Reporting Guidelines for Controlled Experiments in Software Engineering

Andreas Jedlitschka
Fraunhofer Institute for Experimental Software Engineering
Fraunhofer-Platz 1
67663 Kaiserslautern, Germany
jedl@iese.fraunhofer.de

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

Abstract

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. Reporting guidelines are 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. The objective of this paper is to survey the most prominent published proposals for reporting guidelines, and to derive a unified standard that which can serve as a starting point for further discussion. We provide detailed guidance on the expected content of the sections and subsections for reporting a specific type of empirical studies, i.e., controlled experiments. Before the guidelines can be evaluated, feedback from the research community is required. For this purpose, we propose to adapt guideline development processes from other disciplines.

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 demands:

- 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 parts of these issues by providing a framework for goal-oriented research, leading to a common body of knowledge (BoK) and, based on that, comprehensive problem-oriented decision support regarding SE technology selection.

One major problem for integrating study results into a common BoK is the heterogeneity of study reporting [3]: (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.

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 [4].

As already identified by Kitchenham et al. [5], 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., [3], [6], [7], [8], [9], [10], [11], [12]. At the same time, several reporting guidelines have been proposed, e.g., [13], [5]. 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. [14]).

0-7803-9508-5/05/$20.00

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 standard, which can serve as a starting point for further discussion.

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 [4], [15], clinical practice guidelines [16], and empirical results from psychological research [17].

In the field of software engineering (SE) research, in 1999, Singer [13] described how to use the "American Psychological Association (APA) Styleguide" [17] for publishing experimental results in SE. In 2001, Kitchenham et al. [5] 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 [18] 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. [19] and Juristo et al. [20], address the issue of reporting guidelines. Wohlin et al. suggest an outline for reporting the results of empirical work. Juristo et al. 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 unifying and enhanced reporting guideline. The second row of the

Table 1. Characterization of different Guidelines for empirical SE

	Singer [13]	Wohlin et al. [19]	Kitchenham et al. [5]	Juristo et al. [20]	Kitchenham [21]	New Proposal
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	Motivation
		Problem Statement				
		Experiment Planning	Experimental Context			
	Introduction	Problem Statement	Experimental Context	Goal Definition	Background	Related Work
	Method	Experiment Planning	Experimental Design	Design	Review Questions	Design
					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	Interpretation
	Discussion	Discussion and Conclusion	*	Experimental Analysis	Conclusion	Discussion and Conclusions
		Discussion and Conclusion	*	Experimental Analysis		Future Work
					Acknowledgments	Acknowledgements
					Conflict of Interest	
	References	References	*	*	References	References
	Appendices	Appendix			Appendices	Appendices

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. Proposed Guideline for Controlled Experiments

Our work started with the collection and integration of existing guidelines. As indicated in Table 1, the resulting reporting guideline comprises the following elements: *Structured Abstract, Motivation, Related Work, Design, Execution, Analysis, Interpretation, Discussion, Conclusion and Future Work, Acknowledgements, References,* and *Appendices*. The structuring elements are discussed in detail in the following subsections.

3.1 Structured Abstract

The need for an 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 [21]. The exact form 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 [18]. In other disciplines, e.g., medicine and psychology, a special form of the abstract, the so-called structured abstract [27], 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" [16].

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 used 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 principal limitations of the research. Example: "Generalization of results is limited due to the fact that undergraduate students participated in the study".

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

The inclusion of the element *Limitations* follows a suggestion made in [23], 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 [23].

It is important to use only a few sentences for each structuring element of the abstract. Hartley [24] found that the number of words will increase by about 30%. 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 [25], [21].

3.2 Motivation

The purpose of the motivation section is to set the scope of the work and to give the potential reader good reasons for reading the remainder of the publication. 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 [19].

We suggest subsections for the *Problem Statement*, the *Research Objectives*, and the description of the *Context* of the research.

With the exception of Wohlin et al. [19], who demand 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 motivation section, fast readers will not risk missing this important information.

Following the suggestions of Wohlin et al. [19] and Kitchenham et al. [5], we suggest to explicitly describe the context of the study. While Wohlin et al. describe the context as part of the experimental planning, we decided to encapsulate this topic in a separate subsection.

3.2.1 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? The problem statement should end with a brief description of the solution idea and the (expected) benefits of this solution.

