This is a preliminary version of a chapter in Shull, F., Singer, J., and Sjøberg, D.I. (eds.); *Advanced Topics in Empirical Software Engineering*, Springer, 2007. (to appear)

Reporting Experiments in Software Engineering

Andreas Jedlitschka*, Marcus Ciolkowski*, and Dietmar Pfahl**

**Fraunhofer Institute for Experimental Software Engineering*
Fraunhofer-Platz 1
67663 Kaiserslautern, Germany
[surname.name]@iese.fraunhofer.de

***University of Calgary*
Schulich School of Engineering
ICT 540, 2500 University Dr. N.W.
Calgary, Alberta T2N 1N4, Canada
dpfahl@ucalgary.ca

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 checklist 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 checklist 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 paper presents a unification of a set of guidelines for reporting experiments in software engineering.

Limitation: The checklist has not been evaluated broadly, yet.

Conclusion: The resulting checklist provides detailed guidance on the expected content of the sections and subsections for reporting a specific type of empirical studies, 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 [1], 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 aks for:

- a taxonomy of available methodologies, based upon their strengths and weaknesses;
- 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).

The evidence-based software engineering (EBSE) paradigm [2] promises to solve some of these issues partly by providing a framework for goal-oriented research, leading to a common body of knowledge and, based on that, comprehensive problem-oriented decision support regarding SE technology selection.

One major problem for integrating study results into a common body of knowledge, e.g., by performing systematic reviews [3], is the heterogeneity of study reporting [4]: (1) It is difficult to find relevant information because the same type of information is located in different sections of

different study reports; (2) important information is often missing - for example, context information is reported differently and without taking into account further generalizability. Furthermore, specific information of interest for practitioners is often missing, like a discussion on the overall impact of the technology on the project or business goals.

One way to avoid heterogeneity is to introduce and establish reporting guidelines. More generally speaking, reporting guidelines are expected to 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 [5], 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 already identified by Kitchenham et al. [6], reporting guidelines are necessary for all relevant kinds of empirical work, and they have to 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, e.g., [4], [7], [8], [9], [10], [11], [12], [13]. At the same time, several reporting guidelines have been proposed, e.g., [14], [6]. Even though each of these proposals has its merits, none of these proposals 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. [15]).

The first version of a guideline for reporting controlled experiments [16] was presented during a workshop [17]. Feedback from the workshop participants as well as from peer reviews was incorporated in the second version of the guideline that has been presented at the International Symposium on Empirical Software Engineering (IESE2005) [18]. In parallel the guideline was evaluated by means of a perspective-based inspection approach [19]. The authors of this evaluation claim to have detected 42 issues where the guideline would benefit from amendment or clarification and eight defects. The feedback from this report and direct discussion with some of the authors led to a second iteration of the guideline, where we have incorporated the amendments if appropriate and removed defects that we agreed upon.

The goal of this paper is to survey the most prominent published proposals for reporting guidelines and to derive a unified and – where necessary – enhanced checklist for reporting experiments comprising of controlled experiments and quasi-experiments.

2. Related Work

Empirical software engineering research is not the first area encountering problems with regard to extracting crucial information from empirical research and to insufficient reporting. Other disciplines, such as medicine and psychology, have experienced similar problems before and have achieved various improvements by standardizing and instantiating reporting guidelines, e.g., for randomized controlled trials in biomedical research [5], [20], psychology [21], clinical practice guidelines [22], and empirical results from psychological research [23].
In the field of software engineering (SE) research, in 1999, Singer [14] described how to use the "American Psychological Association (APA) Styleguide" [23] for publishing experimental results in SE. In 2001, Kitchenham et al. [6] 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. In 2003, Shaw [24] 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

Table 1. Characterization of different proposal of guidelines for empirical SE

	Singer [14]	Wohlin et al. [25]	Kitchenham et al. [6]	Juristo and Moreno [26]	Kitchenham [3]	New Proposal
Title	*	*	*	*	Title	Title
Authorship	*	*	*	*	Authorship	Authorship
Keywords		*	*	*	Keywords	Keywords
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	Abstract	*	*	*	Executive Summary or Structured Abstract	Structured Abstract
	Introduction	Introduction	*	Goal Definition	Background	Introduction
		Problem Statement				
		Experiment Planning	Experimental Context			
	Introduction	Problem Statement	Experimental Context	Goal Definition	Background	Related Work
	Method	Experiment Planning	Experimental Design	Design	Review Questions	Experiment Planning
					Review Methods	
	Procedure	Experiment Operation	Conducting the Experiment and Data Collection	Experiment Execution	Included and Excluded Studies	Execution
	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 & Future Work
					Acknowledgments	Acknowledgements
					Conflict of Interest	
	References	References	*	*	References	References
	Appendices	Appendix			Appendices	Appendices

SE, such as Wohlin et al. [25] and Juristo and Moreno [26], address the issue of reporting guidelines. Wohlin et al. suggest an outline for reporting the results of empirical work. Juristo and Moreno provide a list of "most important points to be documented for each phase" in the form of "questions to be answered by the experimental documentation".
Table 1 gives a characterization of 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 last column of the table contains our proposal of a unified and enhanced checklist for reporting controlled experiments. 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 the type of study in focus. The remaining rows list the structuring elements as they are mentioned 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 disjoint parts of these elements can be mapped to two different elements of our new proposal.
An asterix (*) indicates that the authors do not explicitly mention or describe details for this element, but it is assumed that the elements are implicitly required.

