances where we thought the guidelines were defective.

Issues arising from the Author's perspective identified problems with potential duplication of information. Guidelines need to be very clear about what information goes into which section. This is a problem for the "Experimental Design" and "Execution" sections as well as the numerous validity sections.

Our results suggest that the main problems with the current version of the guidelines are:

1. Relationships among the individual elements are not clear in the case of reporting validity issues and the reporting of planned tasks versus actual conduct. Thus, it is difficult to be sure what information to put in which section. There is also a risk that the guidelines will result in unnecessary duplication that would make experimental reports less readable.
2. In places, the guidelines require us to adopt reporting standards that are inconsistent with those of other disciplines. For example the suggested headings are inconsistent with

Table 15 Defects identified by perspective-based reviews

Question	ID	Defect	Guidelines reference
Res-1	D1	The guidelines omit any reference to keywords	None
Res-3	D2	The guidelines do not conform to the classic reporting structure used by most other scientific domains (e.g. the IMRAD (Introduction, Material & Methods, Results, and Discussion) format see Harris 2002 and Moher et al. 2001). Furthermore the deviation from the standard structure is not justified	All
Res-3	D3	The guidelines advice discussing validity and generalizability in five separate places and thus introduce the possibility of considerable duplication and redundancy	Context 3.2.3
P-10	D4	The guidelines do not address the needs of the practitioner. This issues is recorded against P-10 but arose because it applied to many of the previous practitioner questions	None
P-16	D5	The guidelines do not require that the tasks the technology addresses are described	Context 3.2.3
P-18	D6	The guidelines do not require that the treatments (or levels) be defined in operational terms	Related work 3.3
M-2	D7	The heading "Data set reduction" is too specific since the author needs to report procedures for data transformation, handling missing values etc. Change to "Data set preparation"	Section 3.6.2
M-2	D8	In several places the information accompanying the guidelines references the advice of other authors. However, it does not specify whether the advice should be followed or not (nor whether advice from different authors is contradictory)	None

the IMRAD standard (see Harris 2002 and Moher et al. 2001). We need to be absolutely certain that this is a good idea.

Our results suggest that the guidelines need to be revised. Any revised guidelines will need to be subjected to further theoretical and empirical validation if they are to be generally accepted. We also need to review research results in other disciplines that might provide additional justification for the guidelines structure and contents. For example, as noted by Jedlitschka and Pfahl (2005), Hartley (2004) provides a summary of the numerous studies that have assessed the value of structured abstracts.

A limitation of our evaluation methodology (review using perspective-based checklists) is that we started our evaluation with perspectives that included general questions and ended it with perspectives that included mainly perspective-specific questions. Furthermore we did not check whether some questions were in essence the same but were asked in different ways. We believe that it is preferable to have a separate list of general questions and another list of specific questions for each perspective. Table 16 identifies a set of 17 general questions cross-referenced to the perspectives from which they were obtained; the questions that are the same or similar in other perspectives; and the perspectives to which they apply. Analyzing Table 16 with respect to questions addressed by the guidelines identified in Table 9 shows that the guidelines provide very good coverage of general questions, with 15 of the 17 general questions (88%) addressed by the guidelines.

Our choice of evaluation method seemed to work well for an initial theoretical validation. Our approach of multiple reviews fitted well with the training element of our evaluation exercise but is not an essential element of a review based evaluation. It would be much quicker to perform a single review with individuals each taking a different perspective. We suggest that a similar review-based evaluation should be performed on the revised guidelines. This type of evaluation would be appropriate for any research group

Table 16 General questions

Original Question Id	Question	Replaces Question	Relevant to perspective
Res-1	Is the paper easy to find?	P-1	All except Rev
Res-2	Is it a relevant paper?	P-2	All
Res-3	Is the overall structure of the paper appropriate?		All
Res-4	Is the research problem hypothesis easy to identify?	Rep-2	All
Res-8	Is required background knowledge referenced?		Rep & P
Res-9	Is the research related to other relevant research?	P-6, Rev-4	All
Res-10	Is the experimental design appropriate?		All
Res-11	Is the statistical analysis correct?		All
Res-12	Is the raw data available?		All
Res-13	Is it easy to identify the findings/results of the experiment?	P-4	All
Res-14	Do the conclusions arise from the results?	P-5	All
Res-16	Are limitations of the experiment made clear?		All
Res-17	Is there any discussion of required further research?	P-22	All
P-18	Is the new approach/technique well defined?	M-3 & Rep-4	All
P-19	Does the paper make it clear who is funding the experiment and whether they have any vested interests?		All
M-6	Is the experimental procedure well-defined?	Rep-5	All
M-14	Are any side-effects or risk associated with the treatments defined	P-14	All

