Lessons Learned from Evaluating a Checklist for Reporting Experimental and Observational Research

Roel Wieringa
Dept. of Computer Science
University of Twente
The Netherlands
R.J.Wieringa@utwente.nl

Nelly Condori-Fernandez
Dept. of Computer Science
University of Twente
The Netherlands
n.condorifernandez@utwente.nl

Maya Daneva
Dept. of Computer Science
University of Twente
The Netherlands
M.Daneva@utwente.nl

Bela Mutschler
University of Applied Sciences
Ravensburg-Weingarten
Germany
bela.mutschler@hs-weingarten.de

Oscar Pastor
Universidad Poltécnica de Valencia
Spain
opastor@dsic.upv.es

ABSTRACT

This short paper summarizes and discusses the result of an iterative construction and evaluation of a checklist for writing and reading reports about experimental and observational research.

Categories and Subject Descriptors

D.2.0 [**Software Engineering**]: General—*Miscellaneous*

General Terms

Documentation, Experimentation

Keywords

Unified checklist; Experimental and observational research

1. INTRODUCTION

In the past decades, several checklists for empirical research in software engineering (SE) have been proposed, mostly for experimental research [3, 4, 5, 15] but also for case study research [8]. The question that motivated the research summarized in this paper is: What are the differences and commonalities between these checklists? This question is relevant for teachers who want to recommend a checklist to their students and have to answer questions about which parts of which checklists are relevant for a particular research problem.

One obvious difference between checklists for experimental and for observational research is that the first refer to an experimental treatment and the second do not. But beyond this, it is less obvious what are the differences between the checklists. Does the checklist of Jedlitschka & Pfahl [3] supersede all earlier experimental checklists? Is there a reason why it differs from the CONSORT [9] checklist? Why are there so many difference between the non-treatment part of the experimental checklists and the case study checklist [8]? Understanding the differences and commonalities between checklists would help students to interpret the checklists and apply them in an informed manner.

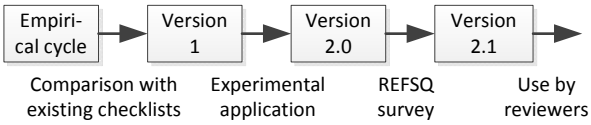

Figure 1: Iterative development of a unified checklist.

The effort to identify a common core of a cluster of checklists for reporting experimental and observational research has led in an iterative way to a new, unified checklist (figure 1). This short paper reports on the status of this iterative development. The effort started with the empirical cycle, which is a rational choice cycle in which the empirical researcher faces the problem to produce support for or against a knowledge claim about a population. This expresses a view of science as critical and rational knowledge acquisition [7]. The first author has used this as a framework for teaching empirical methods since 2007 [14]. Its use to define a checklist for empirical research based on extant checklists is described in [12]. In section 3, we review a number of lessons learned from evaluating the checklist. To save space, we frame the discussion in terms of Version 2.1 of the checklist (see figure 1).

2. VERSION 2.1 OF THE CHECKLIST

Version 2.1 assumes the research model of figure 2. This shows that there is an object of study (OoS), which represents elements of a population of interest. In statistical

ESEM'12, September 19–20, 2012, Lund, Sweden.

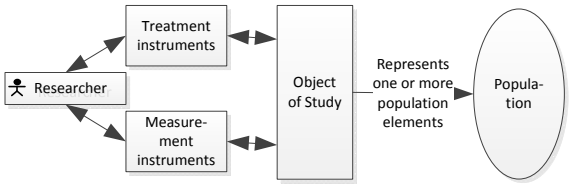

Figure 2: Research model assumed by Versions 2.0 and 2.1.

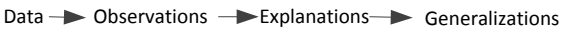

Figure 3: Analysis model assumed by Versions 2.0 and 2.1.

research, the population is the set of entities from which a sample is drawn; in case study research, the population is the set of cases similar to the case under study. Generalization from case studies is based on reasoning by similarity of structure [10], and is fallible, like all generalizations.

