A Unified Checklist for Observational and Experimental Research in Software Engineering (Version 1)

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

***Abstract*—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. In this report we design a unified checklist for empirical research, and identify commonalities and differences between experimental and case study research. We design the unified checklist as a specialization of the general engineering cycle, which itself is a special case of the rational choice cycle. We then compare the resulting empirical research cycle with two checklists for experimental research, and with one checklist for case study research. The resulting checklist identifies important questions to be answered in experimental and case study research design and reports. The checklist provides insights in two different types of empirical research design and their relationships. Its limitations are that it ignores other research methods such as meta-research or surveys. It has been tested so far only in our own research designs and in teaching empirical methods. Future work includes expanding the comparison with other methods and application in more cases, by others than ourselves.**

***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 two book-length introductions to experiment design [4], [5] as well as Kitchenham's checklist for systematic reviews [6]. Recently, also Runeson and Höst published a checklist for case studies [7], itself based on an analysis of existing checklists for case study research in different disciplines [8].

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 report is to identify these commonalities and produce 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 II). Next, we will present empirical research itself as an engineering problem, namely the problem how to acquire knowledge about the real world. This perspective allows us to sketch a high-level version of the *empirical research 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. In addition, we have used the checklist in our own research and in our methodology courses.

To illustrate the meaning of this checklist in terms of both experimental and of case study research, we compare it with the checklists for experimental research of Jedlitschka and Pfahl [3] and of the CONSORT group [9], [10], and of Runeson and Höst for case study research [7] (section V).

I agree with Moher et al [10] that "the format of articles should abide by journal style; editorial directions; the tradition of the research field addressed; and, where possible, author preferences", and so we do not discuss guidelines for structuring papers, such as section headings. The unified checklist that we come up with are lists of questions that should be asked when preparing research, and that should be asked when writing or reading a research report.

We take the perspective of researchers who want to design their research, to write, or evaluate a report about research, or to replicate the research [11], [12]. Due to page limitations, we ignore the practitioner perspective in this report. In section VI we discuss the validity of what we know about the checklist and section VII summarizes what has been achieved and what still needs to be done.

II. The engineering cycle

Our starting point is the engineering cycle (figure 1) discussed more in detail elsewhere [13], [14]. This is a

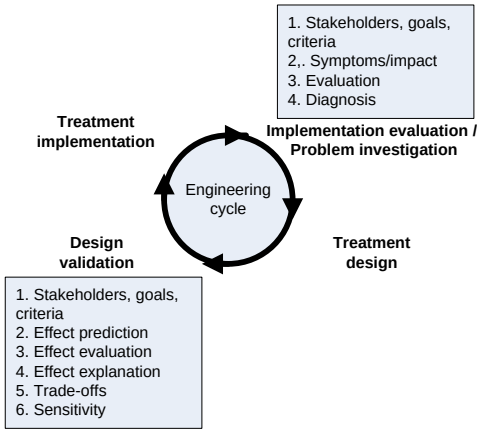

Figure 1. The engineering cycle.

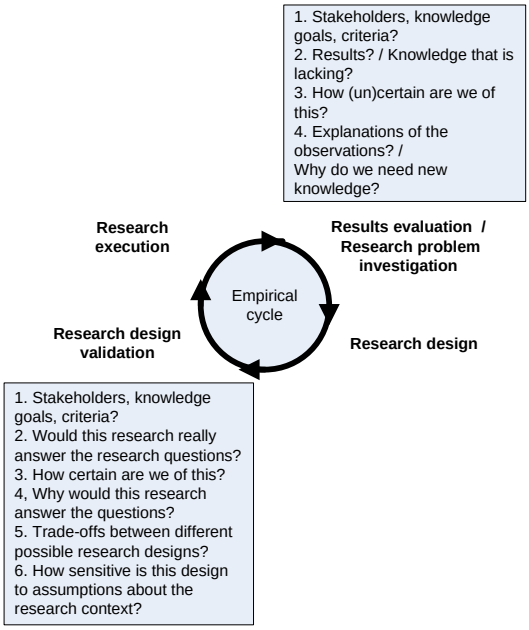

Figure 2. The empirical cycle.

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 [15], [16], [17].

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 also asked in the validation task. But in validation there is no instance of the treatment used in practice 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. Modelling and simulation as research methods are not considered in this report.