3. Checklist for reporting Controlled Experiments

Our work started with the collection and integration of existing guidelines. During the evaluation, the issue of redundancies arose, therefore, we advise authors to reduce redundancies. The structure of the report as presented in this section provides options, especially with regard to the description of the design and the execution. For a conference paper (which is usually much shorter than a journal paper) it is proposed to combine the description of the experiment planning and the execution as well as the description of the analysis procedure and the analysis. For a journal paper it is requested to separate the concerns of these sections. The first ones describe a plan where as the latter ones describe deviations from the plan and what actually has happened.
As indicated in Table 1, the resulting reporting guideline comprises the following elements: *Title*, *Authorship*, *Structured Abstract, Keywords*, *Introduction, Related Work, Experiment Planning, Execution, Analysis, Discussion, Conclusion and Future Work, Acknowledgements, References,* and *Appendices*. The structuring elements are discussed in detail in the following subsections. The classical structure as proposed for reporting experiments in psychology (e.g., by APA [23] and Harris [27]), comprise the following sections: Title, Abstract, Introduction, Method (composed of subsections for Design, Participants, Apparatus or Material, and Procedure), Results, Discussion, References, and Appendices. Singer has adopted this structure in her proposal. But as shown in Table 1, there is no common agreement for reporting SE experiments. 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 guidelines incorporates wording as it is common for experiments in empirical SE to also support the reading of already published reports.
One general remark with regard to the level of detail of the report is that all information has to be provided that would enable someone else to judge the reliability of the experiment, e.g., by replicating the study (in the best case, without approaching the author). The need for detailed provision of information is not specific for SE. It is, for example, also pointed out by Harris [27]. 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 that intention in mind while compiling the report.

3.1 Title

The title of the report has to be informative, first off all, because as Harris [27] describes it the title (together with the abstract) "alert potential readers to the existence of an article of interest to him". In order to attract readers from industry it would be of a great benefit to use commonly used industry terms rather than artificial wordings and abbreviations. The treatments and dependent variables have to be specified avoiding unnecessary redundancy. In that sense, Harris [27] 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, to add "- a controlled experiment" (- a replicated controlled experiment, - a quasi experiment) if there are no limitations with regard to the title length. This will help the reader to easily identify controlled experiments in future. If the title length is limited it is more important to include treatments and the dependent variables.

3.2 Authorship

All individuals making a significant contribution should be in the author list or at least acknowledged (c.f. Section 3.12).
Most report styles require contact details. If not so, provide at least the e-mail of the responsible author. In case the author is likely to move provide sustaining contact information or to be on the save 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, often is the only part of a publication that is freely accessible [3]. Regardless of the format of the abstract, authors have to ensure that all relevant interventions, or conditions (i.e. independent variables) and dependent variables are mentioned.
The exact format of the abstract needs more discussion. For example, Shaw 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 [24]. For reporting experiments in psychology, Harris [27] suggests for an abstract the following aspects to be described: (1) the problem under investigation, (2) the participants, (3) the empirical method, (4) the findings, and (5) the conclusions.
In other disciplines, e.g., medicine and psychology, a special form of the abstract, the so-called structured abstract [33], has been imposed on authors by a huge number of journals in order 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 suggest using a structured abstract consisting of the elements listed below:
Background: Give a brief introducing notice about 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" [22].
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 2x2 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 need regarding cost, benefits, risks, and transition issues shall be described.
The recommendation to include the element *Limitations* follows a suggestion made in [29], since every piece of evidence has its limitations. This additional information will help readers judge transferability of the results to their own context. It will also help to prevent uncritical acceptance by the reader [29].
It is important to use only a few sentences for each structuring element of the abstract. Hartley [30] found that the number of words will increase by about 30% if structured abstracts are used. But he claims that these "extra costs" will pay back because, with the additional, valuable information given in the abstract, a wider readership might be encouraged and increasing citation rates will improve (journal) impact factors. Several researchers who compared structured abstracts with traditional ones found advantages with regard to the information, but no real disadvantages [31], [3].
To attract readers from industry it would be of a great benefit to use commonly used industry terms rather than artificial wordings and abbreviations.
Some publishers limit the length of the abstract by number of words or number of lines. In this case we suggest prioritizing the traditional elements: *background (one sentence), objective, method, results, and conclusion* and to omit the structuring words by keeping the structure.

3.4 Keywords

A list of keywords is not necessarily requested by publication styles. Nevertheless, if provided (and if free of any pre-defined characterization, like ACM) it should contain besides the areas of research the treatments, dependent variables, and study type.

3.5 Introduction

The purpose of the introduction section is to set the scope of the work and to give the potential reader good reasons for reading the remainder of the publication (motivation). The first section of the introduction needs to set the research into a wider context before introducing the specific problem. The survey of existing guidelines shows variations with regard to the content of this section. In most cases, this section starts with a broader introduction to the research area [25].
We suggest subsections for the *Problem Statement*, the *Research Objectives*, and the description of the *Context* of the research. In particular the description of the context is an essential input for aggregating studies.
With the exception of Wohlin et al. [25], who recommend a special section to describe the problem under study, most of the proposed guidelines include the description of the problem within a more comprehensive section, often labeled "Introduction". Our proposal tries to capitalize on the advantages of these alternatives. On the one hand, we recognize the importance of the problem statement by highlighting the topic by means of an explicit subsection. On the other hand, by including the problem statement in the first section, fast readers will not risk missing this important information.
Following the suggestions of Wohlin et al. [25] and Kitchenham et al. [6], we suggest to explicitly describe the context of the study. While Wohlin et al. describe the context as part of the experiment planning, we decided to encapsulate this topic in a separate subsection.

3.5.1 Problem Statement

The problem statement is important because it supports the readers in comparing their problems with the problem investigated in the reported experiment. In addition, this section helps to judge the relevance of the research. In general, we would expect answers to the questions: What is the problem? Where does it occur? Who has observed it? Why is it important to be solved? Any underlying theory, causal model, or logical model has to be specified in this section. The problem statement should end with a brief description of the solution idea and the (expected) benefits of this solution.

3.5.2 Research Objective

With regard to the research objective, or, as Wohlin et al. called it, the "Definition of the Experiment", the description should follow the goal template of the Goal/Question/Metric (GQM) method [24]:

Analyze <Object(s) of study> for the purpose of <purpose> with respect to their <Quality Focus> from the point of view of the <Perspective> in the context of <context>.