The researcher interacts with the OoS, usually mediated by instruments, through measurement interactions in which the researcher collects data from the OoS and aims to minimize influencing the OoS. In experimental research, the researcher additionally applies a treatment to the OoS, and aims to avoid any other interactions than the treatment.

Version 2.1 (table 1) asks for the relation of the research to a **context** of knowledge and practice in questions 1 and 14. The **research problem** is characterized by three items of information: a conceptual framework, research questions stated in terms of this framework, and a population about which the questions are asked. **Research design** follows the research model of figure 2 plus a question that asks how the report describes the reasoning from measurement to conclusions, and questions about the justification (i.e. validation) of design choices. A research report may include information about what actually happened during **execution** of the research design.

The questions under **Results Analysis** assume a list of abstraction levels as shown in figure 3. Each of the arrows represents a fallible inference that must be justified but also critically assessed. *Data* are the measurements taken, consisting of numbers, interview transcripts, etc. *Observations* are summaries of these, obtained by quantitative or qualitative analysis of the data. *Explanations* add some theoretical insight about what could have caused these observations; *generalizations* are claims beyond the OoS.

3. LESSONS LEARNED SO FAR

Version 1 (figure 1) had the same layout as Version 2.1 but has more questions. To construct it, three existing checklists were compared with each other by allocating their items to places in the empirical cycle. The checklists are referred to as JP [3], CT [9] and RH [8] henceforth. JP was selected because it is itself assembled from previous checklists in software engineering. CT was selected because it is the current end point of the development of a checklist for reporting experiments in a different field with a rich experimental tradition. RH was selected because it is the current state of the art of checklists for case study reporting in software engineering. Version 1 is basically the union of all three checklists [13] and contains more questions than Version 2.1 The following observations about Version 1 are also valid for Version 2.1.

- The Population question does not occur explicitly as an item in these other checklists [13]. The two checklists for experiment reports consider this to be part of the description of the sampling procedure. RH take the position that case studies have no populations but that case study findings could be generalized analytically to similar cases [8, page 154]. We agree, and consider this set of similar cases to be the population, so this item is included in the unified checklist.

- CT and RH ask about possible explanations consistent with the observations; JP does not mention explanations. Explanations of observations relate observations to more general theory, and it is theory-building and -validation that advances scientific understanding [11]. So we have kept this item in our unified checklist.

- The two experiment report checklists (JP and CT) ask about generalizations from the observations; RH do not ask about this. However, RH support the idea of analytical generalization to similar cases, as we do [8, page 154], so we include this item in our unified list.

These observations capture the points in which Version 1, and hence Version 2.1, is *not* a common core but adds something with respect to at least one checklist. In addition, Version 1 contains detailed questions about validation and research execution. As explained below, these questions have been removed in Version 2.

We have evaluated Version 1 in a small experiment where seven PhD students and three senior researchers in three different research groups answered the checklist questions about a paper reporting about an observational case study and one reporting about an experiment [2].We give a brief summary of the conclusions.

- The number of questions answered by the subjects decreased when advancing through the 40 questions of the checklist. Our explanation is that Version 1 is too long.

- The answers and comments by the subjects showed us that important terms such as "Unit of Data Collection" (the term used for Object of Study in Version 1) and "treatment" are not understood by the subjects in the way we intended. For example, in our research model (figure 2), which was *not* made available to the subjects, a treatment is an intervention on the OoS by the researcher. This differs from the concept of a treatment that seemed to be used by some of the subjects, namely that a treatment is a level of an independent variable. The effect of an independent variable on a dependent variable can be studied observationally, by blocking on different levels of the independent variable. But these levels are then given to the researcher by nature rather than set by the researcher, as in our treatments. In Version 2, we tried to avoid this confusion by including figure 2 in the checklist.