3.2.2 Research Objectives

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 [18]:

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 [28] or [19]. The common use of this formalized format would increase the comparability of research objectives, e.g., for the purpose of systematic reviews.

3.2.3 Context

Similar to the CONSORT Statement [1], our *Context* subsection requests that the setting and locations of a study have to be 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 (e.g., application domain, type of company, experience of the participants, time constraints, process, tools, and size of project) that might affect the generality and utility of the conclusions.

It is sufficient to describe the context factors informally within the *Context* subsection. A precise (formal) definition will be given later on in the section *Experimental Design* (cf. Section 3.4.).

3.3 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 [6].

There is no common consensus on where this section fits best. In contrast to Singer [13], Juristo et al. [20], Wohlin et al. [19], and Kitchenham et al. [5], we suggest presenting related work as a special section. This section should consist of: *Description of the Investigated Technology* (or tool, method), *Description of Alternative Solutions*, and *Related Experiments*.

In most cases the technology and the alternatives to be described here will be the levels 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 description depends on the availability of earlier publications. 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. Shaw [18] demands that the related work should not only be a simple list of experiments but an objective description of the main findings relevant to the work at hand.

Especially in the case of an experiment that compares different approaches, it is crucial to objectively describe the alternative approaches. Additionally, other alternatives might be mentioned. If available, existing evidence, in the form of earlier experiments, should be described. The relation to alternative approaches and other experiments (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. Appropriate citation is absolutely mandatory.

3.4 Experimental Design

This section should describe the outcome of the experiment planning phase. It is important because, as Singer stated, this section is the "receipt for the experiment" [13]. It should provide all information that is necessary to replicate the study or to integrate it in the BoK. 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 [21], [5].

We suggest subsections for the formulation of the *Goals, Hypotheses, Parameters, and Variables*, the *Design***,** the *Subjects*, the *Objects*, the *Instrumentation*, the *Data Collection Procedure***,** the *Analysis Procedure*, as well as the subsection *Evaluation of the Validity*. The format of the *Experimental Design* section is inspired by the structure of the "Experiment Planning" phase suggested by Wohlin et al. [19].

3.4.1 Goals, Hypotheses, Parameters, and Variables

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 ([28] can serve as an example). Regarding the naming of the types of variables, we follow Juristo et al. [20].

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 context of an experiment needs to be described by listing parameters that represent characteristics that are invariable throughout the conduct of experiment and do not influence the results of the experiment. A precise description of the context via measurable parameters is essential, because the results yielded by the experiment will be true locally for the conditions reflected in the parameters.

There are two types of variables that need to be described: response variables (aka. dependent variables) and factors (aka. independent variables or 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*. For each factor, its corresponding levels (aka. alternatives, treatments) have to be specified in measurable form.

For the definition of measures, we follow Kitchenham et al. [5] who suggest using as many standard measures as possible. Besides approaches to obtain the respective measures such as GQM [18] and a conceptual Entity-Relationship model proposed by Kitchenham et al. [29], 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 et al. [20]. More specialized sets are available for the field of defect reduction [6], [9], [11], [12] and maintenance [30].

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 have to be described. Based on Juristo et al. [20], Wohlin et al. [19], and Kitchenham et al. [5], Table 2 gives a schema for the description of variables.

For subjective measures Kitchenham et al. request that a measure of inter-rater agreements is presented, such as the kappa statistics or the intra-class correlation coefficient for continuous measure [5].

3.4.2 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 by further selection criteria (e.g., randomization, blocking, and balancing). Kitchenham et al. [5] 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. [19] and Juristo et al. [20].

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

Table 2. Schema for the description of variables

Name of the variable	Abbreviation	Class (product, process, resource, method) [19],[20]	Entity (instance of the class) [20]	Type of attribute (internal, external) [19],[20]	Scale type (nominal, ordinal …) [29]	Unit [29]	Range or, for nominal and restricted ordinal scales, the definition of each scale point. [29]	Counting rule in the context of the entity [29]

available subjects". This point is also referred to by Kitchenham et al. [5], 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 subjects and objects are assigned to levels (treatments) in an unbiased manner [5]. If any kind of blinding has been used, the details need to be provided; this applies to the execution and the analysis. In case the experiment is a replication, the adjustments and their rationales need to be discussed.