For examples of the use of the goal definition template, see [34] or [25]. The common use of this formalized format would increase the comparability of research objectives, e.g., for the purpose of systematic reviews.

3.5.3 Context

Similar to the CONSORT Statement [1], our *Context* subsection suggests that the setting and locations of a study are described. The author should provide information that will help the readers understand whether the research relates to their specific situations. After having executed the experiment, the context is needed to evaluate external validity, i.e., transferability of results from one context to another. The context consists of all particular factors that might affect the generality and utility of the conclusions, like:

* application type (e.g., real-time system),
* application domain, (e.g., telecommunication),
* type of company (e.g., small and medium sized),
* experience of the participants (e.g., professionals with on average five 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 has to be mentioned if there are specific requirements with regard to the support environment. It is sufficient to describe the context factors informally.

3.6 Related Work

Published guidelines state the importance of clarifying how the work to be reported relates to existing work. Researchers as well as practitioners need to get fast access to related work, because it facilitates drawing a landscape of alternative approaches and relations between different experiments [7]. Appropriate citation is absolutely mandatory.

There is no common consensus on where this section fits best. In contrast to Singer [14], Juristo and Moreno [26], Wohlin et al. [25], and Kitchenham et al. [6], we suggest presenting related work as a special section. This section should consist of the following subsections: description of *Technology* (or tool, method)[1] *under Investigation*, description of *Alternative Solutions, description of related studies,* as well as levels of *Relevance to Practice*.

[1] For the ease of reading, we use technology as an umbrella for technology, method, and tool. Subheadings have to be adopted.

3.6.1 Technology under Investigation

In most cases the technology and the alternatives to be described here will be the treatments (aka. levels or alternatives) in the experiment. For example, if one intends to compare two reading techniques, descriptions would have to be provided with regard to the research objectives. The detail of the required description depends on the availability of earlier publications. After all, the application of the technology should be reproducible. For readers that have no specific background a more general reference, e.g., to a textbook might be helpful. With regard to the content of the description, it is most important that all identifying characteristics are provided; for example, for each level used in the study (e.g., reading techniques in inspection experiments), a description of the pre- and post-conditions for the application is needed. Pre-conditions describe what is necessary to apply the technique, whereas post-conditions describe the (expected) effects.

3.6.2 Alternative Technologies

The relation to alternative approaches in the field will help to arrange this work in a larger context. Shaw [24] recommends that the related work should not only be a simple list of research (i.e., experiments) but an objective description of the main findings relevant to the work at hand. We therefore advise authors to report related work, whether supportive or contradictory.
Especially in the case of an experiment that compares different approaches, it is crucial to objectively describe the alternative approaches. Additionally, other possible alternatives and superseding techniques shall be mentioned.

3.6.3 Related Studies

If available, existing evidence, in the form of earlier studies and especially, experiments, should be described. The relation to other studies (existing evidence) in the field will help to arrange this work in a larger context and supports reuse of this study for replication or systematic review, improving the value of the research and providing a sound basis for this work.
In case the reported study is a replication, the parental study and its findings also have to be described. This will help the reader follow the comparison of the findings.

3.6.4 Relevance to Practice

In addition, if one of the treatments (technologies) has been applied to real software projects or under realistic circumstances, it is important to provide related references.

3.7 Experiment Planning

This section, sometimes referred to as experimental design or protocol, should describe the outcome of the experiment planning phase. It is important because, as Singer stated, this section is the "recipe for the experiment" [14]. In other words, the purpose of this section is to provide the description of the plan or protocol that is afterwards used to perform the experiment and to analyze the results. Therefore, it should provide all information that is necessary to replicate the study, to integrate it into the ESE body of knowledge. In addition, it allows readers to evaluate the internal validity of the study, which is an important selection criterion for systematic review or meta-analysis [3], [6].
We suggest subsections for the formulation of the *Goals*, the *Experimental Units*, the *Experimental Material,* the *Tasks*, the *Hypotheses, Parameters, and Variables*, the *Experiment Design*, the *Procedure*, as well as the subsection *Analysis Procedure*. The order of subsections differs from Harris' proposal, because it was recognized that in case variables arise from the Experimental units of material it is easier to read if they have been introduced before.

3.7.1 Goals

In this subsection the research objective should be refined, e.g., with regard to the facets of the quality focus, if different aspects of the quality focus are of interest ([34] can serve as an example). The purpose is to define the important constructs (e.g., the quality focus) of the experiment's goal.

3.7.2 Experimental Units

In this subsection, information on the sampling strategy (how the sample will be selected), on the population from which the sample is drawn, and on the planned sample size should be provided. In case a statistical power calculation has been used, assumptions, estimates, and calculations have to be provided.

A side comment on the section's title: In empirical SE, many experiments are performed involving human subjects. If this is the case, it is more convenient to talk about participants [14]. Moreover, the commonly used term subjects may be too restrictive. Therefore we propose to use the generic term experimental units, which would cover individual subjects, teams or other experimental items such as tasks and systems.

The instantiation of the sampling strategy and the resulting samples need to be described, including number of participants (per condition), kind of participants (e.g., computer science students). All measures for randomization have to be reported here, especially the random allocation of experimental units to treatments. Random allocation is a necessary requirement for controlled experiments, otherwise it is a quasi-experiment. Furthermore, all characteristics that might have an effect on the results, e.g., experience (with regard to the techniques to be applied (e.g., range of experience in years, together with mean) and educational level. As Singer states, all important characteristics related to the experimental units (subjects) have to be provided [14]. These characteristics can be understood as restrictions to the sample. For instance, if a certain level of experience is required, the sample might be drawn from fourth-term computer science students. A description of the motivation for the subjects to participate is mandatory. For instance, it should be stated whether the participants will be paid (how much) for taking part in the experiment, or whether they will earn educational credits. Additionally, the answers to the following questions are of interest [25]: How were the participants committed? How was consent obtained? How was confidentiality assured? How participation was motivated (induced)?