III. The Empirical Cycle

We view empirical research as an application of the engineering cycle to a particular kind of problem, namely the problem to acquire justified true knowledge about the real world. The resulting empirical cycle (figure 2) consists of the following tasks:

- **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. It is, however, useful to indicate a few of the most important choices to be made in research design. Figure 3 shows a few. It shows two kinds of

decisions to be made about the *Unit of data Collection* (UoDC), which is the part of the world where the researcher will get his or her data from. The UoDC can be a sample of the population, or it can be a model of elements of the population. A *sample* is a subset of the population. Case studies may choose a sample of size one, but in statistical studies the sample is always larger. A *model* is an entity that represents a population element with respect to a property. This means that observations of this property in the model provide information about this property in population elements. For example, with respect to the property of being able to use a notation, students may be used as models of professional software engineers, and with respect to its behavior in a queue, a software agent can be constructed and used to represent an airplane taxiing at an airport. In statistical simulations, a model is a set of elements, each of which models an entity in the population.

One kind of decision is whether or not the UoDC will be treated by the researcher, e.g. will receive a stimulus, input, medicine, instruction, etc. with a view to finding out how the UoDC responds to this. In *observational research,* the researcher administers no treatment, in *experimental research,* he does.

The other dimension is whether the researcher intends to use statistical reasoning or case-based reasoning about the UoDC. In *statistical reasoning,* the UoDC is a sample from some population and the researcher aims to statistically estimate some feature of a population distribution from observations of a sample. In a variation of this, the researcher can do statistical experiments with a statistical model, which is a set of models of elements of the population.

An example of observational research with statistical reasoning is a survey. An example of experimental research with statistical reasoning about a sample is a randomized controlled trial, where the effect of a treatment applied to one group is compared to the effect of another treatment applied to another group. An example of experimental research with a statistical model is a simulation of the behavior of airplanes taxiing on an airport by a set of software agents.

Case-based reasoning is reasoning by analogy between cases. For example, an *observational study* of a single case could be the basis of a generalization to other, analogous, cases.

An *action case study* is a single case to which the researcher applies a treatment in order to help a real-world person or organization. An example is a researcher who has designed a new technique for relating business objectives to enterprize architectures, and is now using this in a company to solve a problem of that company, in order to help the company and to learn how the technique performs in practice.

	No treatment by researcher (observational study)	**Treatment by researcher (experimental study)**
Statistical reasoning	• Survey of a sample	• Statistical experiment with a sample; • Statistical experiment with a statistical model
Case-based reasoning	• Observational case study of a small sample	• Action case study; • Case-based modelling experiment

Figure 3. Some kinds of research. Only statistical experiments and observational case studies are discussed in this report.

A *case-based modelling experiment* is similar, except that the UoDC is not a person or organization to be helped, but an artifact used as model of an arbitrary element of a population. For example, an experiment performed on a prototype of an algorithm can be used to learn how all future implementations of the algorithm would behave in the real world. The researcher is then reasoning by analogy from the behavior of this prototype to the behavior of future implementations.

The checklist in this report is aimed at observational case studies and randomized controlled trials, but figure 3 suggests that we can extend it to other kinds of research.

- **Research design validation.** Before the research design is implemented, we check whether it would really answer the research questions (validation question 2), how certain we are 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.

IV. A UNIFIED CHECKLIST

The first version of this checklist has been published in 2007 [18] and since then I have used it to compare and analyze many other checklists [19], [20], [21], [8], [3], [5], [2], [1], [22], [7], [4]. This has led to successive versions of the unified checklist, which in our research group we have

- **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?
    - 13.1 Instrument validity?
    - 13.2 External validity?
    - 13.3 Ethics?
  - U14 Validity of measurement?
    - U14.1 Validity of measurement procedures?
    - U14.2 Instrument validity?
  - U15 Validity of reasoning?
    - 15.1 Conclusion validity?
    - 15.2 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 4. A checklist for empirical research.

applied in our own research as well in teaching empirical research methods to Master's and PhD students. The result is listed in figure 4. Before we discuss it, a note about its use:

If the checklist is used for research design, then the questions in research problem investigation, research design and design validation must at least be asked, and for any question that cannot be answered, there should be a valid reason why the question is not relevant. For example, in observational case study research design, no treatment is applied, and so question U9 is not relevant.

If the checklist is used for research reporting, then also the parts about research execution and results evaluation should be used. The checklist is *not* an outline for a research protocol (research plan) nor an outline of a research report, because we ignore formatting of protocols and reports here. Rather, it indicates which questions must be answerable by the plan or report, and if some question is not answerable, the list reminds the researcher that he or she must have a valid reason for not being able to answer it. For example, in many cases not all answers to the questions about research execution may be relevant enough to include in a report.

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; in a case study, it will be a

single UoS. 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 [23], [24]; 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 [25].

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 then to participate in the experiment? (U13.1) Is the treatment applied similar the 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 [26].

