2009 17th IEEE International Requirements Engineering Conference

How to Write and Read a Scientific Evaluation Paper

Roel Wieringa
Department of Computer Science
University of Twente
The Netherlands
Email: roelw@cs.utwente.nl

Hans Heerkens
School of Management and Governance
University of Twente
The Netherlands
Email: j.m.g.heerkens@utwente.nl

Björn Regnell
Department of Computer Science
Lund University
Sweden
bjorn.regnell@cs.lth.se

***Abstract*—Scientific evaluation papers investigate existing problem situations or validate proposed solutions with scientific means, such as by experiment or case study. There is a growing amount of literature about how to report about empirical research in software engineering, but there is still some confusion about the difference between a scientific evaluation paper and other kinds of research papers. This is related to lack of clarity about the relation between empirical research, engineering, and industrial practice. In this minitutorial we give a brief rundown on how to structure a scientific evaluation papers as a special kind of research paper, using experiment reports and case study reports as examples. We give checklists of items that a reader should be able to find in these papers, and sketch the dilemmas that writers and readers of these papers face when applying these checklists.**

***Keywords*-Research methodology, Research reporting, Scientific evaluation papers**

I. Introduction

Scientific evaluation papers evaluate existing problem situations or validate or refute proposed solutions by means of scientific research, ranging from formal, mathematical analysis to empirical research. In this minitutorial we are concerned with scientific evaluation by means of empirical research. There is a growing amount of literature about how to report about empirical research in software engineering. Kitchenham et al. [1] provide guidelines for conducting experiments, and Jedlitschka and Pfahl [2] integrate a number of extant guidelines for reporting about them. Kitchenham et al. [3] provide a constructive evaluation of these guidelines and suggest some further extensions. Runeson and Höst [4] integrate extant guidelines for conducting and reporting about case studies. These papers provide ample advice on how to conduct and report about empirical research.

However, scientific evaluation papers often do not relate the research design to the engineering goals of the technical research in which they are embedded. For example, it is often not made clear what exactly the investigated artifacts are, how the research questions relate to theories about the investigated artifacts or how the answers to the questions relate to stakeholder goals, or even what these goals could be. And standard engineering research questions such as trade-off analysis and sensitivity analysis are often omitted.

There is also some confusion about the difference between an evaluation paper reporting on a case study on the one hand, and an industrial experience report on the other. Scientific case study reports often do not contain sufficient information for readers to distinguish them from experience reports; Reflection on validity are often missing.

A third problem is that readers of a scientific evaluation paper sometimes lack the information required to put it into their context. Researchers may not be able to evaluate the paper, for example because the research design or research goal are not clear. Practitioners evaluating the paper on its relevance to their practice may lack information that would allow them to relate it to their own goals and context of practice.

In this minitutorial we give a brief rundown on how to structure a scientific evaluation papers to avoid these problems. In section II we discuss the characteristics of scientific evaluation papers, in particular what distinguishes them from other kinds of research papers. In section II we give a checklist for reports about scientific evaluation, and apply this to an experiment report and a case study report. The checklist is consistent with the checklists mentioned above but less detailed. The reader is encouraged to consult the above papers to find out about more detailed checklists. In section IV we mention some dilemma's that writers and readers face when applying these checklists.

II. Scientific evaluation papers

We view research as the *critical* acquisition of knowledge. The word "critical" means that the researcher does not claim more than he or she can justify, and to this end reflects on all possible ways in which he or she could be wrong, and also exposes the result results and the support for it to a critical peer group of researchers. This is true for conceptual research, such as mathematics, which defines and analyzes concepts, and for empirical research, which investigates observable phenomena. In this paper we are concerned with empirical research.

This view implies that scientific *reporting* is essential to achieve the aim of scientific research. Basically, a scientific research report describes what the researcher wanted to know, what he or she did to answer this question, and

1090-705X/09 $25.00 © 2009 IEEE
DOI 10.1109/RE.2009.17

- Problem statement
  - Research questions
    - . What do we want to know?
  - Unit of study
    - . About what?
  - Relevant concepts & theory
    - . Meaning of key terms
    - . What do we know already?
  - Research goal
    - Why do we want to know?
- Research design
  - What are we going to do to answer the questions?
- Validation of the design
  - Is this going to answer our questions?
- What actually happened
- Analysis of results
- Interpretation

Figure 1. Characteristics of all research papers, including scientific evaluation papers.

