Chapter 8
Reporting Experiments in Software Engineering

Andreas Jedlitschka, Marcus Ciolkowski, and Dietmar Pfahl

Abstract
Background: One major problem for integrating study results into a common body of knowledge is the heterogeneity of reporting styles: (1) It is difficult to locate relevant information and (2) important information is often missing.
Objective: A guideline for reporting results from controlled experiments is expected to support a systematic, standardized presentation of empirical research, thus improving reporting in order to support readers in (1) finding the information they are looking for, (2) understanding how an experiment is conducted, and (3) assessing the validity of its results.
Method: The guideline for reporting is based on (1) a survey of the most prominent published proposals for reporting guidelines in software engineering and (2) an iterative development incorporating feedback from members of the research community.
Result: This chapter presents the unification of a set of guidelines for reporting experiments in software engineering.
Limitation: The guideline has not been evaluated broadly yet.
Conclusion: The resulting guideline provides detailed guidance on the expected content of the sections and subsections for reporting a specific type of empirical study, i.e., experiments (controlled experiments and quasi-experiments).

1. Introduction

In today's software development organizations, methods and tools are employed that frequently lack sufficient evidence regarding their suitability, limits, qualities, costs, and associated risks. In Communications of the ACM, Robert L. Glass (2004), taking the standpoint of practitioners, asks for help from research: "Here's a message from software practitioners to software researchers: We (practitioners) need your help. We need some better advice on how and when to use methodologies." Therefore, he asks for:

- A taxonomy of available methodologies, based upon their strengths and weaknesses

F. Shull et al. (eds.), *Guide to Advanced Empirical Software Engineering*.

- A taxonomy of the spectrum of problem domains, in terms of what practitioners need
- A mapping of the first taxonomy to the second (or the second to the first)

Empirical software engineering (ESE) addresses some of these issues partly by providing a framework for goal-oriented research. The aim of this research is to build an empirically validated body of knowledge and, based on that, comprehensive problem-oriented decision support in the software engineering (SE) domain.

However, one major problem for integrating study results into a body of knowledge is the heterogeneity of study reporting (Jedlitschka and Ciolkowski, 2004). It is often difficult to find relevant information because the same type of information is located in different sections of study reports and important information is also often missing (Wohlin et al., 2003; Sjøberg et al., 2005; Dybå et al., 2006; Kampenes et al., 2007). For example, in study reports, context information is frequently reported differently and without taking into account further generalizability. Furthermore, specific information of interest for practitioners is often missing, like a discussion of the overall impact of the technology on project or business goals.

One way to avoid this heterogeneity of reporting is to introduce and establish reporting guidelines. Specifically, reporting guidelines support a systematic, standardized description of empirical research, thus improving reporting in order to support readers in (1) finding the information they are looking for, (2) understanding how an experiment is conducted, and (3) assessing the validity of its results. This claim is supported by the CONSORT statement (Altman et al., 2001), a research tool in the area of medicine that takes an evidence-based approach to improve the quality of reports of randomized trials to facilitate systematic reuse (e.g., replication, systematic review, and meta analysis).

As identified by Kitchenham et al. (2002, 2004), reporting guidelines are necessary for all relevant kinds of empirical work, but they must address the needs of different stakeholders (i.e., researchers and practitioners). The specific need for standardized reporting of controlled experiments has been mentioned by different authors for a long time, e.g., Lott and Rombach (1996), Pickard et al. (1998), Shull et al. (2003), Vegas et al. (2003), Wohlin et al. (2003), and Sjøberg et al. (2005). At the same time, several more or less comprehensive and demanding reporting guidelines have been proposed, e.g., by Singer (1999), Wohlin et al. (2000), Juristo and Moreno (2001), and Kitchenham et al. (2002). Even though each of these proposals has its merits, none has yet been accepted as a de-facto standard. Moreover, most of the existing guidelines are not explicitly tailored to the specific needs of certain types of empirical studies, e.g., controlled experiments a comprehensive classification of empirical studies is given by Zelkowitz et al. (2003).

The goal of this chapter is to survey the published proposals for reporting guidelines and to derive a unified and – where necessary – enhanced guideline for reporting controlled experiments and quasi-experiments. Nevertheless, many of the elements discussed throughout this chapter will also make sense for reporting other types of empirical work.

2. Background

Empirical software engineering research is not the first research domain to encounter problems with insufficient reporting. Other disciplines, such as medicine and psychology, have experienced similar problems and have achieved various improvements by standardizing and instantiating reporting guidelines, e.g., for randomized controlled trials in biomedical research (Altman et al., 2001; Moher et al., 2001), psychology (Harris, 2002), clinical practice guidelines (Shiffman et al., 2003), and empirical results from psychological research (American Psychological Association, 2001).

In the field of SE research, in 1999, Singer (1999) described how to use the "American Psychological Association (APA) Styleguide" (2001) for publishing experimental results in SE. In 2002, Kitchenham et al. (2002) provided initial guidelines on how to perform, report, and collate results of empirical studies in SE based on medical guidelines as well as on the personal experience of the authors. Shaw (2003) provided a tutorial on how to write scientific papers, including the presentation of empirical research as a special case. Additionally, standard text books on empirical SE, such as Wohlin et al. (2000) and Juristo and Moreno (2001), address the issue of reporting guidelines. Wohlin et al. (2000) suggest an outline for reporting the results of empirical work. Juristo and Moreno (2001) provide a list of the "most important points to be documented for each phase" in the form of "questions to be answered by the experimental documentation."

Jedlitschka et al. presented a first version of a guideline for reporting controlled experiments (2005a) during a workshop on empirical software engineering (Jedlitschka, 2005). Feedback from the workshop participants, as well as from peer reviews, was incorporated into a second version of the guideline (2005b). In parallel, the guideline was evaluated by means of a perspective-based inspection approach (Kitchenham et al., 2006). This evaluation highlighted 42 issues where the guideline would benefit from amendment or clarification and eight defects. The feedback from the perspective-based inspection and discussions with its authors led to a second iteration of the guideline, where the amendments were incorporated if we found them appropriate and defects were removed (Jedlitschka and Ciolkowski, 2006). Additional feedback from individual researchers was also incorporated (Jedlitschka et al., 2007).

Table 1 characterizes the existing proposals for guidelines on reporting empirical work in SE. The first row of the table lists the proposals, arranged with regard to their publication date. The second row of the table describes the focus of the guidelines. The entry "Empirical Research" indicates that the guidelines are not tailored to a specific type of empirical research. Otherwise, the specific type is explicitly mentioned, e.g., "Controlled Experiment" or "Systematic Review." The third row describes the phases of an experiment covered by the guideline. The entry "All" indicates that the guideline covers all phases of a study. The remaining rows list the structuring elements in the proposed guidelines and map them to the structure of our proposal (last column). Elements of existing proposals occurring twice in a column indicate that these elements can be mapped to two different elements of our new proposal.