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. Comparison with Other Checklists

The checklist in figure 4 is applicable to all experimental research; if we drop U9 (treatment of UoDC), it is applicable to observational research too. For different kinds of observational and experimental research, further specialization is needed. I illustrate this by comparing it with the checklists for experimental research provided by Jedlitschka & Pfahl [3] (J&D henceforth) and the CONSORT group [10], [9], and with the checklist for observational case studies provided by Runeson & Höst [7] (R&H henceforth). The CONSORT group (Consolidated Standards of Reporting Trials) is a group of scientists and editors in medical research that aims to improve reporting of RCT's in medical research. I repeat that these checklists have been used to construct the unified checklist; so what follows is not a validation of the unified checklist, but an illustration of how it relates to extant checklists.

The checklist items in figure 5 to 9 have been numbered as they are in the references used. Excluded are items related to report formatting[1] and also items that summarized earlier parts of a checklist from a reader's perspective[2].

A. Research problem investigation

U1 The three checklists tend to ignore the possibility that empirical research may be part of a higher level engineering cycle (U1). J&D do mention a problem statement and from their explanation of this item it is apparent that they have in mind a problem and solution with benefits for stakeholders. We interpret this as the identification of a higher-level engineering cycle.

U2 By "research objectives", all three checklists mean high-level versions of the research questions, rather than an indication of the knowledge goal in a higher-level engineering cycle, which is intended in the unified checklist.

U3 J&D and CONSORT's conceptual model of the world is that it exists of variables. They require these variables to be defined (U3). J&D additionally view the experiment as being structured in terms of subjects who perform tasks on objects. R&H do not require such a conceptual model for the case to be defined, although one can view such a model to be part of the theoretical basis of the case.

U4 We regard the requirement of the two experimental checklists that the definitions of variables used in experiments be standardized, as part of a validity check of these definitions.

U5 All three checklists ignore the concept of unit of study. Yet the specification of the UoS (population) is important, as it may not be clear what the intended target of generalization is: For example, all software engineering projects that have existed and could possibly ever exist in the world, or all agile projects, etc. The CONSORT statement assumes that the population is all people, or all people with a certain disease, or possibly all instances of some disease.

U6 All three checklists require the statement of research questions.

U7 All three checklists require the specification of current knowledge in the form of related work, scientific background, or theoretical basis (U7).

B. Research design

U8 The UoDC in the two experimental checklists is a sample, and eligibility criteria for a UoS to be in the sample must be specified (CONSORT 4a), grouping and assignment of treatment must be specified, and settings and locations specified. For case studies, the

[1] "Structured abstract", "Motivation", "Acknowledgments", "References" and "Appendices" from J&F. "Title and abstract" from CONSORT. Items 28-38 (reporting checklist) from R&H.

[2] Items 39-50 (reader's checklist) from R&H.

Unified checklist: Research problem investigation	**Jedlitschka & Pfahl [3]**	**CONSORT [10], [9]**	**Runeson & Höst [7]**
U1 Engineering cycle	3.2.1 Problem statement 3.3 Related work		
U2 Knowledge goals	3.2.2 Research objective	2a Background and objectives	2 Research objectives
U3 Conceptual model	3.4.1 Variables, Parameters 3.4.3 Subjects 3.4.4 Objects	6a Outcome measures	
U4 CM validity	3.4.1 standardized measures	6a Public and standardized outcome measures	27 Threats to construct validity
U5 Unit of study (population)			
U6 Research questions	3.4.1 Research goals & hypotheses	2b Specific objectives or hypotheses	2 Research questions, and hypotheses (if any)
U7 Current knowledge	3.3 Related work	2a Scientific background	3 Theoretical basis

Figure 5. Research problem investigation. For ease of reference, the three quoted checklists are numbered as in the papers referenced to.

