Towards a Unified Checklist for Empirical Research in Software Engineering: First proposal

Roel Wieringa
University of Twente
http://www.cs.utwente.nl/~roelw

Abstract—Background: Current checklists for empirical software engineering cover either experimental research or case study research but ignore the many commonalities that exist across all kinds of empirical research. Identifying these commonalities, and explaining why they exist, would enhance our understanding of empirical research in general and of the differences between experimental and case study research in particular. Aim: In this short paper we design a unified checklist for empirical research, that identify commonalities and differences between experimental and case study research. Method: We design the unified checklist as a specialization of the general engineering cycle, which itself is a special case of the rational choice cycle. The unified checklist is based on an analysis and integration of a number of existing checklists. Results: The current version of the checklist exhibits a shared structure of experimental and case study research. Conclusions: Although the checklist exhibits a shared underlying structure of empirical research, its limitations are that it ignores other research methods such as meta-research or surveys and that very little empirical validation of usability and utility of the checklist has been performed so far. We are currently performing one validation, and are planning additional ones. These validations will likely lead to improvements of the current proposal.

Keywords—Empirical research methodology, unified checklist, experimental research, observational research

I. INTRODUCTION

Since 1995, several checklists for experimental software engineering have been published [1], [2], culminating for the time being in a proposal for an integrated checklist by Jedlitschka and Pfahl [3], who also included in their sources Kitchenham’s checklist for systematic reviews [4]. Recently, also Runeson and Höst published a checklist for case studies [5], itself based on an analysis of existing checklists for case study research in different disciplines [6].

Although experimental and case study research are different kinds of research, they have more in common than one would expect at first sight; after all, there is a reason to call these different kinds of knowledge acquisition activities research and this reason should show up in common parts of these checklists. Identifying and explaining these commonalities would produce insight in the underlying structure of empirical research, and this insight in turn could help practicing software engineering researchers to make justified decisions about what to include in their research designs and reports. The goal of this short paper is to identify these commonalities and present a unified checklist that brings out as much as possible of the underlying, shared, structure of different kinds of empirical research.

We start by sketching how empirical research fits into the logical structure of engineering tasks, called the engineering cycle (section III), which in fact is an engineering cycle of which the goal is to acquire knowledge. The unified checklist presented here (section IV) is the result of applying the empirical cycle as a template to compare and analyze various checklists for experimental and case study research [7].

II. THE ENGINEERING CYCLE

Our starting point is the engineering cycle (figure 1) discussed more in detail elsewhere [7], [8]. This is a rational choice cycle in which the engineer investigates improvement possibilities (stakeholders and goals, problematic phenomena and their causes), and then designs, implements and evaluates a treatment (figure 1). This is easily understood in medical terms, where a medical researcher may investigate a disease, design and validate a treatment, after which the treatment is implemented by transferring it to the market, and then evaluated by continuous monitoring. But the engineering cycle is more generally applicable and has been recognized as the logical structure of any engineering activity [9], [10], [11].

The treatment usually consists of the interaction between an artifact (medicine), which in our case may be a physical device, software, techniques, notations, etc., and the problem context (human body), which in our case may be a software project, software system or some physical system.

Validation consists of estimating what effects the treatment would have if implemented (validation question 2), whether this would meet the stakeholder goals (question 3), why the effects would occur (question 4), what trade-offs are involved (5) and how sensitive this is to changes in the problem context (6).
Implementation is transfer to practice. It is not just building a prototype, but transferring the treatment to the problem context where stakeholders will apply it (patients will take the medicine).

In implementation evaluation, exactly the same questions are asked as in problem investigation, but this time with the goal to find out whether the treatment has produced the desired effects.