3.7.3 Experimental Material

This subsection describes the experiment material; that is, the objects used in the experiment. For example, the document used for the application of the reading technique should be presented if terms of its length, complexity, seeded faults (number, type, interactions …) etc.. As stated above, all characteristics that might have an impact on the results should be mentioned here as formally as possible.

3.7.4 Tasks

In this subsection, the tasks to be performed by the experimental units (subjects) have to be described. In case of different treatments and groups, appropriate subsections can be used to ease the reading. Redundancies with regard to the description of the technology in the related work section (c.f., Section 3.6) have to be avoided. The tasks have to be described at a level of detail, so that a replication is possible without the consultation of the authors. If this requires too much space, recommend the use of a technical report or web resources to make the information available.

3.7.5 Hypotheses, Parameters, and Variables

The purpose of this section is to describe the constructs and their operationalization.

For each goal the null hypotheses, denoted H_{0ij}, and its corresponding alternative hypotheses, denoted H_{1ij}, need to be derived, where i corresponds to the goal identifier, and j is a counter in the case that 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 case of ancillary hypotheses a hierarchical system is more appropriate.

Hypotheses need to state the treatments, in particular the control treatments. The description of any control treatment should be sufficient for readers to determine whether the control is realistic. 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.
There are two types of variables that need to be described: dependent variables (aka. response variables) and independent variables (aka. predictor variables). Response variables should be defined and related measures should be justified in terms of their relevance to the goals listed in the section *Research Objectives*. Independent variables are variables that may influence the outcome (the dependent variable); for example, independent variables include treatments, material, and some context factors. In this section, only independent variables are described that are suspected to be causal variables; that is, the variable is either manipulated or controlled through the experiment design. For each independent variable, its corresponding levels (aka. alternatives, treatments) have to be specified in operational form; for example, how the constructs effectiveness or efficiency are measured.
Furthermore, moderating variables like those describing the context of an organization (e.g., size by number of employees, application domain, maturity).
For the definition of measures, we follow Kitchenham et al. [6] who suggest using as many standard measures as possible. Besides approaches to obtain the respective measures such as GQM [24] and a conceptual Entity-Relationship model proposed by Kitchenham et al. [35], no common taxonomy for measures is available yet, although the need has been reported by different authors. A first set of candidate attributes and metrics is presented in Juristo and Moreno [26]. More specialized sets are available for the field of defect reduction [7], [10], [12], [13] and maintenance [36].
Nevertheless, experimenters should be aware of the measurement issue and define their measures carefully. In particular, if a standardized set of metrics is available, authors have to explain which of them are used, not used, or why new ones have been introduced. If existing measures are tailored, the need for the tailoring and the tailored variable has to be described. Based on Juristo and Moreno [26], Wohlin et al. [25], and Kitchenham et al. [6], Table 2 gives a schema for the description of variables.
For subjective measures Kitchenham et al. suggest that a measure of inter-rater agreements is presented, such as the kappa statistics or the intra-class correlation coefficient for continuous measure [6].

3.7.6 Experiment Design

The hypothesis and the variables influence the choice of the experimental design. In the *Experiment Design* subsection the selection of the specific design has to be described. This selection is supported

Table 2. Schema for the description of variables

Name of the variable	Type of the variable (independent, dependent, moderating)	Abbreviation	Class (product, process, resource, method) [25],[26]	Entity (instance of the class) [26]	Type of attribute (internal, external) [25],[26]	Scale type (nominal, ordinal …) [35]	Unit [35]	Range or, for nominal and restricted ordinal scales, the definition of each scale point. [35]	Counting rule in the context of the entity [35]

by further selection criteria (e.g., randomization, blocking, and balancing). Kitchenham et al. [6] propose selecting a design that has been fully analyzed in the literature. If such a design is not appropriate, authors are recommended to consult a statistician. In this case, more details about the background of the design are needed. Descriptions of design types can be obtained from Wohlin et al. [25] and Juristo and Moreno [26].
The design and corresponding hypotheses need to be simple and appropriate for the goal. Wohlin et al. stress that "it is important to try to use a simple design and try to make the best possible use of the available subjects". This point is also referred to by Kitchenham et al. [6], who point out that in many SE experiments, the selected design is complex, and the analysis method is inappropriate for coping with it.

Moreover, authors should describe how the units (i.e., subjects or participants) and material (i.e., objects) are assigned to levels (treatments) in an unbiased manner [6].
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). In case the experiment is a replication, the adjustments and their rationales need to be discussed. Additionally, training provided to the units has to be described, if applicable. Any kind of threat mitigation should also be addressed. For example, a typical strategy to reduce learning effects is to apply a design where subjects apply techniques in different order.

3.7.7 Procedure

In this subsection, it has to be described what precisely will happen to the participants from the moment they arrive to the moment they leave. To be consequent with the description of the planning phase, this is in contrast to Harris who asks for a description of what really happened during the experiment [27]. The description should include the setting (e.g., Where was the experiment conducted? Was it in a separate room for a limited period of time or was it done at home) and the schedule of the experiment as well as the timing for each run of the experiment. Furthermore, details of the data collection method have to be described, including measurement instruments; that is, means for collecting data (e.g., questionnaires, data collection tools). It is important to describe where (e.g., in which phase of the process) the data will be collected, by whom, and with what kind of support (e.g., tool). This is also in accordance with Kitchenham et al. [6], who state that the data collection process describes the "who?", the "when?", and the "how?" of any data collection activity. Activities involving marking the outcomes of the experimental tasks (e.g., mark "true" defects in defect lists) and training provided for the markers have to be described.