Unified checklist: Research design	**Jedlitschka & Pfahl [3]**	**CONSORT [10], [9]**	**Runeson & Höst [7]**
U8.1 Unit of data collection: acquisition	3.4.3 Subjects sampling strategy	4a Eligibility criteria for participants	1 What is the case, what are its units of analysis
U8.2 Unit of data collection: structure	3.4.3 Sample size, subject characteristics relevant to similarity 3.4.4 Objects 3.5.2 Preparation of the sample (e.g. grouping) 3.4.2 Experiment design (structuring the groups) 3.2.3 Context	7a Sample size 3a Trial design (structuring the groups, unit of randomization, allocation ratio) 4b Settings and locations	4 Author's'intentions clear? 5 Is the case adequately defined? 8 Rationale for selection?
U9.1 Treatment of UoDC: Treatment specification		5 Intervention specifications (sufficient detail for replicator) 11b Similarity of treatment and placebo (if relevant)	(Not applicable)
U9.2 Treatment of UoDC: Treatment assignment	3.4.2 Experiment design 3.5.2 Preparation (e.g. training)	8a Random allocation sequence generation method 8b Type of randomization 9 Allocation concealment mechanism 11a Blinding (if done) 13a Participant flow	
U9.3 Treatment of UoDC: Treatment plan	3.5.3 Data collection performed (includes experiment schedule)	5 Interventions: how and when administered	
U9.4 Treatment of UoDC: Treatment instruments	3.4.4 Instrumentation		
U10.1 Measurement: procedures	3.4.6 Data collection procedure	6a How and when are outcomes to be assessed	7 Data or method triangulation 11 Protocol for data collection 12 Data triangulation 13 Measurement (instruments and) procedures
U10.2 Measurement: instruments	3.4.5 Instrumentation		13 Measurement instruments (and procedures)
U11 Kind of reasoning (statistical, case-based)	3.4.7 Analysis procedure	12a Statistical methods to compare groups on outcomes 12b addition statistical methods (e.g. subgroup analysis)	6 Is cause-effect relation under study? 11 Procedures for data analysis defined? 22 Analysis methodology defined?

Figure 6. Research design. In the original checklists, the items with question marks (?) are categorized under research reporting only, but we consider them to be part of design too.

case must be selected and described, and the rationale for its selection must be explained.

U9 The CONSORT statement is most detailed in its requirements for specifying treatment of the UoDC: The intervention must be specified, meaning that the treatment and placebo must be specified, and it must be assessed whether the participants can discern any difference between them. CONSORT requires specification of the type of randomization and indication of the way in which the randomized assignment of treatment to participants performed so far, is concealed from those doing assignment. In some experiments, blinding is possible too, and must then be described. (In software engineering experiments blinding is not possible.) A treatment plan must be designed. The CONSORT checklist does not contain an item explicitly mentioning instrumentation used during treatment.

U10 All three checklists mention measurement procedures. R&H stress triangulation as a way to reduce observer bias. Note that this indicates that the researchers have a qualitative case study in mind, in which data consists of words to be interpreted by the researcher. CONSORT does not contain an item that explicitly mentions measurement instruments.

U11 The two experimental checklists are about statistical experiments and so contain items asking for a specification of the statistical techniques to be used, and of the way they will be used. The case study checklist contains items that ask whether analysis procedures are defined without asking which procedures are defined.

C. Research design validation

- The two experimental checklists mention validity in general terms but do not decompose the concept in subconcepts.
- The case study checklist explicitly mentions ehtics and integrity of subjects, which indicates that the cases studied contain people.

D. Research execution

U16 All three checklists ask for various kinds of information about hat happened during the execution of the research, such as what actually happened when acquiring the sample (U16.1) and why an experiment was stopped early, if applicable. The case study checklists also asks whether the case actually acquired is actually suitable to answer the research questions and whether the phenomenon of interest, for example agile development, was implemented correctly. We interpret this as the question whether the UoDC (a single case) is sufficiently similar to the UoS's in the population to support interesting generalizations from the UoDC to the UoS.

U17 CONSORT is very detailed about what happened during application of the treatment: who actually enrolled the participants, flow of participants, how many dropped out, why etc. Changes in the treatment method must be reported too, with reasons for change.

U18 All checklists require a report about what happened during data collection. Here too CONSORT allows changes in measurement variables, but requires these to be reported and explained with reasons.

U19 CONSORT and R&H require information about the availability of raw data.

U20 Only the case study checklist requires traceability of data to observations and onwards to explanations (theories) and to answers of research questions.

E. Results evaluation

U21 No one requires all data to be available in a report, but data reduction should be described (J&D) and an illustrative synopsis can be given (CONSORT)

U22 J&D require descriptive statistics of the data; CONSORT requires baseline data, which are demographic data that allow a reader to see if the outcomes would be applicable to his or her own case. CONSORT also requires an exact description of participant flow to show whether the analysis done at the end, was applied to all participants that went into the experiment. And CONSORT requires any mention of adverse effects of the treatment on participants.

U23 All checklists require observations to be explained. R&H explicitly state that more than one explanation may be applicable.

U24 All three checklists have an item for answering research questions. In the two experimental checklists, the research questions are hypotheses, and answering them here means statistically testing these hypotheses.

U25 Generalizability (external validity) is mentioned in the two experimental checklists but not in the case study checklist.