There are two important research tasks in the engineering cycle: (1) problem investigation/implementation evaluation and (2) design validation. The research questions asked in these tasks are numbered 1-6 in figure [1] Note that questions 1-4 are similar to the validation questions, but they have a different form because in validation there is no instance of the treatment yet. The task of validation is rather to predict what effects the treatment of the treatment yet. The task of validation is rather to predict what effects the treatment would have if it would be implemented in practice. Engineering researchers typically build prototypes of the designed artifact, and exercise this in simulated problem contexts, to be able to make these predictions.

III. THE EMPIRICAL CYCLE

We view empirical research as a rational approach to solving the problem to acquire justified true knowledge about the real world. Rational problem-solving consists of a sequence problem investigation, design of alternative solutions, validating these and choosing one, implementing it and evaluating the result of implementation. The empirical cycle of figure 2 has this structure.

- Research problem investigation. The stakeholders are at least the researchers themselves and anyone else who depends on the knowledge to be acquired. Their knowledge goal, in the context of a higher-level engineering cycle, is to investigate an engineering problem, or to evaluate an implementation, or to validate a newly proposed design not transferred to practice yet. The criteria to be applied to the acquired knowledge are always the same: Is it true? Is it justified?—both suitably qualified according to uncertainty of the researcher’s answers.

- Research design. There are many possible research designs, and any attempt to classify them would not do justice to the almost infinite variety of possible designs. So far, the unified checklist has been assembled out of checklists for observational cases studies and experiments.

- Research design validation. Before the research design is implemented, we check whether it would really answer the research questions (validation question 2), how certain are we about this (3) and what justification we have for this (4). We will consider alternative designs (5) and also how sensitive the design is to assumptions about the research context. For example must it be executed in the field or could it be executed in the laboratory?

- Research execution. While the design is executed, unexpected events may occur, and deviations from the design or partial redesigns may be called for. If not covered by a contingency plan, these events must be responded to on-the-fly, maintaining validity of the choices.

- Results evaluation. Evaluation of the results includes answering the research questions and assessing our (un)certainty about these, as well as explanations of the observations in terms of existing or newly postulated theories.
### Research problem investigation

- U1 What is the higher-level engineering cycle?
- U2 Knowledge goal in that cycle?
- U3 Conceptual model of the phenomena?
- U4 Conceptual model validity? (including construct validity)
- U5 Unit of study (population)?
- U6 Research questions?
- U7 Current knowledge?

### Research design

- U8 Unit of data collection? (sample, model or case)
  - 8.1 Acquisition?
  - 8.2 Structure?
- U9 Treatment of unit of data collection?
  - U9.1 Treatment specification?
  - U9.2 Treatment assignment?
  - U9.3 Treatment plan?
  - U9.4 Treatment instruments?
- U10 Measurement of unit of data collection?
  - U10.1 Measurement procedures?
  - U10.2 Measurement instruments?
- U11 Kind of reasoning? (statistical or case-based)

### Research design validation

- U12 Validity of unit of data collection?
  - U12.1 External validity?
  - U12.2 Ethics?
- U13 Validity of treatment?
  - U13.1 Instrument validity?
  - U13.2 External validity?
  - U13.3 Ethics?
- U14 Validity of measurement?
  - U14.1 Validity of measurement procedures?
  - U14.2 Instrument validity?
- U15 Validity of reasoning?
  - U15.1 Conclusion validity?
  - U15.2 Internal validity?

### Research execution

- U16 Unit of data collection?
  - U16.1 Acquisition?
  - U16.2 Quality?
  - U16.3 History?
- U17 Execution of treatment?
- U18 Execution of measurements?
- U19 Availability of data?
- U20 Provenance of data?

### Results evaluation

- U21 Data?
- U22 Observations?
- U23 Explanations?
- U24 Answers to research questions?
- U25 Generalizations?
- U26 Limitations?
- U27 Contribution to knowledge goals?
- U28 Contribution to engineering goals?

---

Figure 3. A checklist for empirical research.

## IV. A unified checklist

The current proposal has been assembled from a number of other checklists [6, 3, 2, 12, 1, 5]. This has led to successive versions of the unified checklist, which in our research group we have applied in our own research as well in teaching empirical research methods to Master’s and PhD students. The result is listed in figure 3.