3.4.3 Subjects

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. As Singer states, all important subject-related characteristics have to be provided [13]. 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 for taking part in the experiment, or whether they will earn educational credits.

In empirical SE, many experiments are performed involving human subjects. If this is the case, it is more convenient to talk about participants [13].

3.4.4 Objects

In this subsection, the objects used in the experiment, for example the document used for the application of the reading technique (length, complexity …), and faults (number, type, interactions …) should be presented. As stated above, all characteristics that might have an impact on the results should be mentioned here as formally as possible.

3.4.5 Instrumentation

In this subsection, information about the instrumentation that might have an impact on the results should be provided. There are two types of instruments: guidelines and measurement instruments (e.g., questionnaires, data collection tools). It is also important to describe which kind of training, if any, the participants will get.

3.4.6 Data Collection Procedure

In this subsection, the schedule of the experiment as well as the timing for each run of the experiment has to be provided. Furthermore, details of the collection method have to be described. Examples vary from manual collection by the participants to automatic collection by 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. [5], who state that the data collection process describes the "who?", the "when?", and the "how?" of any data collection activity.

3.4.7 Analysis Procedure

Since the analysis depends on the design, the mathematical analysis model should be presented in this subsection. 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. [28]).

3.4.8 Validity Evaluation

In this subsection it has to be described whether any particular steps will be taken to increase the reliability of the measurements, meaning that the data is reasonable and that complete and appropriate (correct) collection of data is ensured [5], [19]. Actions that could be taken in advance could be, for instance, specific training, double checks, and automatic measurements; if any actions are planned, they have to be described, too.

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 [20], is necessary. In addition, actions established to improve the reliability and validity of data collection methods or tools have to be described.

3.5 Execution

According to Singer [13], 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. So, besides the who and the when, the specific instantiations of the sampling, randomization, instrumentation, apparatus, execution, data collection, and validation have to be described. The most important point is to describe whether any deviations from the plan occurred and how they were treated.

We suggest structuring the *Execution* section into the following subsections: *Sample*, *Preparation*, *Data Collection Performed*, and *Validity Procedure*.

3.5.1 Sample

In this subsection, the instantiation of the sampling strategy and the resulting sample needs to be described, including number of participants, kind of participants (e.g., computer science students), and all characteristics that might have an effect on the results, e.g., experience (with regard to the techniques to be applied) and educational level. Additionally, the answers to the

following questions are of interest [19]: How were the participants committed? How was consent obtained? How was confidentiality assured? How was participation motivated (induced)?

3.5.2 Preparation

In this subsection, it has 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.

3.5.3 Data Collection Performed

The purpose of this subsection is to describe how the collection process was followed and, if any deviations occurred, how they were solved. The general schedule of the experiment needs to be described as well as how much time the participants were given to run the experiment. Kitchenham et al. [5] demand information about subjects who drop out from the study.

3.5.4 Validity Procedure

The purpose of this subsection is to describe how the validity process was followed and, if any deviations occurred, how they were solved. A description of what kind of actions were taken and what their effect was is necessary.

3.6 Analysis

According to Singer [13], 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 [13] and (2) data should be analyzed in accordance with the design [5]. 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 collection. 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 Reduction*, and *Hypothesis Testing*.

3.6.1 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 (cf. Section 3.11).

3.6.2 Data Set Reduction

In this subsection the reduction of the data set as a consequence of the descriptive statistics should be discussed, i.e., the removal of outliers.

3.6.3 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. Singer [13] demands 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, the p-value, alpha-value, and confidence interval for each finding has to be presented. For each hypothesis, quantitative results should be presented. If a null hypothesis is rejected, it has to be described on which significance level. Kitchenham et al. [5] present a checklist for reporting inferential results.

3.7 Interpretation

The purpose of this section is to interpret the findings from the analysis 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. We suggest structuring the *Interpretation* section into the following subsections: *Evaluation of Results and Implications*, *Limitations of the Study*, *Inferences*, and *Lessons Learned*.

3.7.1 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. Kitchenham et al. [5] point out that it is important to distinguish between statistical

significance and practical importance. The theoretical implications of the findings should be described.