3.7.8 Analysis Procedure

This section shall discuss what kind of statistical tests will be used to analyze the hypotheses and why. Since there are, for example, several variants of ANOVA for different designs it is required to report the type of ANOVA precisely [27]. If different goals are investigated, information for each goal needs to be provided separately. If any additional influences are expected, their analysis needs to be described, too (e.g., see Ciolkowski et al. [34]). As a side comment authors are advised that they should focus experiments to specific hypotheses and perform the minimum not the maximum number of tests. According to Kitchenham [6], testing subsets of the data and fishing for results are bad practice. If there a limitations with regard to the numbers of pages, this section might be combined with the analysis section.

3.8 Execution

According to Singer [14], the purpose of this section is to describe "each step in the production of the research". From our perspective, execution describes how the experimental plan (design) was enacted. The general schedule of the experiment needs to be described as well as how much time the participants were given to run the experiment As stated earlier, in case of limited number of pages this section can be integrated with the experiment planning section. Then, this section can be omitted and a general statement confirming the process conformance in the analysis section (c.f., Section 3.9) will be sufficient.
The most important point is to describe whether any deviations from the plan (aka. protocol) occurred and how they were treated. This counts for all aspects of the experiment plan. Besides the who and the when, the specific instantiations of the sampling, randomization, instrumentation, execution, and data collection have to be described.
We suggest structuring the *Execution* section into the following subsections: *Preparation* and *Deviations.*

3.8.1 Preparation
The purpose of this subsection is to describe the preparation of the experiment. For example, it ought to be described how the experimental groups were formed, how the randomization was performed, what kind of training, if any, was provided to the participants, and how long the training took.

3.8.2 Deviations
The purpose of this subsection is to discuss any deviations occurred during the execution of the experiment, i.e., regarding the planned instrumentation and the collection process and how they were solved. In addition, information about subjects who drop out from the study should be presented, e.g., that five subjects did not attend the final (as, recommended for by Kitchenham et al. [6]).

3.9 Analysis
According to Singer [14], the *Analysis* section summarizes the data collected and the treatment of the data. The most important points for this section are: (1) It is not allowed to interpret the results [14] and (2) data should be analyzed in accordance with the design [6]. If multiple goals were investigated, separate analysis subsections and an overlap 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. If any deviations from the plan occur, they have to be discussed here, e.g., in case no statistically significant results were found, it is necessary to describe what was done to cope with this circumstance.
We suggest structuring the *Analysis* section into the following subsections: *Descriptive Statistics*, *Data Set Preparation*, and *Hypothesis Testing*. If appropriate a sensitivity analysis should be reported in the hypothesis testing section.

3.9.1 Descriptive Statistics
The purpose of this subsection is to present the collected data with the help of appropriate descriptive statistics, including number of observations, measures for central tendency, and dispersion. 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 to provide raw data in the appendices or to describe where the data can be acquired from, e.g., from a web site.

3.9.2 Data Set Preparation
In this subsection the preparation of the data set as a consequence of the descriptive statistics should be discussed. This includes data transformation, outlier identification and their potential removal, and handling of missing values as well as the discussion of drop outs, e.g., the data from the drop outs where completely removed from the data set.

3.9.3 Hypothesis Testing
In this subsection, it has to be described how the data was evaluated and how the analysis model was validated. Special emphasis should be placed on constraints that would hinder the application of a planned analysis method (e.g., normality, independence, and residuals). Any resulting deviations with regard to the hypothesis test from the original plan (e.g., a different test was used because of data set constraints) should be described. Moreover, it has to be described which methods were used to determine statistical significance.
To understand the interpretation and conclusion based on the analysis, it is important to present inferential statistics. Harris provides examples about what has to be reported for different kind of statistical tests [27]. Singer [14] recommends that “inferential statistics are reported with the value of the test, the probability level, the degrees of freedom, the direction of effect”, and the power of the test. More precisely, p-value, alpha-value, and confidence interval for each finding have to be presented. Statistics of effect size shall also be reported to facilitate meta-analysis or comparison of results across experiments. Kitchenham et al. [6] present a checklist for reporting inferential results.

For each hypothesis, quantitative results should be presented. If a null hypothesis is rejected, it has to be described on which significance level. Furthermore, if several different aspects like individual performance, group performance were investigated, a separate subsection for each analysis shall be used.

3.10 Discussion

The purpose of the discussion section is to interpret the findings from the analysis presented in the previous section. Authors have to make it clear how they arrive at their interpretation given the specific results. 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 [27] suggests starting this section (Discussion) with a description of what has been found and how well the data fit the predictions. We suggest structuring the *Discussion* section into the following subsections: *Evaluation of Results and Implications*, *Threats to Validity*, *Inferences*, and *Lessons Learned*.

3.10.1 Evaluation of Results and Implications

In this subsection, the results should be explained. In case it was not possible to reject the null hypotheses, assumptions about the reasons why this happened should be given. Also, any other unexpected result should be described in this subsection. It is pointed out by several authors, that it is important to distinguish between statistical significance and practical importance [6] or meaningfulness [27].

In case of a rejected null hypothesis it has to be assessed that this came through the variation of the independent variable. In case of not being able to reject the null hypothesis the causal relationship between the independent and dependent variables has to be assessed. A more detailed discussion about what has to be incorporated into the evaluation of the results section is given in [27].

The second part of this section shall be dedicated to the discussion of the theoretical implications of the findings for the area.

The author has to refer to the *Introduction* and the *Related Work* section. This is the place where the relation of the results to earlier research (experiments) has to be provided. The contribution of the study should be discussed here; that is to consider how the results contribute to the underlying theory as described in the introduction section.

We point out that it is important to (1) ensure that conclusions follow from the results [6], (2) differentiate between the results of the analysis and the conclusions drawn by the authors [6], and (3) not indulge in fanciful speculation, conjectures have to be made with caution and kept brief [27].