that includes staff with research and industrial experience. It would be useful for any research group intending to adopt the guidelines to undertake such an evaluation. With respect to the other evaluation options listed in Table 2, we believe that most of the evaluation options are useful and viable for specific stakeholders:

- Evaluation of each guideline element (i.e. T1 in Table 2) and determining the mapping between the new guidelines and existing experimental guidelines (i.e. T2) should be performed by the guideline developers.
- Evaluation of the guideline development process (i.e. T3) is the responsibility of the research community, so could be undertaken by research networks such as the International Software Engineering Research Network (ISERN, http://isern.iese.de/network/ISERN/pub/).
- An empirical evaluation method based on comparing the completeness of papers prepared using the guidelines with those that do not (i.e. E1), cannot be undertaken until guidelines are more widely adopted.
- An empirical evaluation method based on rewriting existing papers in order to conform to the guidelines and comparing them with the original versions (i.e. E2), requires a substantial research effort and would be best addressed by a research network. However, experimental validation involving re-writing existing experimental reports poses a number of practical problems. A significant problem is that it is difficult to assess how well written any experimental report is, so it may be difficult to assess the before and after versions of a report objectively. In addition, re-writing an existing report will depend on the expertise of the researchers doing the re-writing and the quality of the original report, not just the quality of the guidelines.

An important issue raised by the evaluation exercise is that of the Practitioner/Consultant viewpoint. The guidelines did not fit this perspective well. Attempts to address this perspective would make papers much longer and probably more complex. Would it be better to have different standards for practitioner-oriented papers? On the one hand, it can be argued that experiments in software engineering are not relevant to practitioners because they usually involve students, and/or simplified tasks and materials, and/or unrealistic settings. This would suggest practitioners only want to read case studies or industrial surveys. On the other hand, even if controlled experiments are not representative of industry practice, they provide proof of concept information without which industry is unlikely to undertake any realistic case studies. One course of action may be to re-write research results for practitioner-oriented magazines (as long as copyright issues are addressed). However, it may also be beneficial to identify the issues that are most important to practitioners and ensure they are covered by the current guidelines.

This paper has evaluated guidelines for controlled experiments. However, we believe that software engineering needs reporting guidelines for other types of empirical studies, in particular, case studies performed in industrial settings and industry surveys, not least because these types of study are of most relevance to practitioners. We believe that many of the perspective-based questions related to Researchers, Practitioners, and Reviewers are quite general (with the exception of questions that relate specifically to the methodology used for formal experiments) and can be used to help evaluate reporting guidelines developed for other forms of empirical study. Even the Meta-analyst perspective and the Replicator perspective are relevant to other forms of study although the questions would need to be revised. In particular, any attempt to construct and evaluate guidelines for industrial case studies and surveys should ensure that the Practitioner perspective is fully considered.

Acknowledgement NICTA is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

Appendix

Table 17 Meeting schedule

Meeting Number	Date	Purpose	Outcome
1	5th October	Agreement to undertake evaluation activity and identification of evaluation method	Agreed to perform a perspective-based review. Identified the perspectives we would use
2	12 October 2005	Specify a perspective-based checklist for Researchers and Practitioner/ consultants	Checklist questions identified
3	19 October 2005	Review the guidelines from the perspective of researchers	A list of problems and defects
4	26 October 2005	Review the guidelines from the perspective of practitioner/ consultants	A list of problems and defects
5	2 November 2005	Review progress and Decide whether to continue with remaining perspectives	Developed checklist questions for Meta-analyst, Replicator and Review. Decided to treat the author perspective differently
6	23 November 2005	Review from the Meta-analyst perspective	A list of problems and defects
7	7 December 2005	Review from the perspective of a replicator and a reviewer. Note We were joined for this meeting by Professor Martin Shepperd and Carolyn Mair from Brunel University	Two lists of problems and defects
8	14 December 2005	Review from the perspective of an author	A list of problems and defects

References

Abdelnabi Z, Cantone G, Ciolkowski M, Rombach D (2004) Comparing code reading techniques applied to object-oriented software frameworks with regard to effectiveness and defect detection rate *Proceedings ISESE 04*.