The list is a classification of information items. Workshop papers present work started, not work finished, and typically contain a subset of the items listed below. Use your own judgment to decide what is acceptable.	
Items to consider	**Possible questions to consider**
Research context	
1. Motivation?	• Is the desired increment of knowledge of the paper clearly stated? • Is the current state of knowledge in this area adequately summarized? • Is the research goal motivated by a practical improvement goal?
Research problem	
2. Conceptual framework?	• Are relevant concepts common knowledge among the authors and readers of the paper? • Are relevant concepts that are not common knowledge, defined and motivated? • Are relevant concepts operationalized?
3. Research questions?	• Are research questions clearly stated?
4. Population?	• Is the target of generalization clear, i.e. for which population are these questions relevant?
Research design and justification	
5. Object of study?	• Is it clear what the object of study is, i.e. the sample, case(s) or model studied? • Is it clear how it was acquired or constructed? • Is it clear what structure it has? • Is validity of the OoS justified, i.e. is it clear why the OoS is representative of the population (e.g. sample size, representativeness of the case or model)? • If people are involved, is ethics discussed (informed consent) ?
6. Reasoning?	• Is it clear what reasoning will be applied to extract observations from the data? (descriptive statistics, and/or qualitative coding of interviews, etc.) • Is it clear what reasoning will be used to answer the research questions? (statistical hypothesis testing, and/or qualitative analysis, etc.) • Is the validity of this reasoning justified?
7. Treatment specification?	• In experimental studies: Is the treatment (i.e. intervention by researcher) specified, including any instruments used? • Is the validity of the treatment discussed? • If people are involved, is ethics discussed (fairness, absence of harm)?
8. Measurement specification?	• Are measurements described, including measurement procedures and instruments? • Is the validity of these measurement procedures and instruments discussed? • If people are involved, is ethics discussed (privacy, respect for people)?
Research execution	
9. What happened?	• Is the report of what actually happened during research useful for the reader?
Results analysis	
10. Observations?	• Are the observations clearly described (graphs, box plots, interview summaries etc.)? • Limitations: Are the interpretations made in extracting observations from data critically assessed on validity?
11. Explanations?	• Are observations explained in terms of underlying mechanisms or of available theories? • Limitations: Are explanations assessed on plausibility?
12. Generalizations?	• Are observations or explanations generalized to the population? • Are population hypotheses tested? • Limitations: Is the plausibility of the hypotheses assessed?
13. Answers?	• Are the research questions answered explicitly? • Limitations: Is the plausibility of these answers assessed?
Research context	
14. Impact?	• Is future research discussed? • Are implications for practice discussed?

Table 1: Version 2.1 of the checklist, to be handed to PC members of a workshop.

- The validity questions in Version 1 were separated from the research design questions, even though they asked about the same items of information; and as stated above, the validity questions were very detailed and non-standard. This confused the subjects. In Version 2 we merged the validation questions with the design questions (as in table 1) and returned to the usual validation concepts: Conclusion, internal and external validity are about the three inferences in the ascent from data in figure 3; Construct validity is the discussion of the validity of the conceptual framework and its operationalizations.

Next, we have conducted a survey among participants of requirements engineering conference[1] in which we asked about understandability and perceived utility of Version 2 [1]. The outcome of this survey encouraged us to do a next evaluation. We have asked program committee members of a workshop[2] to use Version 2.1 when reviewing papers. Version 2.1 differs from Version 2 only in that the Context questions have been reduced to two in total.

4. DISCUSSION

Important differences between different checklists that we found are the absence of the concept of a population and generalization in the case study checklist (RH) and of an item asking for explanations in one experiment checklist (JP). In our proposal, these items are included, because we view research as the critical answering of knowledge questions about a population. The answers are never definitive and empirical research can only provide support that increases or decreases the plausibility of these answers.