Table 1 Overview on structuring proposals for reporting controlled experiments

	Singer (1999)	Wohlin et al. (2000)	Kitchenham et al. (2002)	Juristo and Moreno (2001)	Kitchenham (2004)	Jedlitschka et al. (2007)
Type of study	Empirical research	Empirical research	Empirical research	Controlled experiment	Systematic review	Controlled experiment
Phases of study	Reporting	All	All	All	All	Reporting
Structure	*	*	*	*	Title	Title
	*	*	*	*	Authorship	Authorship
	*	*	*	*	Keywords	Keywords
	Abstract	*	*	*	Executive summary or structured abstract	Structured abstract
	Introduction	Introduction Problem statement Experiment planning	*	Goal definition	Background	Introduction
	Introduction	Problem statement	Experimental context	Goal definition	Background	Background
	Method	Experiment planning	Experimental design	Design	Review questions Review methods	Experiment planning
	Procedure	Experiment operation	Conducting the experiment and data collection	Experiment execution	Included and excluded studies	Deviations from the plan
	Results	Data analysis	Analysis	Experimental analysis	Results	Analysis
	Discussion	Interpretation of results	Interpretation of results	Experimental analysis	Discussion	Discussion
	Discussion	Discussion and conclusion	*	Experimental analysis	Conclusion	Conclusions and future work
	–	–	–	–	Acknowledgments Conflict of interest	Acknowledgements
	References	References	*	*	References	References
	Appendices	Appendix	*	*	Appendices	Appendices

An asterisk (*) indicates that the authors do not explicitly mention or describe details for this element, but it is assumed that the elements are implicitly required.

We investigated the structures of published reports of controlled experiments in empirical software engineering and have concluded that, in general, authors do not use a common set of guidelines in determining what information to include in their report. In other disciplines, such as medicine and psychology, editors have agreed on a common reporting style, not only regarding the layout of the report, but also its content. Given that the first publication of a reporting guideline for empirical SE research by Singer (1999) was over 7 years ago and little has progressed since that time, we conclude that significant effort needs to be invested to make sure that guidelines are widely accepted and used. This is what other communities have already learned (Altman et al., 2001; Harris, 2002).

Because of this, this chapter provides a description of the most common elements in the various reporting guidelines, giving guidance to readers where we have diverged from others suggestions. This guideline should be seen as a means for supporting both authors of a report in providing relevant information in the appropriate place and readers of a report in knowing where to look for a certain type of information.

3. Guideline for Reporting Controlled Experiments

In this section, we discuss what information should be presented in reports of experiments. It some cases, it may be necessary to adapt the length of a report depending on the requirements of the publisher. Therefore, the structure as presented in this section provides several options. For example, for a conference paper (which is usually much shorter than a journal paper) it may be appropriate to combine the description of the experiment planning and the deviations from the plan as well as the description of the analysis procedure and the analysis, whereas for a journal paper, it is generally appropriate to separate the content of these sections.

In all reports, however, generally speaking, enough information has to be provided to enable readers to judge the reliability of the experiment. The need for detailed provision of information is not specific for SE. It is, for example, also pointed out by Harris (2002). We are well aware that due to limitations of pages (e.g., for conferences), this is not possible in all cases, but the author should at least keep this intention in mind while compiling the report.

As indicated in Table 1, our reporting guideline comprises the following elements: Title, Authorship, Structured Abstract, Keywords, Introduction, Background, Experiment Planning, Execution, Analysis, Discussion, Conclusion and Future Work, Acknowledgements, References, and Appendices.

Our proposal reflects the requirements of existing standards, such as APA, but provides more structuring elements and asks for specific details that are not relevant for many experiments in psychology, like a technology's impact on the overall project budget or time and on the product's quality. Furthermore, our guideline incorporates wording as it is common for experiments in empirical SE to also

Table 2 Quick reference

Section	Content	Scope	Priority
3.1 Title		<title> + "– A controlled experiment"; Is it informative and does it include the major treatments and the dependent variables?	Required
3.2 Authorship		Does it include contact information, i.e., a valid email?	Required
3.3 Structured abstract	Background	Why is this research important?	Required
	Objective	What is the question addressed with this research?	Required
	Methods	What is the statistical context and methods applied?	Required
	Results	What are the main findings? Practical implications?	Required
	Limitations	What are the weaknesses of this research?	
	Conclusions	What is the conclusion?	Required
3.4 Keywords		Areas of research the treatments, dependent variables, and study type	Might be required by the publisher
3.5 Introduction	Problem statement	What is the problem? Where does it occur? Who has observed it? Why is it important to be solved?	Required
	Research objective	What is the research question to be answered by this study? E.g., by using the GQM goal template: Analyze <Object(s) of study> for the purpose of <purpose> with respect to their <Quality Focus> the point of view of the <Perspective> in the context of <context>	Required
	Context	What information is necessary to understand whether the research relates to a specific situation (environment)?	Required
3.6 Background	Technology under investigation	What is necessary for a reader to know about the technology to reproduce its application?	Required if not published elsewhere
	Alternative technologies	How does this research relate to alternative technologies? What is the control treatment?	Required
	Related studies	How this research relates to existing research (studies)? What were the results from these studies?	If available
	Relevance to practice	How does it relate to state of the practice?	If available
3.7 Experiment planning	Goals	Formalization of goals, refine the important constructs (e.g., the quality focus) of the experiment's goal	Required

(continued)

Table 2 (continued)

Section	Content	Scope	Priority
	Experimental units	From which population will the sample be drawn? How will the groups be formed (assignment to treatments)? Any kind of randomization and blinding has to be described	Required
	Experimental material	Which objects are selected and why?	Required
	Tasks	Which tasks have to be performed by the subjects?	Required
	Hypotheses, parameters, and variables	What are the constructs and their operationalization? They have to be traceable derived from the research question respectively the goal of the experiment	Required (for an explorative studies there might be no hypothesis defined)
	Design	What type of experimental design has been chosen?	Required
	Procedure	How will the experiment (i.e. data collection) be performed? What instruments, materials, tools will be used and how?	Could be integrated with execution
	Analysis procedure	How will the data be analyzed?	Could be integrated with analysis
3.8 Execution	Preparation	What has been done to prepare the execution of the experiment (i.e., schedule, training)	
	Deviations	Describe any deviations from the plan, e.g., how was the data collection actually performed?	
3.9 Analysis	Descriptive statistics	What are the results from descriptive statistics?	Required
	Data set preparation	What was done to prepare the data set, why, and how?	
	Hypothesis testing	How was the data evaluated and was the analysis model validated?	
3.10 Discussion	Evaluation of results and implications	Explain the results and the relation of the results to earlier research, especially those mentioned in the *Background* section	
	Threats to validity	How is validity of the experimental results assured? How was the data actually validated?	Required

(continued)

Table 2 (continued)

Section	Content	Scope	Priority
		Threats that might have an impact on the validity of the results as such (threats to internal validity, e.g., confounding variables, bias), and, furthermore, on the extent to which the hypothesis captures the objectives and the generalizability of the findings (threats to external validity, e.g., participants, materials) have to be discussed	
	Inferences	Inferences drawn from the data to more general conditions	Required
	Lessons learned	Which experience was collected during the course of the experiment	Nice to have
3.11 Conclusions and future work	Summary	The purpose of this section is to provide a concise summary of the research and its results as presented in the former sections	Required
	Impact	Description of impacts with regard to cost, schedule, and quality, circumstances under which the approach presumably will not yield the expected benefit	
	Future work	What other experiments could be run to further investigate the results yielded or evolve the Body of Knowledge	
3.12 Acknowledgements		Sponsors, participants, and contributors who do not fulfil the requirements for authorship should be mentioned	If appropriate
3.13 References		All cited literature has to be presented in the format requested by the publisher	Absolutely required
3.14 Appendices		Experimental materials, raw data, and detailed analyses, which might be helpful for others to build upon the reported work should be provided	Might be made available trough technical reports or web site

support the reading of already published reports. The structuring elements are discussed in detail in the following subsections. Table 2 shows each element, along with the section it is detailed in, and its particular sub-elements.