Abrahao S, Poels G, Pastor O (2004) Assessing the reproducibility and accuracy of functional size measurement methods through experimentation, *Proceedings ISESE 04*.

Dybå T, Kampenes VB, Sjøberg DIK (2006) A systematic review of statistical power in software engineering experiments. Inf Softw Technol 48(8):745–755

Harris P (2002) Designing and reporting experiments in psychology, 2nd edn. Open University Press.

Hartley J (2004) Current findings from research on structured abstracts. J Med Libr Assoc 92(3):368–371

Jedlitschka A, Pfahl D (2005) Reporting guidelines for controlled experiments in software engineering. IESE-Report IESE-035.5/E

Kitchenham B (2004) Procedures for performing systematic reviews. Joint Technical Report, Keele University TR/SE-0401 and NICTA 0400011T.1, July

Kitchenham B, Pfleeger SL, Pickard L, Jones P, Hoaglin D, El Emam K, Rosenberg J (2002) Preliminary guidelines for empirical research in software engineering. IEEE Trans Softw Eng 28(8):721–734
Kitchenham B, Al-Khilidar H, Ali Babar M, Berry M, Cox K, Keung J, Kurniawati F, Staples M, Zang H, Zhu L (2006) Evaluating guidelines for empirical software engineering studies, ISESE06, Brazil
Moher D, Schultz KF, Altman D (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Lancet 357:1191–1194, April 14
Pickard LM, Kitchenham BA, Jones P (1998) Combining empirical results in software engineering. Inform Softw Technol 40(14):811–821
Schroeder PJ, Bolaki P, Gopu V (2004) Comparing the fault detection effectiveness of N-way and random test suites, *Proceedings ISESE 04*
Sjøberg DIK, Hannay JE, Hansen O, Kampenes VB, Karahasanovic A, Liborg N-K, Rekdal AC (2005) A survey of controlled experiments in software engineering. IEEE Trans Softw Eng (9):733–753, September 31
Shull Forest, Rus Ionna, Basili Victor (2000) How perspective-Reading can Improve Requirements Inspection. IEEE Computer 73–78, July
Verelst Jan (2004) The influence of the level of abstraction on the evolvability of conceptual models of information systems. *Proceedings ISESE 04*
Wohlin C, Runeson P, Höst M, Regnell B, Wesslén A (2000) Experimentation in software engineering. An introduction. Kluwer Academic Publishers
Wohlin C, Petersson H, Aurum A (2003) Combining data from reading experiments in software inspections. In: Juristo N, Moreno A (eds) Lecture notes on empirical software engineering. World Scientific Publishing

Barbara Kitchenham is Professor of Quantitative Software Engineering at Keele University in the UK. From 2004–2007, she was a Senior Principal Researcher at National ICT Australia. She has worked in software engineering for nearly 30 years both in industry and academia. Her main research interest is software measurement and its application to project management, quality control, risk management and evaluation of software technologies. Her most recent research has focused on the application of evidence-based practice to software engineering. She is a Chartered Mathematician and Fellow of the Institute of Mathematics and Its Applications, a Fellow of the Royal Statistical Society and a member of the IEEE Computer Society.

Hiyam Al-Kilidar obtained her bachelor degree in civil engineering from Kuwait University in 1986 and Masters of Engineering Science in Project Management from the University of New South Wales, Australia in 1997. She is currently a PhD candidate at the University of New South Wales. Between 1986 and 1991, she

was a design engineer at the Kuwait Institute for scientific research. Since 1996 Hiyam has held a number of teaching and research positions at the University of New South Wales, NICTA, the University of Technology, Sydney and Muscat College of Management Science and Technology, Oman. Her research interests are in the areas of project management, quality management, pair programming and design and software quality.