So far our claim about this checklist is that this may be useful for readers of a paper to understand what the paper is about. Researchers may use our checklist as a reference point to compare or combine other checklists, or as an independent checklist to read or write papers. In addition, they can use it when they design an experiment of case study, possibly in combination with other checklists geared to those research designs. The checklist does *not* give criteria for evaluating the research reported in a paper: Our claim is that in order to understand what research a paper reports about, a reader, and therefore a reviewer, needs this information. *Judging* research is a different matter. It is difficult and even expert reviewers may differ in their judgment [6]. We intend to do additional evaluations by communities of experts as suggested by Kitchenham et al. [5].

5. ACKNOWLEDGMENTS

Dr. Condori-Fernandez is supported by the EU Marie Curie Fellowship Grant 50911302 PIEF-2010.

6. REFERENCES

[1] N. Condori-Fernandez, M. Daneva, and R. Wieringa. A survey on empirical requirements engineering research practices. In J. Dörr, editor, *Postproceedings REFSQ*, 2012. ISSN 1860-2770.

[2] N. Condori-Fernandez, R. Wieringa, M. Daneva, B. Mutschler, and O. Pastor. Experimental evaluation of a unified checklist for designing and reporting empirical research in software engineering. Technical Report TR-CTIT-12-12, Centre for Telematics and Information Technology University of Twente, 2012.

[3] A. Jedlitschka and D. Pfahl. Reporting guidelines for controlled experiments in software engineering. In *Proceedings of the 4th International Symposium on Empirical Software Engineering (ISESE 2005)*, pages 94–104. IEEE Computer Society, 2005.

[4] N. Juristo and A. Moreno. *Basics of Software Engineering Experimentation*. Kluwer, 2001.

[5] B. Kitchenham, H. Al-Khilidar, M. Babar, M. Berry, K. Cox, J. Keung, F. Kurniawati, M. Staples, H. Zhang, and L. Zhu. Evaluating guidelines for reporting empirical software engineering studies. *Empirical Software Engineering*, 13:97–121, 2008.

[6] B. Kitchenham, D. Sjøberg, P. Brereton, D. Budgen, T. Dybå, M. Höst, D. Pfahl, and P. Runeson. Can we evaluate the quality of software engineering experiments? In G. Succi, M. Morisio, and N. Nagappan, editors, *ESEM*. ACM, 2010.

[7] R. Merton. The normative structure of science. In *Social Theory and Social Structure*, pages 267–278. The Free Press, 1968. Enlarged Edition.

[8] P. Runeson and M. Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14:131–164, 2009.

[9] K. Schulz, D. Altman, and D. Moher. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Annals of Internal Medicine*, 152(11):1–7, 1 June 2010.

[10] P. Seddon and R. Scheepers. Towards the improved treatment of generalization from knowledge claims in IS research: drawing general conclusions from samples. *European Journal of Information Systems*, pages 1–16, 2011. doi:10.1057/ejis.2011.9.

[11] D. Sjøberg, T. Dybå, B. Anda, and J. Hannay. Building theories in software engineering. In F. Shull, J. Singer, and D. Sjøberg, editors, *Guide to advanced empirical software engineering*, pages 312–336. Springer, 2008.

[12] R. Wieringa. Towards a unified checklist for empirical research in software engineering: First proposal. In T. Baldaresse, M. Genero, E. Mendes, and M. Piattini, editors, *16th International Conference on Evaluation and Assessment in Software Engineering (EASE 2012)*, pages 161–165. IET, 2012.

[13] R. Wieringa. A unified checklist for observational and experimental research in software engineering (version 1). Technical Report TR-CTIT-12-07, Centre for Telematics and Information Technology University of Twente, 2012.

[14] R. J. Wieringa and J. M. G. Heerkens. Designing requirements engineering research. In *Workshop on Comparative Evaluation in Requirements Engineering (CERE'07), Delhi*, pages 36–48, Los Alamitos, October 2007. IEEE Computer Society.

[15] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Weslén. *Experimentation in Software Engineering: An Introduction*. Kluwer, 2002.

[1] www.refsq.org

[2] www.bpm.scitech.qut.edu.au/erbpm2012/