3.1. Title

The title of the report has to be informative, because the title (together with the abstract) "alerts potential readers to the existence of an article of interest" (Harris, 2002). To attract readers from industry, it is important to use commonly used industry terms. Harris (2002) suggests avoiding phrases like "A Study of" or "An Experimental Investigation of." This might be true for psychology, but for ESE, where we do not have explicit journals for experiments, we propose adding "– a controlled experiment" (– a replicated controlled experiment, – a quasi-experiment) if there are no limitations with regard to the title length. This helps the reader to easily identify controlled experiments. Furthermore, if possible, it additionally aides the reader if the dependent variables and treatments can be specified in the title.

In fact, where the title length is limited, we believe it is more important to include treatments and the dependent variables than "a controlled experiment." As an example of a succinct meaningful title, consider the following: The title of a publication describing a controlled experiment to investigate technique X compared to technique Y (the treatments) regarding the maintainability of a product (dependent variable) could be "Comparing the Impact of Technique X and Technique Y on Product's Maintainability – A Controlled Experiment." From the perspective of a reader, both from research as well as from industry, this title would allow for easily identifying the main aspects of the publication.

3.2. Authorship

All individuals making a significant contribution should be in the author list or at least acknowledged (c.f. Sect. 3.12).

Most report styles require contact details. If not, provide at least the e-mail address of the responsible author. As authors might change their job, it is sometimes more appropriate to provide the contact information of the more stable author – for example a professor as opposed to a graduate student (Kitchenham, 2004), or, "to be on the safe side," provide contact information for all authors.

3.3. Structured Abstract

The need for a self-contained abstract is beyond any question. It is an important source of information for any reader, as it briefly summarizes the main points of the study and, moreover, is often the only part of a publication that is freely accessible (Kitchenham, 2004). Abstracts should summarize the broad research questions.

Additionally, for a single experiment, regardless of the format of the abstract, authors should ensure that all relevant interventions or conditions (i.e., independent variables) and dependent variables are mentioned. When more than one experiment is reported in a paper, this may be infeasible, and instead authors will need to describe their experiments in more general terms.

The exact format of the abstract needs more discussion. For example, Shaw (2003) found that there is a common structure for the clearest abstracts consisting of the following elements: (a) the current state of the art, identifying a particular problem, (b) the contribution to improving the situation, (c) the specific result and the main idea behind it, and (d) how the result is demonstrated or defended. For reporting experiments in psychology, Harris (2002) suggests that an abstract should describe the following aspects: (1) the problem under investigation, (2) the participants, (3) the empirical method, (4) the findings, and (5) the conclusions.

A large number of journals in medicine and psychology have imposed a special form of the abstract, the structured abstract (Hayward et al., 1993; Bayley and Eldredge, 2003), on authors to improve the clarity of abstracts. The most common elements of structured abstracts are *Background* or *Context*, *Objective* or *Aim*, *Method*, *Results*, and *Conclusion*.

Inspired by the lessons learned from medicine, we propose using a structured abstract consisting of the elements listed below:

Background: Give a brief explanation of the motivation for conducting the study. Example: "Software developers have a plethora of development technologies from which to choose, but often little guidance for making the decision" (Shull et al., 2003).

Objective: Describe the aim of the study, including the object under examination, the focus, and the perspective. Example: "We examined <technique1> vs. <technique2> with regard to fault detection rates from the viewpoint of a quality engineer."

Method: Describe which research method was used to examine the object (e.g., experimental design, number and kind of participants, selection criteria, data collection and analysis procedures). Example: "We conducted a controlled experiment using a 2×2 factorial design with 24 randomly assigned undergraduate students participating. The data were collected with the help of questionnaires and analyzed using ANOVA."

Results: Describe the main findings. Example: "<technique1> was significantly more effective than <technique2> at an alpha level of 0.05."

Limitations: Describe the major limitations of the research, if any. Example: "Generalization of results is limited since the analyzed technique was applied only to specify systems smaller than 10,000 lines of code."

Conclusion: Describe the impact of the results. Example: "The result reinforced existing evidence regarding the superiority of <technique1> over <technique2>."

Furthermore, to address practitioners' information needs, cost, benefits, risks, and transitions should also be described.

Our recommendation to include the element *Limitations* in a structured abstract follows a suggestion made in The Editors of *Annals of Internal Medicine* (2004), since every piece of evidence has its limitations. This additional information helps readers judge the transferability of the results to their context. It also prevents uncritical acceptance by the reader.

It is important to use only a few sentences for each structuring element of the abstract. Hartley (2003) found that the number of words increases by about 30% if structured abstracts are used. But he claims that these "extra costs" pay back because, with the additional information given in the abstract, a wider readership might be encouraged and citation rates improve as do (journal) impact factors. Several researchers who compared the use of structured abstracts to traditional ones found advantages for structured abstracts, but no real disadvantages (Hartley, 2004; Kitchenham, 2004).

From this discussion, we conclude that experimenters should certainly use structured abstracts, but even if the abstract is written as text (without structuring elements), it should still include all of the aforementioned elements. Where publishers limit the length of the abstract by number of words or number of lines, we suggest prioritizing the traditional elements: *background (one sentence), objective, method, results, and conclusion*, but recommend sticking with the structure.

As a final note, to attract readers from industry, authors should use terms that are commonly used in industry in describing their research.

3.4. Keywords

Except for Kitchenham (2004) and Jedlitschka et al. (2007), existing guidelines do not explicitly address keywords. Furthermore, keywords are not necessarily requested by all publications. Nevertheless, if provided (and if free of any predefined characterization, like ACM), keywords should describe the areas of research, the treatments, dependent variables, and study type. The list of keywords should complement the title, as it was described earlier, especially in cases where it was not possible to include all pertinent information in the title. As with the title, keywords help readers to identify relevant publications. This is especially important because publishers use keywords for categorisation, and they are visible even in cases where full access to the publication is restricted. Finally, keywords should not be idiosyncratic, but should instead reflect common terms used in the field.

3.5. Introduction

The purpose of the introduction is to set the scope of the work and give potential readers good reasons for reading the remainder of the publication (motivation). The introduction needs to place the research into a wider context before introducing the specific problem. As can be seen from Table 1, there are several variations with

regard to the content of the introduction. In most cases, the introduction starts with a broad description of the research area (Wohlin et al., 2000). With the exception of Wohlin et al. (2000), who recommend a distinct section to describe the problem under study, all of the guidelines include the description of the problem in the introduction. Further, Wohlin et al. (2000) and Kitchenham et al. (2002) suggest the introduction include an explicit description of the context of the study (i.e., the environment in which it is run).

Thus, based on the various guidelines, as a minimum the introduction should include a description of the *Problem Statement*, the *Research Objectives*, and the *Context* of the research.

The problem statement supports readers in comparing their problems with the problem investigated in the reported experiment, thereby judging the relevance of the research to their questions. In general, the problem statement should provide answers to the following questions: What is the problem? Where does it occur? Who has observed it? Why is it important to be solved? In addition, any underlying theory, causal model, or logical model should be specified.