Muhammad Ali Babar is a Senior Researcher with Lero, the Irish Software Engineering Research Centre. Previously, he worked as a researcher with National ICT Australia (NICTA). Prior to joining NICTA, he worked as a software engineer and an IT consultant. He has authored/co-authored more than 50 publications in peer-reviewed journals, conferences, and workshops. He has presented tutorials in the area of software architecture knowledge management at various international conferences including ICSE 2007, SATURN 2007 and WICSA 2007. He obtained an MSc in computing sciences from the University of Technology, Sydney and a PhD in Computer Science and Engineering from the University of New South Wale. His current research interests include software product lines, software architecture design and evaluation, architecture knowledge management, tooling supporting, and empirical methods of technology evaluation.

Michael Berry is the principal of EBSE (Evidence Based Software Engineering) Australia. He obtained his PhD from the University of New South Wales for his research into the Assessment of Software Measurement. Mike spent the first twenty-five years of his working life in the software industry. He was a programmer and systems analyst before moving into managing software development infrastructure through standards, process improvement and measurement activities. He has held an academic position at the University of New South Wales, a research position within CSIRO and is currently a Visiting Researcher at National ICT Australia.

Karl Cox is a Senior Researcher at NICTA. His mission is to develop and sell products and solutions that meet industry's biggest IT challenges through innovative research by continually engaging with industry to

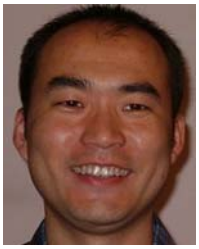
address those challenges. His current research portfolio is in: Strategic IT Requirements Management, Business Process Management, Enterprise Architecture, Requirements for On-Demand Computing. He has published widely in the research community primarily on requirements engineering. He holds a PhD in Computer Science from Bournemouth University, UK.

Jacky Keung is a research scientist at National Information and Communications Technology Australia (NICTA). NICTA is Australia's centre of excellence for Information and Communications Technology R&D. His research interests are in software measurement and its application to project management, cost estimation, quality control and risk management. His most recent research focuses on the application of case-based reasoning approaches to software engineering. He completed his BS (Hons) in Computer Science from the University of Sydney, and received his PhD from the University of New South Wales for his research into the statistical methods of software cost estimation. He also has held an academic position at the University of New South Wales. He is a member of the Australian Computer Society, and a member of the IEEE Computer Society.

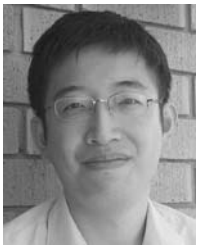
Felicia Kurniawati is currently a Business Analyst with Terra Firma and prior to that, she was a Research Engineer with the Empirical Software Engineering group at National ICT Australia and a Research Assistant with the School of Computer Science and Engineering at the University of New South Wales (UNSW). Her main area of interest is process modelling and improvement. She has worked with software and business processes in both the research and the corporate sectors. Felicia received the degrees of Bachelor of Engineering (first class), majoring in Software Engineering (first class), and Master of Commerce in Marketing from UNSW.

Mark Staples is a Senior Researcher with NICTA. His research interests include software process, software configuration management, and software product line development. Prior to joining NICTA in 2004, he spent

four years working in the software industry. He holds undergraduate degrees in Computer Science and Cognitive Science from the University of Queensland, and a PhD in Computer Science from the University of Cambridge.



He Zhang received his Master of Project Management from University of Sydney, MSc in computer science from Chinese Academy of Sciences and Astronautics, and BEng from Nanjing University of Aeronautics and Astronautics in China. He is currently a PhD candidate in School of Computer Science and Engineering at University of New South Wales, and a research fellow of Empirical Software Engineering Program at National ICT Australia. His research focuses on software process modeling and simulation, software project management, and empirical software engineering. He is a member of ACM SIGSOFT.



Liming Zhu is a researcher at National ICT Australia (NICTA) and a visiting fellow to School of Computer Science and Engineering at University of New South Wales. He worked in several technology lead positions in software industry before obtaining a PhD degree in software engineering at University of New South Wales (UNSW). He organized a number of model driven development workshops, authored book chapters and developed innovative tools in the area. He has worked with Australia government agencies, defense, standardization bodies. His research interests include architecture evaluation, model driven development, service-oriented architecture, business process modeling, and software development processes.