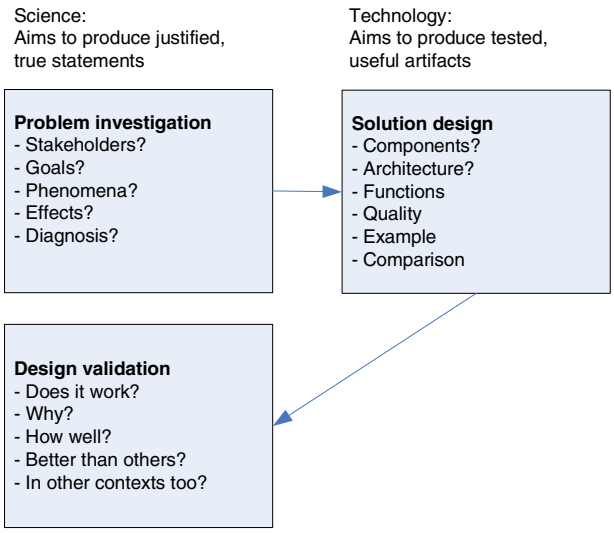

Figure 2. Scientific evaluation papers investigate problems or solution designs.

critically analyzes the conduct and outcome of the research in order to check what could be wrong about it (Figure 1). More in detail, a research report should contain a *problem statement* that lists the research questions as well as the subject of the questions, called the unit of study. Key concepts used in the questions must be defined, and any relevant knowledge already available must be stated. An important part of the problem statement is a statement of research goals, which can be pure curiosity, but in the case of engineering research always also includes utility for some practical purpose [5].

To illustrate, here is the problem statement of an experiment report by Bratthall and Wohlin [6]:

- Investigate an artifact (e.g. a method, technique, software tool, algorithm, notation, . . .)
- that may be already in use (problem investigation) or not yet used (design validation);
- Identify stakeholders interested in the artifact (e.g. requirements engineers, clients, software engineers, . . .)
- Identify relevant goals the stakeholders have (e.g. to interview users, to specify requirements, . . .)
- Describe how the artifact can be used to reach these goals (e.g. interview technique, specification technique, . . .)
- And how this relates to other ways to reach these goals (e.g. it is faster, cheaper, more reliable, easier, . . .)
- Ask technical research questions (figure 2).

Figure 3. Characteristics of scientific evaluation papers in addition to those listed in figure 1.

- Unit of study: Population is all possible software architecture diagrams with decorations to represent qualitative information
- Research questions: Are certain diagram features (e.g. rectangle size) intuitive representations of certain quality attributes (e.g. code size)? Some possible answers are stated in the form of hypotheses.
- Relevant concepts & theory: Software architecture concepts, cognitive psychology concepts (e.g. intuitive is interpreted as cognitive accessibility weight).
- Research goals: Utility. The graphical techniques are proposed to improve representation of properties of an architecture in an architecture diagram (in any architecture diagram language).

And here is the problem statement of a case study report by Regnell et al. [7].

- Unit of study: Distributed requirements prioritization processes in market-driven software product development.
- Research questions: How does a proposed method to prioritize requirements actually work?
- Relevant concepts & theory: Requirements, product development, prioritization.
- Research goals: Utility. Improve distributed requirements prioritization processes.

Scientific evaluation is a particular kind of research, performed as part of the engineering cycle [8]. Figure 2 shows that the goal of these papers is to investigate an existing problem situation or a proposed solution design. Typical research questions in investigating a problem are what the stakeholder goals are, what the problematic phenomena are, how they can be explained and what their effects are. Typical research questions when validating a solution design are whether the solution works at all, why, how well, whether this is better than other solutions, and in which contexts it can work.

For example, the experiment report by Bratthall and

Wohlin [6]

- Investigates an artifact: Graphical decoration of architecture diagrams;
- That is an invention by authors, not used in practice;
- Identifies stakeholders interested in the artifact: Novice programmers and part-time programmers;
- Identifies relevant goals the stakeholders have: to identify candidate areas for reengineering existing software;
- Describes how the artifact can be used to reach these goals: by representing quality aspects of code in an architecture diagram; this is better than having to search for this information in archives;
- Ask technical research questions:
  - Can quality attributes be represented this way? (Does it work?)
  - Which way is most understandable? (How well?)

Another relevant technical research question, not asked by the authors in this paper, is whether the proposed technique would help in identifying candidate areas for reengineering, as claimed in the motivation for the research.

The case study report by Regnell et al. [7]:

- Investigates an artifact: Distributed prioritization in market driven RE;
- Invented by others, used in practice;
- Identifies stakeholders interested in the artifact: requirements engineers, marketing, customer support, product development;
- Identifies relevant goals the stakeholders have: Increasing product sales, finding new customers;
- Describes how the artifact can be used to reach these goals: Find best prioritization that helps stakeholders achieve these goals;
- Asks a technical research question: How is the prioritization related to long-term product strategy? (I.e. How does it work in practice?)