The description of the problem statement should lead directly to the description of the research objective. The research objective starts with a brief description of the solution idea and the (expected) benefits of the solution.

> Example adopted from (Ciolkowski et al. 1997): Recently, it was reported by […] that defects in a software artefact increase cycle time and development costs. One possible solution would be to start defect detection as early in the development cycle as possible, for example by inspecting requirements documents. The benefit would be that the defects from the requirements phase will not be incorporated in the later phases, which will result in reduced cycle times and development costs.

The description of the research objective (or, as Wohlin et al. (2000) call it, the "Definition of the Experiment"), should be as coherent as possible. One way to achieve this is to use the goal template of the Goal/Question/Metric (GQM) method formulated by Basili et al. (2001). This template includes several elements to be filled in as shown below, with an example underneath.

Analyze <…> for the purpose of <…> with respect to their <…> from the point of view of the <…> in the context of <…>.

The following example is adapted from Ciolkowski et al. (1997):

> Analyze perspective-based reading and ad hoc reading techniques
> For the purpose of evaluation
> With respect to their effectiveness
> From the viewpoint of potential users
> In the context of the software engineering class at the University

For further examples of the use of the goal definition template to describe the research objective, see Wohlin et al. (2000).

The description of the context is essential for practitioners as well as for researchers. Practitioners need context information to see if the technique/process/tool under study would be applicable in their own organization. Researchers need context information to understand the limits of the study (e.g., whether the results are generalizable), to replicate results, and to aggregate results or perform meta-

analyses. To describe the context of the research, the CONSORT Statement (Altman et al., 2001; Moher et al., 2001) suggests that the setting and locations of a study are described. In software engineering this could include information about application type (e.g., real-time system), application domain, (e.g., telecommunications), type of company (e.g., small or medium sized), experience of the participants (e.g., professionals with on average 5 years of related practical experience), time constraints (e.g., critical milestones, delivery date), process (e.g., spiral model), tools (e.g., used for capturing requirements), size of project (e.g., 500 person months). Furthermore, it is valuable to know whether there are specific requirements with regard to the environment in which the technique, tool, or method was applied.

A more formal description of context from a researcher's viewpoint comprises context factors that might affect the generality and utility of the conclusions. These are generally detailed when describing the experimental design.

The introduction generally ends with an outline for the remainder of the paper.

3.6. Background

Researchers as well as practitioners need an understanding of the landscape of the reported research, including alternative approaches and relationships between different experiments (Jedlitschka and Ciolkowski, 2004b). Most guidelines require appropriate citation, as described, for example, in the APA style guide (2001).

In contrast to Singer (1999), who includes background information in the Introduction, Wohlin et al. (2000), Juristo and Moreno (2001), Kitchenham et al. (2002), Jedlitschka and Pfahl (2005a, b), and Jedlitschka et al. (2007) suggest presenting background information in a unique section.

At a minimum, the background should present: a description of the *Technology* (or tool, method)[1] *under Investigation*, a description of *Alternative Solutions*, i.e., other reports that address the same problem or are comparable from a technology view point, a *Description of Related Studies*, i.e., empirical studies that have investigated the same or similar treatments, and, if appropriate, levels of *Relevance to Practice*, i.e., how successfully the technique has been applied in industry. In the following, we provide more details on each of these elements.

Because readers need to understand at some level what is being investigated before they can understand how it relates to other work, the background will frequently begin with a brief description of the treatment and control variables of the experiment. The detail of the description depends on the availability of earlier publications and the length of the report. Moreover, for readers who have no specific background in the area, a more general reference, e.g., to a textbook, might be helpful.

[1] For ease of reading, we use technology as an umbrella term for technology, method, and tool.

The description of alternative solutions/approaches helps to frame the work within a larger research context. This description should not simply be a list of related research (Shaw, 2003), but rather an objective description of the main findings relevant to the work currently being reported. Alternative solutions should be reported whether they are supportive of or contradictory to the current research approach. Especially in the case of an experiment that compares different approaches, it is crucial to objectively describe the alternative approaches. Note that a comparison of the results of related work and the current results should be done in the discussion section after the results have been presented (c.f. Sect. 3.10).

In the description of related studies, existing evidence (if available), in the form of earlier studies and, especially, experiments, should be described. As with alternative solutions, the relation of the current research to other studies (existing evidence) helps readers understand where this work fits into a larger research context. Moreover, it supports the reuse of this study for replication or systematic review, providing a sound basis for research and improving its value. If the reported study is a replication, the parental study and its findings also have to be described.

In terms of relevance to practice, if applicable, if one of the treatments (technologies) has previously been applied to real software projects or under realistic circumstances, a short summary of the findings and related references should be provided.

3.7. Experiment Planning

This section, sometimes referred to as experimental design or protocol, describes the plan or protocol that is used to perform the experiment and analyze the results. It is important because, as Singer stated, this section is the "recipe for the experiment" (Singer, 1999). Therefore, it should provide all information that is necessary to replicate the study and integrate it into the ESE body of knowledge. In addition, this section allows readers to evaluate the internal validity of the study, which is an important selection criterion for systematic review or meta-analysis (Kitchenham, 2004; Kitchenham et al., 2002).

According to several guidelines (e.g., Harris, 2002), the experiment planning section should describe the *Goals*, *Participants*, *Experimental Material, Tasks*, *Hypotheses, Parameters, and Variables*, *Experiment Design*, *Procedure* for conducting the study, as well as the *Analysis Procedure*. Using this order allows for successive refinement of the details of the study. In some cases, however, a different order might be appropriate.

The level of detail regarding the various elements depends on the kind of publication, respecting the required length of the report. Therefore, authors should prioritise the information according to what is most relevant for the particular audience. Alternatively, authors may consider combining several sections into one. For instance, it might be appropriate to integrate the description of the procedure with the description of the execution, or to integrate the description of the analysis

procedure with that of the analysis. Furthermore, it might be possible to put all relevant material into an appendix or longer technical report. If this is not possible, archiving the information on a website may be an alternative. To address concerns that arise in sharing protocols, including raw data and material, Basili et al. (2007) propose an initial licensing model.

3.7.1. Goal(s)

Often the original research objective as described in the introduction is not concrete enough. The purpose of this paragraph is, therefore, to define in more concrete terms the main manipulations of the experiment. For example, the GQM template provided in the introduction could be refined into something like:

Example adapted from Ciolkowski et al. (1997):

Goal 1: Analyze perspective-based reading and ad hoc reading techniques
For the purpose of understanding their effectiveness
With respect to the defect detection rate of individual developers

Goal 2: Analyze perspective-based reading perspectives
For the purpose of understanding their effectiveness
With respect to detecting different defect classes

The refinement of the main research question should be described and motivated to allow for traceability down to the hypotheses, which will be described in later in this chapter.

3.7.2. Participants

The participants (often referred to as subjects or, if not humans, experimental units) need to be described in detail. Furthermore, the sampling strategy and the resulting samples need to be described, including the number of participants (per condition), the kind of participants (e.g., computer science students), and the populations from which they were drawn. All measures for randomization have to be reported here, especially the random allocation of participants to treatments. Where a statistical power calculation has been used, assumptions, estimates, and calculations have to be provided.