### A. Research problem investigation

If there is a higher-level engineering cycle in the context of which this empirical research is performed, then this cycle should be identified (U1) and the goal of this research in that cycle should be stated (U2): problem investigation, implementation evaluation, or design validation.

To state the research questions, the relevant conceptual model may have to be described (U3) and validated (U4). A conceptual model is a collection of concepts and their relations. In some cases, the concepts used in research questions are understood and agreed on among the writer and all readers of a research report, but in other cases, there may be ambiguities and relevant concepts must be explicitly defined. For example, in an empirical study of effort estimation practices, relevant concepts such as effort and program size must be defined.

If a conceptual model is explicitly defined, its validity must be motivated (U4). Validity of a conceptual model includes construct validity. For example, the concept of "usability of a notation" must be operationalized in terms of observable indicators, and this operationalization must be valid.

Research questions (U6) presuppose a population about which these questions are asked, such as the population of all distributed software engineering projects or of all service-oriented architectures. To avoid the impression that we are biased towards statistical studies, we use the term unit of study (UoS) to indicate arbitrary elements of the population (U5). When research questions are stated, then extant knowledge apparently is insufficient to answer them satisfactorily. This requires a discussion of current knowledge about these questions (U7).

### B. Research design and its validation

To acquire knowledge about the UoS, the researcher must collect some data from an entity that we call "Unit of Data Collection" (UoDC). In a statistical study, the UoDC will be a sample of existing UoS’s of sufficient size; in a case study, it will be a small set of UoS’s. The intention of the researcher is to study the UoDC and then draw conclusions about UoS’s in general.

The first set of questions to be answered concerns the UoDC: (U8.1) How is it to be acquired? (U8.2) What structure does it have? To answer U8.1 in the case of sampling, the sampling process must be described; for case studies, case study selection must be described. To answer U8.2 for samples, sample size and grouping should be described, and should be related to expected effect size and desired power of the test [13, 14]; for case studies, the structure of the case in terms of units of analysis and other relevant structure information must be described.
These decisions must be motivated by considerations of external validity (U12.1) and, if applicable, ethics (U12.2). For samples, justification of external validity requires justification of the representativeness of the sample for the population with respect to the research questions; this will include justification that the size of the sample is sufficient with respect to the expected effect size to be measured, but it will also include considerations of homogeneity of the sample with respect to the relevant variables as compared to the homogeneity of the population. For cases, justification of external validity requires justification of the similarity of this case to other UoS’s with respect to the research questions, and an argument should be given why, and to what extent, this similarity can be expected to support any generalization from this case to other UoS’s [15].

In experimental research, some control is exercised over the UoDC, consisting of a treatment (U9.1) applied to the UoDC (U9.2) according to a plan (U9.3), possibly using instruments (U9.4). In statistical experiments, the UoDC is a sample, and application of the treatment (U9.2) includes assigning the treatment to subjects. In randomized controlled trials (RCT’s), there are at least two treatments and this requires dividing the sample into groups and assigning different treatments to different groups. Treatment may involve instruments, such as instructional material for human subjects.

The decisions made in treating the UoDC must be motivated in terms of validity of instruments (U13.1), external validity (U13.2) and, if applicable, ethics (U13.3). For example, will the planned instruction to human subjects indeed prepare them to participate in the experiment? (U13.1) Is the treatment applied similar to the treatment applied in practice to all UoS’s? (U3.2) And if the UoDC consists of people, is their integrity respected? (U13.3)

In observational case studies, no treatment is applied, but the researcher may want to design safeguards against exercising any influence on the case. We consider this to be part of the justification of the validity of measurement procedures and so the checklist does not mention it here.

In all kinds of research, the empirical researcher will take measurements. These must be instrumented (U10.2) and measurement procedures must be designed (U10.1). The instruments and procedures must be justified by explaining why the instruments measure the indicators of interest (U14.2) and the procedures must not disturb the phenomenon to be measured (U14.1).