III. Checklists

Figure 4 lists criteria for all elements of research papers. This is an elaboration of a checklist we used earlier [9] and it is consistent with the lists mentioned above. This checklist can be specialized according to the kind of report, e.g. experiment report or case study report. Due to space limitations we will not do that here. The application of the problem statement checklist to Bratthall and Wohlin [6] and Regnell et al. [7] has been given above and we here alsoapply the research design checklist. For Bratthall and Wohlin [6] this becomes:

- Unit of data collection: Some architecture diagrams in simple notation with graphic decorations.
- Environment of data collection: The laboratory. 35 subjects, consisting of master students and PhD students.
- Measurement instruments: A questionnaire.

- Problem statement
  - Unit of study described?
    Research questions given?
  - Relevant concepts & theory described?
  - Research goals stated?
- Research design
  - Unit of data collection described?
  - Environment of data collection described?
  - Measurement instruments described?
  - Measurement procedures described?
  - Data analysis procedures described?
- Validity discussed?
- Research execution described?
- Analysis of results
  - Observations given?
- Interpretation
  - Possible explanations given?
  - Answers to research questions given?
  - Reflection on validity present?
  - Implications for research goals drawn?

Figure 4. Checklist for scientific evaluation paper.

- Measurement procedures: 15 seconds to answer each question.
- Data analysis procedures. AHP to identify a ranking of understandability, removal of outliers, boxplots to describe results, Kruskall-Wallis and ANOVA to reject hypotheses, PLSD tests to analyze ANOVA results.

The research design of Regnell et al. [7] is as follows:

- Unit of data collection. One particular distributed requirements prioritization process at one particular multinational company.
- Environment of data collection. The field: The company.
- Measurement instruments. A questionnaire.
- Measurement procedures. Questionnaire sent by email to 10 stakeholders.
- Data analysis procedures. Qualitative analysis of answers.

IV. Dilemmas

Checklists cannot be a replacement for judgment, and applying them may place the reader or writer for dilemmas. For the writer, these revolve around the question how to balance the investment in paper writing against other possible investments of his or her time. To what level of detail should one adhere to these (or other) checklists? What is the chance of success and what would be the drivers of success?

The reviewer too needs to judge how to invest time. Dilemma's faced by the reviewer include the following: Why should I review a paper rather than do something else? Is the topic interesting? Do I like the author? Do I like the research method? What will be the opinion of others if I do the review, or decline to do it? How will my opinion reflect back on me? How much work is it to understand the paper? How can I do this with minimum effort?

The reviewer also needs to reflect on the validity of his or her opinion. Wrong reasons for rejecting a research paper would be, among others, that the investigated artifact is not novel (rejection for this reason this would prevent accumulation of knowledge about an artifact) or that only a single case was studied (this would rule out case study research). Wrong reasons for acceptance would be, among others, mere interestingness of the results that are otherwise of low validity, or admiration for the sophistication of statistical techniques used to produce results that are insignificant for theory as well as for practice.

The practitioner is faced with similar dilemma's but will look for clues whether the knowledge produced by the paper is applicable to a relevant practical problem. Dilemma's faced by a practitioner include the following: Are the claims of the paper valid in this situation. How much would it cost to apply it? What would be the benefit of using it? Who would bear the costs and who would enjoy the benefits? How can I find this out with minimum effort?

Answers to these questions cannot be given by following yet another checklist. However, the author of a paper can help by providing information in the abstract, introduction or conclusion that will help the reader to find out the answers to these questions quickly.

REFERENCES

[1] B. Kitchenham, S. Pfleeger, D. Hoaglin, K. Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 721–733, August 2002.

[2] 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)*. IEEE Computer Society, 2005, pp. 94–104.

[3] 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*, vol. 13, pp. 97–121, 2008.

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

[5] D. Stokes, *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press, 1997.

[6] L. Bratthall and C. Wohlin, "Is it possible to decorate graphical software design and architecture models with qualitative information? –an experiment," *IEEE Transactions on Software Engineering*, vol. 28, no. 12, pp. 1181–1193, December 2002.

[7] B. Regnell, M. Höst, J. Natt och Dag, P. Beremark, and T. Hjelm, "An industrial case study on distributed prioritisation in market-driven requirements engineering for packaged software," *Requirements Engineering*, vol. 6, pp. 51–62, 2001.

[8] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: A proposal and a discussion," *Requirements Engineering*, vol. 11, no. 1, pp. 102–107, March 2006.

[9] R. Wieringa and J. Heerkens, "The methodological soundness of requirements engineering papers: A conceptual framework and two case studies," *Requirements Engineering Journal*, vol. 11, no. 4, pp. 295–307, 2006.