All participant characteristics that might have an effect on the results or restrict the sample in some way should also be described in this section. This may include experience with the techniques to be applied or mean/range of experience in years, or educational level. For instance, if a certain level of experience is required, the sample might be drawn from fourth-term computer science students (as opposed to first-term students).

A description of the motivation for the participants to participate is mandatory. For instance, it should be stated whether the participants were paid and if so, how much, or whether they earned educational credits for taking part in the experiment. Additionally, the answers to the following questions are of interest (Wohlin et al.,

2000): What was the commitment of the participants? How was consent obtained? How was confidentiality assured? How was participation motivated (induced)?

3.7.3. Experimental Materials

In this section, all experimental materials and equipment should be described. For instance, if the study involves a questionnaire, questions should be described, as should any other characterizations of the questionnaire, e.g., it had five sections focusing on specific topics, with the topics named. As another example, in an experiment looking at different reading techniques, the document used for the application of the reading technique should be described in terms of its length, complexity, seeded faults (number, type, interactions), etc. As with the participant section, all characteristics that might have an impact on the results should be mentioned here as formally as possible. However, in case of conference papers, it is often not possible to present all the materials in detail, so we suggest providing more detail either in the appendix of an associated technical report, or using a website.

Note that in this section, the materials should not be presented verbatim, but rather described with as much detail as necessary for the readers to understand what materials the participants interacted with during the experiment.

3.7.4. Tasks

Here, the tasks performed by the participants should be described in enough detail so that a replication of the experiment is possible without consultation of the authors. Redundancies with regard to the description of the technology in the background section (c.f., Sect. 3.6) should be avoided. If the description requires too much space, the information should be made available in a technical report or as a web resource. When space is a consideration, the task description could be integrated with the description of the procedure. However, separating the two descriptions makes it easier for readers to understand how the hypotheses, parameters, and variables were derived.

3.7.5. Hypotheses, Parameters, and Variables

In this section, hypotheses, parameters, and variables should be described. This description should be linked to the research objective already reported in the introduction.

For each goal stated in the research objective, the null hypotheses, denoted H_{0ij}, and their corresponding alternative hypotheses, denoted H_{1ij}, need to be reported, where i corresponds to the goal identifier, and j is a counter for cases where more

than one hypothesis is formulated per goal. The description of both null and alternative hypotheses should be as formal as possible. The main hypotheses should be explicitly separated from ancillary hypotheses and exploratory analyses. In the case of ancillary hypotheses, a hierarchical system is appropriate. Hypotheses need to state the treatments and the control conditions.

Continuing the example for Goal1 from Sect. 3.7.1 (adapted from Ciolkowski et al. (1997)):

> The goal of the experiment is to determine:
> Q1: Which reading technique produces a higher mean defect detection rate?
> One of the possible hypotheses is:
> H_{011}: Individuals applying a perspective-based reading (PBR) technique detect more defects than individuals using ad hoc reading.

In the example hypothesis H_{011}, the treatment is perspective-based reading and the control condition is ad hoc reading. A further formalization of H_{011} and the alternative hypothesis H_{111} could be written in the following form (where MDDR stands for mean defect detection rate):

$$H_{011} = \text{MDDR(PBR)} > \text{MDDR(ad hoc)}$$

$$H_{111} = \text{MDDR(PBR)} \leq \text{MDDR(ad hoc)}$$

It is important to differentiate between experimental hypotheses and the specific tests being performed; the tests have to be described in the analysis procedure section.

In addition to the hypotheses, there are two types of variables that need to be described in this section: the dependent variable(s) (aka. response variables) and the independent variable(s) (aka. predictor variables). As with the hypotheses, dependent variables need be defined and justified in terms of their relevance to the goals listed in the *Research Objectives*. Dependent variables are the variables that are measured to ascertain whether the independent variable had an effect on the outcome. Likewise, independent variables are variables that are frequently manipulated in the experiment and may influence the dependent variable(s). Independent variables can include treatments, materials, and some context factors. In this section, only independent variables that are manipulated or controlled through the experimental design (i.e., causal variables) are described. For each independent variable, its corresponding levels (aka. alternatives, treatments) have to be specified in operational form. In the example given above, the dependent variable is the MDDR. The independent variable is the type of reading technique, which has two levels, PBR and ad hoc.

With respect to reporting, authors need to describe their metrics clearly. In particular, if a standardized set of metrics is available, authors have to explain which of them are used. If existing metrics are tailored, the need for the tailoring and the tailored metric have to be explicated. Based on Wohlin et al. (2000), Juristo and Moreno (2001), and Kitchenham et al. (2001), Table 3 gives a schema for the description of variables and related metrics.

Table 3 Schema for the description of variables

Name of the variable	Type of the variable (independent, dependent, moderating)	Abbreviation	Class (product, process, resource, method)	Entity (instance of the class)	Type of attribute (internal, external)	Scale type (nominal, ordinal …)	Unit	Range or, for nominal and restricted ordinal scales, the definition of each scale point	Counting rule in the context of the entity
Type of reading technique	independent	RT	Method	Reading Technique	N.A.	nominal	N.A.	PBR; ad hoc	N.A.
Mean defect detection rate	dependent	MDDR	Process	Inspection process	Internal: efficiency; external: quality	ratio	Number of defects per hour	>= 0	Number of agreed upon defects after review meeting / total effort for inspection process in hours

For subjective metrics, a statistic for inter-rater agreements should be presented, such as the kappa statistics or the intra-class correlation coefficient for continuous metrics (Kitchenham et al., 2002).

3.7.6. Experiment Design

In the *Experiment Design* subsection, the specific design has to be described. Elements in this section that need to be described include whether the experiment was a within – or between-subjects design, or a mixed factors design, with a description of each of the levels of the independent variable. Juristo and Moreno (2001) give a comprehensive description of designs for experiments. Moreover, authors should describe how participants were assigned to levels of the treatments (Kitchenham et al., 2002).

If, for example, an experiment examined the effect of PBR versus ad hoc reading techniques on short and long times spent looking for defects on MDDR, with different sets of subjects using the techniques, it would be reported as a 2 (reading technique) × 2 (time period) between-subjects design with reading technique having two levels: PBR and ad hoc, and time also having two levels (15 min and 30 min).

In addition to this formalization of the design, if any kind of blinding (e.g., blind allocation) has been used, the details need to be provided; this applies to the execution (e.g., blind marking) and the analysis (e.g., blind analysis). If the experiment is a replication, the adjustments and their rationales need to be discussed. If applicable, training provided to the participants has to be described. Any kind of threat mitigation should also be addressed, i.e., what measures were used to manage treats to validity. For example, a typical strategy to reduce learning effects is to have subjects exposed to the various levels of a treatment in a random or ordered fashion.

3.7.7. Procedure

The procedure section should describe precisely what happened to the participants from the moment they arrived to the moment they left (Harris, 2002). This includes a description of any training provided (e.g., the participants received a 2-h lecture introducing perspective-based reading). The procedure section should also include a description of the setting (i.e., where the experiment occurred), and the schedule for the experiment. Furthermore, details of the data collection method have to be described, including when the data was collected, by whom, and with what kind of support (e.g., tool). This is in accordance with Kitchenham et al. (2002), who state that the data collection process describes the "who," the "when," and the "how" of any data collection activity. Any type of transformation of the data (e.g., marking "true" defects in defect lists) and training provided for such should also be described

here. If there are limitations with regard to the numbers of pages, the description of the procedure can be integrated with the analysis section.