3.10.2 Threats to Validity

All threats that might have an impact on the validity of the results as such have to be discussed in this subsection. This includes at least (1) *threats to construct validity* (the degree to which inferences can legitimately be made from the operationalizations to the theoretical constructs on which those operationalizations were based), (2) *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 (3) *threats to external validity* (e.g., participants, materials). If applicable, (4) *conclusion validity* (e.g., appropriateness of statistical tests) should also be addressed. Subsections might be used for each threat that has to be discussed by using the kind of threat as heading. A comprehensive classification of threats to validity is given, e.g., in Wohlin et al. [25]. Following the arguments presented by Kitchenham et al. [6], it is not enough to only mention that a threat exists; it also ought to be discussed what the implications are. Furthermore, other issues, like personal vested interests or ethical issues regarding the selection of experimental units (in particular experimenter-subject dependencies) shall be discussed

For internal validity, for example, it ought to be described whether any particular steps were taken to increase the reliability of the measurement instruments. This refers, among other things, to the completeness, appropriateness (correctness), and consistency of the data collection. If any actions were planned, they ought to be described, together with their effects. Actions that have been performed in advance could be, for instance, specific training, double checks, and automatic measurements. A description of the validity of the materials used during the study and the conformance of the participants, e.g., how it is ensured that the participants will follow the guidelines [26] (i.e., process conformance), is necessary.

3.10.3 Inferences

The purpose of this subsection is to describe inferences drawn from the data to more general conditions; that is, to generalize findings (define the scope of validity). This has to be done carefully, based on the findings by incorporating the limitations. It has to be ensured that all claims can be supported by the results. This care includes the definition of the population to which inferential statistics and predictive models apply. This subsection is also the place to describe how and where the results can be used (generalization). For technologies not in use, scale-up issues have to be discussed.

3.10.4 Lessons Learned

In this optional subsection, it ought to be described which experience was collected during the course of the experiment, i.e., during design, execution analysis. The purpose is to describe what went well and what did not. If the reasons for interesting observations are known, they can be described in this subsection, too.

3.11 Conclusions and Future Work

This section shall describe, based on the results and their discussion as presented above, the following aspects as detailed in respective subsections: *Summary, Impact*, and *Future Work*.

3.11.1 Summary

The purpose of this section is to provide a concise summary of the research and its results as presented in the former sections.

3.11.2 Impact

To enable readers to get the most important findings with regard to the practical impact in one place, we emphasize a description of the impact on cost, time, and quality and summary of the limitations.
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 other steps of the development process?
Impact on Quality: Is there any impact on the quality of the product and the products of other steps?
Besides the description of the impact we ask for a discussion of the approach's level of maturity, when the investments will pay back, and consequences arising from the implementation. (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 have to be described, i.e., circumstances under which the approach presumably will not yield the expected benefits or shall not be employed. Furthermore, any risks or side-effects associated with the implementation or application of the approach shall be reported like in a package insert accompanying drugs.

3.11.3 Future Work

In this subsection, it has to be described what other research (i.e., experiments) could be run to further investigate the results yielded or evolve the body of knowledge.

3.12 Acknowledgements
In this section sponsors, participants, and (research) contributors who do not fulfill 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 what conditions the material and preferably the raw data, too, will be made available to other researchers (i.e., technical report, web resource).

4. Conclusion and Future Work
The contribution of this paper is a checklist for reporting experiments in software engineering that 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 checklist provide detailed guidance on how to fill in the various sections and subsections of this template for a specific type of empirical studies, i.e., controlled experiments and quasi-experiments. In some places, for instance for the definition of variables, we suggest a prescriptive formalization schema.
This checklist aims at structured, complete and comprehensive documentation of controlled experiments. We are aware of the fact that in some cases due to page limitations (e.g., conference paper) it might be not possible to provide all the proposed information. Therefore, we envision to tailor this checklist towards different purposes, namely for different lengths of reports (e.g., by combining some sections) and for different audiences (e.g., by providing a checklist for reporting results from controlled experiment to practitioners).
Our proposal has been evaluated, e.g., through a peer review thought members of the program committee of the International Symposium on Empirical Software Engineering and by means of a perspective-based inspection [19]. To our knowledge, although many researchers have promised to use the checklists and report feedback, the checklists have not yet been evaluated by applying it to a significant number of experiments to check its usability. In order to assess the benefits and challenges of the proposed guidelines, it is necessary to use them in two ways: (1) to describe new experiments and (2) to rewrite already published experiments. The first approach, preferably performed by different research groups to reduce expectation bias, leads to feedback with regard to the applicability that will be based on the experience of the very authors. The second approach can be used to compare the availability and accessibility of information between the two descriptions. Further approaches for evaluating guidelines in general are proposed by Kitchenham et al. [19]. The prerequisite is the general availability of information with regard to the specific experiment. We are aware that this guideline is not a static instrument but will, even if it is a standard, need continuous evaluation with regard to changing requirements.
The experience of the last seven years, since the first publication of a reporting guideline for empirical SE research by Singer in 1999 [14], leads us to conclude that significant effort needs to be invested to make sure that guidelines are widely accepted. This is also what other communities have already learned [15]. We propose adopting some of their measures to enact reporting guidelines within the SE community. For example, we believe the SE community needs an organization that is

able to achieve sufficient consensus on the guidelines and that is able to establish the guidelines in review boards for journals, conferences, etc.
At this point in time, the discussion with regard to reporting guidelines has just started. The involvement of different stakeholders is crucial for success. To address this aim, we have set up an initial working group, consisting of eight researchers from five countries, who have committed themselves to the task of defining and disseminating guidelines. The first step on that path was to identify which types of guidelines are needed, and to define the goal for each type of guideline. As the main objective we have identified the further use of empirical reports, namely the aggregation of empirical results from single studies. This objective is supported by different authors who have tried to aggregate single findings into more generic knowledge, e.g., [4], [10], [12], [13].
The working group collected a set of existing guidelines, including those from other disciplines. The authors have committed to refining guidelines for controlled experiments.
One important issue related to defining guidelines is to evaluate and ensure the quality of the proposed guidelines as well as to further evolve them. This will be done by reporting studies, preferably performed by different research groups to reduce bias (to overcome expectation biases, it is important in this phase that the guidelines are used by volunteering authors who were not involved in the definition), following the guidelines and trials to perform a systematic review [3] (as a specific form of aggregation). The systematic reviews should be done by groups of experts in the specific field. We will then qualitatively compare the ease of extracting relevant information from the report following the guidelines with the attempts we made before with study reports that do not follow the guideline. Further needs that might not be foreseen today may require evolution of the guidelines.
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 that a researcher can benefit from applying them. For example, one benefit could be that the integration into the Body of Knowledge will be easier if studies are reported using the guidelines. 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.