3.7.2 Limitations of the Study

As already mentioned in the *Experimental Design* section, all 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 in this subsection. According to Kitchenham et al. [5], it is not enough to mention that a threat exists; it also has to be discussed what the implications are. A classification of threats to validity is given, e.g., in Wohlin et al. [19].

3.7.3 Inferences

The purpose of this subsection is to describe inferences drawn from the data to more general conditions. This has to be done carefully, based on the findings and the limitations. 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).

3.7.4 Lessons Learned

In this subsection, it has 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.

3.8 Conclusions and Future Work

This section presents a summary of the study. We suggest structuring the *Conclusion and Future Work* section into the following subsections: *Relation to Existing Evidence*, *Impact*, *Limitations*, and *Future Work*.

3.8.1 Relation to Existing Evidence

This subsection is the place where the relation of the results to earlier experiments, especially those mentioned in the *Related Work* section, has to be provided. The contribution of the study should be discussed here. Kitchenham et al. [5] point out that it is important to (1) ensure that conclusions follow from the results and (2) differentiate between the results of the analysis and the conclusions drawn by the authors.

3.8.2 Impact

To enable readers to get the most important findings, we emphasize a description of the impact on cost, time, and quality.

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?

3.8.3 Limitations

In this subsection, principal limitations of the approach have to be described, i.e., circumstances under which the approach presumably will not yield the expected benefits.

3.8.4 Future Work

In this subsection, it has to be described what other experiments could be run to further investigate the results yielded or evolve the BoK.

3.9 Acknowledgements

In this section sponsors, participants, and contributors who do not fulfill the requirements for authorship should be mentioned.

3.10 References

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

3.11 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.

4. Conclusion and Future Work

The contribution of this paper is a (preliminary) proposal of a standardized reporting guideline that unifies the most prominent existing guidelines published by various authors (cf. Table 1). In addition to providing a uniform structure of a reporting template, we have tried to 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. In

some places, for instance for the definition of variables, we suggest a prescriptive formalization schema.

Our proposal has not yet been evaluated, e.g., through a peer review by stakeholders, or by applying it to a significant number of controlled 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. The prerequisite is the general availability of information with regard to the specific experiment. We are aware that this proposal can only be a first step towards a standardized reporting guideline.

The experience of the last 6 years, since the first publication of a reporting guideline for empirical SE research by Singer in 1999 [13], 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 [14]. 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 consent 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., [3], [9], [11], [12].

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 [21] (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 BoK 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.

Acknowledgements

We would like to thank Marcus Ciolkowski, Reidar Conradi, Tore Dybå, Natalia Juristo, Barbara Kitchenham, Dieter Rombach, Janice Singer, Sira Vegas, Claes Wohlin, and many others for fruitful discussions, and the anonymous reviewers for giving valuable feedback, thus helping to improve the paper. Furthermore, we are grateful to Sonnhild Namingha from the Fraunhofer Institute for Experimental Software Engineering for reviewing a previous version of this paper.

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] 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

[4] 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

[5] 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.

[6] 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

[7] 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

[8] 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

[9] 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.

[10] 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 [31], pp. 39-84

[11] 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

[12] Wohlin, C.; Petersson, H.; Aurum, A.: Combining Data from reading Experiments in Software Inspections; In [31], pp. 85-132

[13] Singer, J.: Using the American Psychological 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)

[14] Zelkowitz, M.V.; Wallace, D.R.; Binkley, D.W.; Experimental Validation of New Software Technology; In [31], pp. 229 – 263

[15] 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

[16] 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

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

[18] 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.

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

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

[21] 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

[22] 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

[23] The Editors; Addressing the Limitations of Structured Abstracts (Editorial); In *Annals of Internal Medicine Vol. 140*, No.6 March 2004

[24] Hartley, J; Improving the Clarity of Journal Abstracts in Psychology: The Case for Structure; *In Science Communication, Vol 24*, 3, *2003,*pp.366-379.

[25] Hartley, J; Current findings from research on structured abstracts; *In Journal of the Medical Library Association, 92, 3, 2004,*pp. 368-371.

[26] 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.

[27] 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

[28] 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

[29] 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

[30] 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

[31] Juristo, N. and Moreno, A. (eds.); *Lecture Notes on Empirical Software Engineering*, Ed. River Edge, NJ, USA: World Scientific Publishing, October 2003