3.7.8. Analysis Procedure

The statistical tests undertaken depend on the experimental design; therefore, the experimental plan is finalized with a description of the analysis procedure detailing which methods were used to test the hypotheses in analysing the data. If different hypotheses are investigated, information for each hypothesis needs to be provided separately. If any additional influences are expected, their analysis also needs to be described, e.g., see Ciolkowski et al. (1997). If there are page limitations, the analysis procedure can be combined with the analysis section.

3.8. Deviations from the Plan

In an ideal situation, the experiment was conducted exactly as it was planned. Then the description in the procedure section (c.f., Sect. 3.7.7) is both, the representation and the instantiation of the plan. In that case, this section is not needed. However, deviations regarding the original plan are often experienced. Because this might have an impact on both the validity of the results and the replicability of the study, it is necessary to describe those deviations by describing the original plan when deviations occurred. This includes all differences between the instantiated procedure and the plan, for instance, regarding instrumentation and the collection process. Deviations can occur regarding participation (who actually participated), schedule (e.g., the time participants were given for the tasks), or data collection. In addition, information about subjects who do not complete the study should be presented, for example, five subjects did not attend the final session; as recommended by Kitchenham et al. (2002). If possible, reasons for the non-completion should be given; that information is worthwhile when replicating the study.

In the case of a limited number of pages, this description can be integrated with the procedure section (c.f. Sect. 3.7.7). In addition, a general statement confirming the process conformance could be given in the description of the analysis.

3.9. Analysis

According to Singer (1999), the *Analysis* section summarizes the data collected and its treatment. In this section, the results should be described devoid of any interpretation. When there are limited pages, authors might tend to add some interpretation to the analysis section. However, according to existing guidelines, especially from other disciplines, interpretation and results belong to clearly distinct sections. If it

is necessary to include interpretation in the analysis section, we strongly favour establishing a clear distinction between the two (e.g., by using textual measures or subsections).

If multiple goals were investigated, separate analysis subsections and an overall (summarizing) analysis are required. Since the analysis procedures are already described in the design section, the purpose of this section is to describe the application of the analysis methods to the data collected. The Analysis section generally contains three types of information: *Descriptive Statistics*, *Data Set Preparation*, and *Hypothesis Testing*. When appropriate, a sensitivity analysis should be reported in the hypothesis testing section.

Presenting the data by using appropriate descriptive statistics, including number of observations, measures for central tendency, and dispersion, gives the reader an overview of the data. Mean, median, and mode are example measures for central tendency. Standard deviation, variance, and range, as well as interval of variation and frequency are example measures for dispersion. To facilitate meta-analysis, it is highly recommended [e.g., by Kitchenham et al. (2002)] to provide raw data in the appendices or to describe where the data can be acquired, e.g., from a website.

Additional processing (or preparation) of the data set may be required. Such preparations should be discussed here. This includes, if appropriate, data transformation, outlier identification and their potential removal, and handling of missing values, as well as the discussion of dropouts (i.e., data from participants who were not present for all experimental sessions). Chap. 7 details methods for dealing with missing values.

For hypothesis testing, special emphasis should be placed on how the data was evaluated (e.g., by an ANOVA) and how the analysis model was validated. The violations of the statistical assumptions underlying the analysis method (e.g., normality, independence, and residuals) should also be described. The values of the resulting statistics also need to be reported. Harris outlines what has to be reported for different kinds of statistical tests (Harris, 2002). Singer (1999) recommends that "inferential statistics are reported with the value of the test (effect size), the probability level, the degrees of freedom, the direction of effect," and the power of the test. To this list, we add the alpha value and the confidence interval where appropriate (Dybå et al., 2006; Kampenes et al., 2007).

3.10. Discussion

The purpose of the discussion section is to interpret the findings presented in the previous section. This includes an overview of the results, threats to validity, generalization (where are the results applicable?), as well as the (potential) impact on cost, time, and quality. Harris (2002) suggests starting this section with a description of what has been found and how well the data fit the predictions. Related to this, authors should discuss whether the hypotheses were confirmed or not. The discussion

section should include information about each of the following three elements: *Evaluation of Results and Implications*, *Threats to Validity*, and *Inferences*.

3.10.1. Evaluation of Results and Implications

The purpose of the evaluation of results and implications is to explain the results. All findings, including any unexpected results, should be described in this subsection. Moreover, if the null hypothesis was not rejected, authors may include reasons for why they believe this is the case. Several authors point out that it is important to distinguish between statistical significance and practical importance (Kitchenham et al., 2002) or meaningfulness (Harris, 2002). The results should also be related to both theory and practice.

Although it is still very rare for SE experiments to develop theory, the implications of the findings should be related to the larger theory being developed, and how they further explicate or illuminate that theory (see Chap. 12 for more information about theory). The results should be discussed in the light of the objectives stated in the introduction and also related to the previous work described in the background section. These two together should help to build a broader theoretical foundation for the work.

With respect to practice, the results should be related to current and potential practice, outlining how practice can be improved by applying the results. If the null hypothesis was not rejected, it is not possible to give an interpretation in any direction; in particular, it does not mean that the null hypothesis is true, only that not enough evidence exists to reject it. In some cases, the value of the effect is so small that there may actually be no relevant application to current practice. This has to be explicated as well.

In writing the discussion, it is important to (1) clearly state the results of the analysis separately from any inferences or conclusions based on those results (Kitchenham et al., 2002), (2) to ensure that the conclusions follow from the results (Kitchenham et al., 2002), and (3) that conjectures be made with caution and kept brief, leaving out fanciful speculation (Harris, 2002).

3.10.2. Threats to Validity

All threats that might have an impact on the validity of the results need to be discussed. This includes at least (1) *threats to construct validity*, (2) *threats to internal validity*, (3) *threats to external validity*, and if applicable, and (4) *threats to conclusion validity*. A more comprehensive classification of threats to validity is given in Wohlin et al. (2000). Each of these four types of threats to validity is defined below, and needs to be covered in a research paper. Ignoring the threats can lead to the wrong conclusions regarding the validity of the results. For example, a practitioner might assume that the results would apply

to his situation where the external validity could indicate problems regarding generalizability.

Construct validity. Construct validity refers to the degree to which the operationalization of the measures in a study actually represents the constructs in the real world. For instance, in measuring readability, a researcher may look at the time required to read source code. The construct validity of this measure is the extent to which the readability of source code is actually related to the time required to read it. There are a number of threats to construct validity outlined in Wohlin et al. (2000).

Internal validity. Internal validity refers to the extent to which the treatment or independent variable(s) were actually responsible for the effects seen to the dependent variable. Unknown factors may have had an influence on the results and therefore put limitations on the internal validity of the study. Note that it is possible to have internal validity in a study and not have construct validity. For instance, it could be true that the manipulations in the study did actually affect the outcome, and yet the manipulations did not map/represent the desired entity in the real world.

External validity. External validity refers to the degree to which the findings of the study can be generalized to other participant populations or settings. External validity can often be a problem for controlled experiments in artificial environments where the same conditions may not hold in the real world. Wohlin et al. describe three types of threats to internal validity dealing with people, place, and/or time.