Finally, a plan must be made for reasoning from raw data to observations, and from observations to explanations (U11). Quantitative data, consisting of numbers, may have to be transformed, and can be described using descriptive statistics, using diagrams and other representations. We consider these representations to be the observations made. For hypothesis testing, the statistical inference procedures to be used must be planned and justified (U15.1), which is conclusion validity.

Qualitative data, consisting of words, must be coded and translated into observations by a process that must not insert any of the beliefs of the coders about the topic of the qualitative data. The coding procedures must be planned ahead, and their validity justified (U15.1). This is the qualitative analogue of conclusion validity.

Once the observations have been extracted from the data, the researcher wants to explain them in terms of preceding causes or underlying mechanisms. Internal validity (15.2) is the question whether these explanations are valid. During research design, internal validity must be justified by excluding as many controllable causes that could explain the effects to be measured other than the treatment applied. In statistical inference, randomization of samples is a major tool to exclude any other explanation of the observed effect than the applied treatment. If the subjects are people, then some of the factors to be controlled for are for example maturation and history [16].

In case study research, additional causes that could explain what is observed, cannot be controlled, and therefore they must be documented in the case description, so that the reader can assess whether these are plausible alternative explanations of the observed effect.

C. Research execution

During research execution, events may occur that influence the interpretation of results and are therefore relevant for the researcher trying to understand the results, or may be relevant for the researcher aiming to replicate the research. The checklist again follows the elements of research design.

Events during acquisition of the UoDC may be reported (U16.1), and events that impact the quality of the UoDC as a source of data may be reported (U16.2). For example, the sample finally assembled may not be the intended size, or be more heterogeneous than originally hoped for. Or the case actually acquired may not exhibit all features that would make it similar to the UoS’s in the population of interest. Also, during the execution, events may occur to the UoDC, such as drop out of subjects, that are worth reporting (U16.3).

The implementation of the treatment of the UoDC may contain events worth reporting about too (U17), and similarly the measurement may contain unexpected events relevant for the researcher and would-be replicator of the research (U18).

Finally, data, as far it is not confidential, should be made available in some way (U19) and the provenance (traceability) of data to the points of measurement should be recorded (U20).

D. Results evaluation

The full data set is rarely published in a report, but any transformations (e.g. data set reduction) should be reported,
and a brief summary can be given (U21). Observations can be reported by means of descriptive statistics, characteristic fragments from interviews can be reported, etc. (U22). Explanations for these observations in terms of previously known theories or mechanisms can be provided, or new theories or mechanisms can be postulated that would explain the observations (U23). All of this must be used to provide answers to the research questions (U24), which may include the outcome of tests of hypotheses.

From observations and explanations applicable to the UoDC, generalizations about the UoS can be inferred (U25). All of these results, from observations to generalizations, are uncertain, and the uncertainties have to be summarized as limitations of the study (U26).

Finally, a research report should identify contributions to knowledge (U27), which refers back to the state of knowledge reported in answering U7, and contributions to the engineering goal (U28), which refers back to any higher level engineering cycle identified in answering U1.

V. EVALUATION

The unified checklist has been compared in detail [17] with the checklist for experiments by Jedlitschka and Pfahl [3], for randomized controlled trials by the CONSORT group [12], and the checklist for case studies by Runeson and Höst [5]. Since the time of writing, our research group has conducted an empirical evaluations of the unified checklist by asking a sample of researchers to apply it to the analysis of two papers, one paper reporting about an observational case study, and the other reporting about a controlled experiment. In a second empirical evaluation we have asked a sample of researchers to give their opinions about the usability and usefulness of the checklist. We will report about the results of these experiments in the near future.

ACKNOWLEDGMENT

I would like to thank Maya Daneva and Nelly Condori-Fernandez for their useful comments on an earlier version of the report.

REFERENCES