5. Acknowledgements

We would like to thank Claes Wohlin who gave valuable insights and comments, Barbara Kitchenham and her team at NICTA for their valuable feedback, which helped to improve the guidelines, and many others for fruitful discussions. Furthermore, we are grateful to Sonnhild Namingha from our Institute for reviewing a previous version of this paper.

6. References

[1] Glass, R.L.: Matching Methodology to Problem Domain; In Column Practical Programmer in Communications of the ACM /Vol. 47, No. 5, May 2004, pp. 19-21
[2] Kitchenham, B.A.; Dybå, T.; Jørgensen, M.; Evidence-based Software Engineering; In Proc. of 26th Intern. Conf. on Software Engineering (ICSE'04); May 2004; Edinburgh, Scotland, United Kingdom, 2004, pp. 273-281
[3] Kitchenham, B.A.; 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 July, 2004
[4] Jedlitschka, A.; Ciolkowski, M.: Towards Evidence in Software Engineering; In Proc. of ACM/IEEE Intern. Symposium on Software Engineering 2004 (ISESE2004), Redondo Beach, California, August 2004, IEEE CS, 2004, pp. 261-270

[5] Altman, D.G.; Schulz, K.F.; Moher, D.; Egger, M.; Davidoff, F.; Elbourne, D.; Gøtzsche, P.C. and Lang, T. for the CONSORT Group; The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration, in Annals of Internal Medicine, Volume 134, Nr 8, April 2001, pp. 663 – 694

[6] Kitchenham, B.A.; Pfleeger, S.L.; Pickard, L.M.; Jones, P.W.; Hoaglin, D.C.; El Emam, K.; Rosenberg, J.: Preliminary guidelines for empirical research in software engineering; IEEE Transactions on Software Engineering, Vol. 28, No. 8 , Aug 2002, pp. 721 -734.

[7] Jedlitschka, A.; Ciolkowski, M.; Towards a Comprehensive Summarization of Empirical Studies in Defect Reduction; In Proc. of ISESE 2004 Vol.II: Posters and Fast Abstract Sessions, Redondo Beach, California, August 2004, pp. 5-6

[8] Lott, C.M.; Rombach, H.D.; Repeatable software engineering experiments for comparing defect- detection techniques; Empirical Software Engineering Journal, 1996, Vol. 3.1, pp. 241-277

[9] Pickard, L.M.; Kitchenham, B.A.; Jones, P.W.: Combining empirical results in software engineering; Information and Software Technology, 40(14): 1998, pp. 811-821

[10] Runeson, P.; Thelin, T.: Prospects and Limitations for Cross-Study Analyses – A Study on an Experiment Series. In Jedlitschka, A.; Ciolkowski, M. (eds): The Future of Empirical Studies in Software Engineering, Proc. of 2nd Int. Workshop on Empirical Software Engineering, WSESE 2003, Roman Castles, Italy, Sept. 2003, Fraunhofer IRB Verlag, 2004. pp. 141-150.

[11] Shull, F., Carver, J., Travassos, G. H., Maldonado, J. C., Conradi, R., and Basili, V. R.; Replicated Studies: Building a Body of Knowledge about Software Reading Techniques; in [37], pp. 39-84

[12] Vegas, S.; Juristo, N.; Basili, V.: 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 2003, pp. 199-230

[13] Wohlin, C.; Petersson, H.; Aurum, A.: Combining Data from Reading Experiments in Software Inspections; In [37], pp. 85-132

[14] Singer, J.: Association (APA) Style Guidelines to Report Experimental Results; In Proc. of Workshop on Empirical Studies in Software Maintenance, Oxford, England. September 1999. pp. 71-75. (dec.bmth.ac.uk/ESERG/WESS99/singer.ps)

[15] Zelkowitz, M.V.; Wallace, D.R.; Binkley, D.W.; Experimental Validation of New Software Technology; In [37], pp. 229 – 263

[16] Jedlitschka, A.; Pfahl, D.; Reporting Guidelines for Controlled Experiments in Software Engineering. IESE-Report IESE-035.5/E, 2005

[17] Jedlitschka, A.; Minutes from Third International Workshop on Empirical Software Engineering "Guidelines for Empirical Work in Software Engineering". IESE-Report 052.05/E, Oulu, June 2005

[18] Jedlitschka, A.; Pfahl, D.; Reporting Guidelines for Controlled Experiments in Software Engineering; In Proc. of ACM/IEEE Intern. Symposium on Software Engineering 2005 (ISESE2005), Noosa Heads, Australia, Nov 2005, IEEE CS, 2005, pp. 95-104

[19] Kitchenham, B.; Al-Khilidar, H.; Ali Babar, M.; Berry, M.; Cox, C.; Keung, J.; Kurniawati, F.; Staples, M.; Zhang, H.; Zhu, L.; Evaluating Guidelines for Empirical Software Engineering Studies; In Proc. of ACM/IEEE Intern. Symposium on Software Engineering 2006 (ISESE2006), Rio de Janerio, Brazil, Sep 2006, IEEE CS,

[20] Moher, D.; Schulz, K.F.; Altman, D.; for the CONSORT Group; 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, April 18, 2001; pp. 1987-1991

[21] Harris, P.; Designing and Reporting Experiments in Psychology. 2nd Edition, Open University Press, 2002

[22] Shiffman, R.N.; Shekelle, P.; Overhage, J.M.; Slutsky, J.; Grimshaw, J.; and Deshpande, A.M.; Standardized Reporting of Clinical Practice Guidelines: A Proposal from the Conference on Guideline Standardization; Annals of Internal Medicine; Volume 139 Issue 6; September 2003; pp. 493-498

[23] American Psychological Association. 2001. Publication Manual of the American Psychological Association, (5th ed.). Washington, DC: American Psychological Association.