Conclusion validity. Conclusion validity refers to whether the conclusions reached in a study are correct. For controlled experiments, conclusion validity is directly related to the application of statistical tests to the data. If the statistical tests are not applied correctly, this is a threat to the conclusion validity. Thus, examples of threats to conclusion validity involve anything that causes a Type I or Type II error.

To facilitate reading, subsections might be appropriate for each threat that has to be discussed. Following the arguments presented by Kitchenham et al. (2002), it is not enough to mention that a threat exists; the implications of the threat with respect to the findings also need to be discussed.

Other threats than those listed above may also need to be discussed, such as personal vested interests or ethical issues regarding the selection of participants (in particular, experimenter-subject dependencies).

3.10.3. Inferences

In this section, the findings can be generalized, within the scope of validity, to broader research questions or settings. This should be done carefully, based on the

findings, by incorporating the limitations. All claims need to be supported by the results. For technologies not currently in use, scale-up issues should be discussed.

3.11. Conclusions and Future Work

The final section of the report should describe, based on the results and discussion, the following elements: *Summary, Impact*, and *Future Work*.

The conclusion section begins with a concise summary of the research and its results as presented in the former sections. Unique to the domain of software engineering – in order to enable readers to get the most important findings with regard to the practical impact in one place – in the conclusion we emphasize a description, where possible, of the impact on cost, time, and quality, and a summary of the limitations. Note that these conclusions can only be drawn if they were directly investigated in the experiment.

Impact on Cost: What effort was necessary to introduce and perform the technique (e.g., what are the costs of detecting a defect of a certain type with this technique? Is there any impact on the cost of other steps of the development process, positive or negative ones (e.g., reduced cost for rework)?)

Impact on Time: Is there any positive or negative impact on the time of the proposed solution/technology/technique on other steps of the development process?

Impact on Quality: Is there any impact on the quality of the proposed solution/technology/technique on the quality of other steps of the development process?

Besides the description of the impact, where possible and appropriate, a discussion of the approach's level of maturity, when the investments will pay back, and consequences arising from the implementation will help readers to assess the technology. (Although in most cases artificial, we assume a rough estimate is better than no information.)

If applicable, *limitations* of the approach with regard to its practical implementation should also be described, i.e., circumstances under which the approach presumably will not yield the expected benefits or should not be employed. Furthermore, any risks or side-effects associated with the implementation or application of the approach should also be mentioned.

Finally, an outlook to future work should be given. It should describe what other research (i.e., experiments) could be carried out to further investigate the results yielded or evolve the body of knowledge and theoretical constructs.

3.12. Acknowledgements

In this section, sponsors, participants, and (research) contributors who do not fulfil the requirements for authorship should be mentioned.

3.13. References

In this section, all cited literature has to be presented in the format requested by the publisher.

3.14. Appendices

In this section, material, raw data, and detailed analyses that might be helpful for others to build upon the reported work should be provided (i.e., meta-analysis).

If the raw data is not reported, the authors should specify where and under which conditions the material and the raw data could be made available to other researchers (i.e., technical report, web resource). Here a license model, such as the one proposed by Basili et al. (2007) can be used to ensure to all parties that their contribution is acknowledged and the material is only used for the defined purposes. The licensor can, for example, require that any publication based on the delivered data has to be sent to him.

4. Conclusion

In this chapter, we have motivated the importance of reporting standards for maturing empirical software engineering research. The contribution of this chapter is a guideline for guiding researchers while reporting experiments in software engineering. The presented guideline unifies and extends the most prominent existing guidelines published by various authors (cf. Table 1). In addition to providing a uniform structure of a reporting template, the guideline provides detailed guidance on which information should be provided in the various sections of a report. This guideline was developed for a specific type of empirical study, i.e., controlled experiments and quasi-experiments. Nevertheless, many aspects discussed throughout this chapter have to be reported in other empirical study reports, like case studies.

Thus, this chapter provides researchers with a means for structured and comprehensive documentation of empirical studies, especially experiments. In some cases, due to page limitations (e.g., conference paper), it might not be possible to provide all the proposed information. Although each paper should stand for itself, we have discussed possible shortcuts by integrating certain sections. Furthermore, authors should make use of technical reports or web resources to provide additional information, including material, raw data, and detailed analysis.

During our work on guidelines, we learned that issues are related not only to structure and comprehensiveness, but also to the information needs of stakeholders. In this chapter, we presented, from our perspective, a quite comprehensive model, addressing several stakeholders. To especially attract decision makers in industry, we envisage tailoring this guideline for different audiences (e.g., by providing a

guideline for reporting results from empirical research to practitioners). Researchers doing replications or performing a systematic review certainly have different information needs than practitioners looking for candidate techniques for solving their problems. Researchers need more technical information regarding the study as such, whereas practitioners require information regarding the potential of the technique to actually solve their problems; that is, information on development costs, product quality, and development schedule.

An important issue related to the dissemination task is to ensure that the guidelines are used in research practice. One possibility to enforce the usage of reporting guidelines could be that program committees of SE workshops and conferences as well as editorial boards of SE journals make the application of a standard reporting scheme mandatory.

To facilitate the adoption of the guidelines, it would help to stress the benefits that accrue to researchers who apply them. For example, one benefit could be simpler integration of individual results into a common body of knowledge. We also assume that, generally, the SE publication process will become more efficient, since crucial information will be found by reviewers (and other researchers) in the same place every time.

Thus, we would like to conclude this chapter with a call for adherence to guidelines. Whenever reporting results of any kind of empirical studies, it is wise to think about who shall read the publication for what purposes. This way, the report will deliver the information needed for different stakeholder groups and audiences. The guidelines will assist writers to emphasize the right information and the empirical software engineering community to mature.

Acknowledgements While preparing the guidelines, we got valuable feedback from many people. Only the names of some of them can be listed here. We thank Janice Singer for her feedback and support, while finalizing this chapter, the unknown reviewers of the preliminary version of this chapter, Claes Wohlin, who gave valuable insights and comments on an earlier version of the guidelines, Barbara Kitchenham and her team at NICTA for their valuable feedback from the perspective-based reading of an earlier version, which helped to improve the guidelines, and many others from the International Software Engineering Research Network (ISERN) for fruitful discussions. Furthermore, we are grateful to Sonnhild Namingha from Fraunhofer IESE for reviewing a previous version of this chapter.

References

Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gøtzsche, P.C., Lang, T. for the CONSORT Group (2001). The Revised CONSORT Statement for Reporting Randomized Trials, Explanation and Elaboration. Annals of Internal Medicine, Vol. 134, No. 8, pp. 663–694.

American Psychological Association (2001). Publication Manual of the American Psychological Association, 5th edn, American Psychological Association, Washington, DC.

Basili, V.R., Caldiera, G., Rombach, H.D. (2001). Goal Question Metric Paradigm, in Marciniak, J.J. (Ed.), *Encyclopedia of Software Engineering*, Vol. 1, Wiley, New York, pp. 528–532.

Basili, V.R., Zelkowitz, M., Sjøberg, D.I.K., Johnson, P., Cowling, T. (2007). Protocols in the use of Empirical Software Engineering Artifacts. Journal of Empirical Software Engineering, 12(1), pp. 107–119.

Bayley, L., Eldredge, J. (2003). The Structured Abstract, An Essential Tool for Researchers, In Hypothesis. The Journal of the Research Section of the Medical Library Association, Vol. 17, No. 1, 4 pp.