[24] Shaw, M.: Writing Good Software Engineering Research Papers - Minitutorial; In Proc. of the 25th Intern. Conf. on Software Engineering (ICSE'03), Portland, Oregon, IEEE Computer Society, 2003, pp. 726-736.

[25] Wohlin, C.; Runeson, P.; Höst, M.; Ohlsson, M.C.; Regnell, B.; Wesslén, A.; Experimentation in Software Engineering - An Introduction; Kluwer Academic Publishers, 2000

[26] Juristo, N. and Moreno, A.; Basics of Software Engineering Experimentation; Kluwer Academic Publishers, 2001

[27] Harris, P.; Designing and Reporting Experiments in Psychology; Open University Press, Berkshire, UK, second edition 2000

[28] Hayward, R.S.A.; Wilson, M.C.; Tunis, S.R.; Bass, E.B.; Rubin, H.R.; Haynes, R.B.; More Informative Abstracts of Articles Describing Clinical Practice Guidelines; In Annals of Internal Medicine Vol. 118 Issue 9; May 1993; pp. 731-737

[29] The Editors; Addressing the Limitations of Structured Abstracts (Editorial); In Annals of Internal Medicine Vol. 140, No.6 March 2004
[30] Hartley, J; Improving the Clarity of Journal Abstracts in Psychology: The Case for Structure; In Science Communication, Vol 24, 3, 2003,pp.366-379.
[31] Hartley, J; Current findings from research on structured abstracts; In Journal of the Medical Library Association, 92, 3, 2004,pp. 368-371.
[32] Basili, V.R., Caldiera, G., Rombach, H.D.: Goal Question Metric Paradigm; in: Marciniak J.J. (ed.), Encyclopedia of Software Engineering, Vol.1, John Wiley & Sons, 2001, pp.528–532.
[33] Bayley, L.; Eldredge, J.; 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, Spring 2003, 4 pages
[34] Ciolkowski, M.; Differding, C.; Laitenberger, O.; Münch, J.; Empirical Investigation of Perspective-based Reading: A Replicated Experiment; Fraunhofer Institute for Experimental Software Engineering, Germany, 1997, ISERN-97-13
[35] Kitchenham, B.A.; Hughes, R.T.; Linkman, S.G.; Modeling Software Measurement; IEEE Transactions on Software Engineering, Vol.27, No.9, September 2001, pp. 788-804
[36] Kitchenham, B.; Travassos, G.; von Mayrhauser, A.; Niessink, F.; Schneidewind, N.F.; Singer, J.; Takada, S.; Vehvilainen, R.; Yang, H.; "Towards an Ontology of Software Maintenance," J. Software Maintenance: Research & Practice, 11:. 1999, pp. 365-389
[37] Juristo, N. and Moreno, A. (eds.); Lecture Notes on Empirical Software Engineering, Ed. River Edge, NJ, USA: World Scientific Publishing, October 2003

Appendix: Structure for reporting experiments in software engineering

Section	Sub-Section	Scope
Title		<title> + "- A controlled experiment"; Is it informative and does it include the treatments and the dependent variables?
Authorship		Does it include contact information?
Structured Abstract	Background	Why is this research important?
	Objective	What is the question addressed with this research?
	Methods	What is the statistical context and methods applied?
	Results	What are the main findings? Practical implications?
	Limitations	What are the weaknesses of this research?
	Conclusions	What is the conclusion?
Keywords		areas of research the treatments, dependent variables, and study type
Introduction	Problem Statement	What is the problem? Where does it occur? Who has observed it? Why is it important to be solved?
	Research Objective	Analyze <Object(s) of study> for the purpose of <purpose> with respect to their <Quality Focus> from the point of view of the <Perspective> in the context of <context>
	Context	What information is necessary to understand whether the research relates to a specific situation (environment)?
Related Work	Technology under Investigation	What is necessary for a reader to know about the technology to reproduce its application?
	Alternative Technologies	How this research relates to alternative technologies?
	Related Studies	How this research relates to existing research (studies)?
	Relevance to Practice	How does it relate to state of the practice?
Experiment Planning	Goals	Formalization of goals, define the important constructs (e.g., the quality focus) of the experiment's goal
	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.
	Experimental Material	Which objects are selected and why?
	Tasks	Which tasks have to be performed by the subjects?
	Hypotheses, Parameters, and Variables	What are the constructs and their operationalization?
	Design	What type of experimental design has been chosen?
	Procedure	How will the experiment (i.e. data collection) be performed? What instruments, materials, tools will be used and how?
	Analysis Procedure	How will the data be analyzed?
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?
Analysis	Descriptive Statistics	What are the results from descriptive statistics?
	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?
Discussion	Evaluation of Results and Implications	Explain the results and its relation of the results to earlier research, especially those mentioned in the *Related Work* section
	Threats to Validity	How is validity of the experimental results assured? How was the data actually validated? 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
	Lessons Learned	which experience was collected during the course of the experiment
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.
	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
Acknowledgements		sponsors, participants, and contributors who do not fulfill the requirements for authorship should be mentioned
References		all cited literature has to be presented in the format requested by the publisher
Appendices		material, raw data, and detailed analyses, which might be helpful for others to build upon the reported work should be provided