Ciolkowski, M., Differding, C., Laitenberger, O., Münch, J. (1997). Empirical Investigation of Perspective-based Reading, A Replicated Experiment, Fraunhofer Institute for Experimental Software Engineering, Germany, ISERN-97-13.

Dybå, T., Kampenes, B.V., Sjøberg, D.I.K. (2006). A Systematic Review of Statistical Power in Software Engineering Experiments, A Survey of Controlled Experiments in Software Engineering. Information and Software Technology, Vol. 48, pp. 745–755.

Glass, R.L. (2004). Matching Methodology to Problem Domain. Communications of the ACM, Vol. 47, No. 5, pp. 19–21.

Harris, P. (2002). Designing and Reporting Experiments in Psychology, 2nd edn, Open University Press, Buckingham.

Hartley, J. (2003). Improving the Clarity of Journal Abstracts in Psychology, The Case for Structure. Science Communication, Vol. 24, No. 3, pp. 366–379.

Hartley, J. (2004). Current Findings from Research on Structured Abstracts. Journal of the Medical Library Association, Vol. 92, No. 3, pp. 368–371.

Hayward, R.S.A., Wilson, M.C., Tunis, S.R., Bass, E.B., Rubin, H.R., Haynes, R.B. (1993). More Informative Abstracts of Articles Describing Clinical Practice Guidelines. Annals of Internal Medicine Vol. 118, No. 9, pp. 731–737.

Jedlitschka, A. (2005). Minutes from Third International Workshop on Empirical Software Engineering "Guidelines for Empirical Work in Software Engineering". IESE-Report 052.05/E, Oulu.

Jedlitschka, A., Ciolkowski, M. (2004). Towards Evidence in Software Engineering, In Proceedings of ACM/IEEE International Symposium on Software Engineering 2004 (ISESE2004). Redondo Beach, California, pp. 261–270.

Jedlitschka, A., Pfahl, D. (2005a). Reporting Guidelines for Controlled Experiments in Software Engineering. IESE-Report IESE-035.5/E.

Jedlitschka, A., Pfahl, D. (2005b). Reporting Guidelines for Controlled Experiments in Software Engineering, In Proceedings of ACM/IEEE International Symposium on Software Engineering 2005 (ISESE2005). Noosa Heads, Australia, pp. 95–104.

Jedlitschka, A., Ciolkowski, M. (2006). Reporting Guidelines for Controlled Experiments in Software Engineering, Fraunhofer Institute for Experimental Software Engineering, Germany, ISERN-06-1.

Jedlitschka, A., Ciolkowski, M. Pfahl, D. (2007). Reporting Guidelines for Controlled Experiments in Software Engineering, Fraunhofer Institute for Experimental Software Engineering, Germany, ISERN-07-1.

Juristo, N., Moreno, A. (2001). Basics of Software Engineering Experimentation, Kluwer Academic Publishers, Boston, MA.

Kampenes, B.V., Dybå, T., Hannay, J., Sjøberg, D.I.K. (2007). A Systematic Review of Effect Size in Software Engineering Experiments. Information and Software Technology, Vol. 49, No. 11–12, pp. 1073–1086.

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews, Keele University Joint Technical Report TR/SE-0401, ISSN,1353–7776 and National ICT Australia Ltd. NICTA Technical Report 0400011T.1.

Kitchenham, B., Al-Khilidar, H., Ali Babar, M., Berry, M., Cox, C., Keung, J., Kurniawati, F., Staples, M., Zhang, H., Zhu, L. (2006). Evaluating Guidelines for Empirical Software Engineering Studies, In Proceedings of ACM/IEEE International Symposium on Software Engineering 2006 (ISESE2006).

Kitchenham, B., Dybå, T., Jørgensen, M. (2004). Evidence-Based Software Engineering, In Proceedings of 26th International Conference on Software Engineering (ICSE'04), pp. 273–281.

Kitchenham, B.A., Hughes, R.T., Linkman, S.G. (2001). Modeling Software Measurement, IEEE Transactions on Software Engineering, Vol. 27, No. 9, pp. 788–804.

Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K., Rosenberg, J. (2002). Preliminary Guidelines for Empirical Research in Software Engineering, IEEE Transactions on Software Engineering, Vol. 28, No. 8, pp. 721–734.

Lott, C.M., Rombach, H.D. (1996). Repeatable Software Engineering Experiments for Comparing Defect – Detection Techniques, Empirical Software Engineering Journal, Vol. 3.1, pp. 241–277.

Moher, D., Schulz, K.F., Altman, D. for the CONSORT Group (2001). The CONSORT Statement, Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials, Journal of the American Medical Association (JAMA) Vol. 285, No. 15, pp. 1987–1991.

Pickard, L.M., Kitchenham, B.A., Jones, P.W. (1998). Combining Empirical Results in Software Engineering, Information and Software Technology, Vol. 40, No. 14, pp. 811–821.

Shaw, M. (2003). Writing Good Software Engineering Research Papers – Minitutorial, In Proceedings of the 25th International Conference on Software Engineering (ICSE'03). IEEE Computer Society, Portland, Oregon, pp. 726–736.

Shiffman, R.N., Shekelle, P., Overhage, J.M., Slutsky, J., Grimshaw, J., Deshpande, A.M. (2003). Standardized Reporting of Clinical Practice Guidelines, A Proposal from the Conference on Guideline Standardization, Annals of Internal Medicine, Vol. 139, No. 6, pp. 493–498.

Shull, F., Carver, J., Travassos, G.H., Maldonado, J.C., Conradi, R., Basili, V.R. (2003). Replicated Studies, Building a Body of Knowledge about Software Reading Techniques, In Juristo, N., Moreno, A. (Eds.), Lecture Notes on Empirical Software Engineering, World Scientific Publishing, River Edge, NJ, USA, pp. 39–84.

Singer, J. (1999). Using the APA Style Guidelines to Report Experimental Results, In Proceedings of Workshop on Empirical Studies in Software Maintenance, pp. 71–75. (dec.bmth.ac.uk/ESERG/WESS99/singer.ps)

Sjøberg, D.I.K., Hannay, J., Hansen, O., Kampenes, B.V., Karahasanovic, A., Liborg, N.-K., Rekdal, A. (2005). A Survey of Controlled Experiments in Software Engineering. Transactions on Software Engineering, Vol. 31, No. 9, pp. 733–753.

The Editors of Annals of Internal Medicine (2004). Addressing the Limitations of Structured Abstracts (Editorial). Annals of Internal Medicine, Vol. 140, No. 6, pp. 480–481.

Vegas, S., Juristo, N., Basili, V. (2003). A Process for Identifying Relevant Information for a Repository, A Case Study for Testing Techniques. In Aurum, A., Jeffery, R.,Wohlin, C., Handzic, M. (Eds.). *Managing Software Engineering Knowledge*, Springer-Verlag, Berlin, pp. 199–230.

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, River Edge, NJ, USA, pp. 85–132.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A. (2000). Experimentation in Software Engineering – An Introduction, Kluwer Academic Publishers, Boston, MA.

Zelkowitz, M.V., Wallace, D.R., Binkley, D.W. (2003). Experimental Validation of New Software Technology. In Juristo, N., Moreno, A. (Eds.), Lecture Notes on Empirical Software Engineering, World Scientific Publishing, River Edge, NJ, USA, pp. 229–263.