U26 The two experimental checklists mention limitations. They both seem to view this as an discussion of all validity issues. Note that this appears *after* research execution, as this item refers to limitations in execution of the experiment and in the analysis of the results.

U27 All three checklists have an item asking for the addition to knowledge produced by this research.

U28 The two experimental checklists have an item asking for the impact of the results, which we interpret here as referring back to the engineering goal that motivated the research (U1). The case study checklist mentions recommendation for practice under the heading of general conclusions.

Unified checklist: Research design validation	Jedlitschka & Pfahl [3]	CONSORT [10], [9]	Runeson & Höst [7]
U12.1 Validity of UoDC: external validity	3.4.8 Validity evaluation 3.5.4 Validity procedure		27 Threats to external validity 21 Can we answer the research questions with these data (?)
U12.2 Validity of UoDC: Ethics			15 Approval by review board? 10 Integrity of subjects?
U13.1 Validity of treatment Instrument validity			(Not applicable)
U13.2 Validity of treatment: External validity			
U13.3 Validity of treatment: Ethics			
U14.2 Validity of measurement: Validity of measurement procedures			14 Are the methods of measurement sufficient to fulfill the objectives of the study?
U14.1 Validity of measurement: instrument validity			
U15.1 Validity of reasoning: Conclusion validity			
U15.2 Validity of reasoning: Internal validity			

Figure 7. Research design validation.

Unified checklist: Research execution	Jedlitschka & Pfahl [3]	CONSORT [10], [9]	Runeson & Höst [7]
U16.1 UoDC: Acquisition	3.5.1 What happened during sampling?	14a Recruiting 25 Funding	
U16.2 UoDC: Quality			9 Is the case relevant for the research questions? 17 Observed phenomenon correctly implemented?
U16.3 UoDC: History		7b When applicable, explanation of any interim analyses and stopping guidelines 14b: Why the trial ended or was stopped.	
U17 Execution of treatment	3.5.2 Preparation (e.g. what happened during training) 3.5.3 Data collection performed	3b Important changes in methods after trial commencement 5 Interventions: how and when administered 10 Implementation: who generated the allocation sequence, who enrolled the participants, who assigned them to interventions 13a Participant flow: number of participants per group, numbers of participants who were randomly assigned, received treatment, were actually analyzed 13b Participant flow: losses and exclusions after randomization, with reasons	(Not applicable)
U18 Execution of measurements	3.5.3 Data collection performed	6b Changes to outcome variables after trial commenced, with reasons	16 Is data collected according to protocol?
U19 Availability of data		23 Registration number and trial registry 24 Where the full trial protocol can be accessed, if available	18 is data recorded to allow analysis? 19 Are sensitive results identified?
U20 Provenance of data			20 Are data collection procedures traceable? 23 Chain of evidence from data to research questions to theory?

Figure 8. Research execution.

Unified checklist: Results evaluation	Jedlitschka & Pfahl [3]	CONSORT [10], [9]	Runeson & Höst [7]
U21 Data	3.6.2 Data set reduction	20 Brief synopsis of findings	
U22 Observations	3.6.1 Descriptive statistics	15 baseline data: demographic and clinical characteristics of each group 16 For each group, number of participants included in each analysis and whether this was the number in the original group 19 Important harms or unintended effects in each group	
U23 Explanations	3.7.1 Evaluation of results and implications	20 Possible mechanisms 22 Interpretation consistent with results	23 Explanations 24 Alternative explanations 25 If a cause-effect relation is studied, can the cause be distinguished from other factors?
U24 Answers to research questions	3.6.3 Hypothesis testing	17a Outcomes and estimation (including effect size, and precision such as confidence interval) 17b For binary outcomes, both absolute and relative effect sizes 18 Ancillary analyses (e.g. subgroup analyses)	23 Answers to research questions outcome of hyopthesis testing
U25 Generalizations	3.7.3 Inferences to more general conclusions	21 Generalizability (external validity, applicability)	
U26 Limitations	3.7.2 Limitations of the study 3.8.3 Limitations	20 Limitations (e.g. sources of bias, imprecision, multiplicity of analyses)	
U27 Contribution to knowledge goals	3.8.1 Relation to existing evidence	20 Comparison with relevant findings from other published studies	26 Conclusions
U28 Contribution to engineering goals	3.8.2 Impact	20 Clinical implications	26 Conclusions

Figure 9. Results evaluation.

VI. Validity

A checklist is a tool for thought, and like any tool should be judged on usability and utility. *Usability* is here the property of being usable by researchers in their practice. This aspect of the unified checklist has been tested to a limited extent by using the checklist in our own research (we can apparently use it) and by teaching it to PhD students (at least they understand the checklist, which is a prerequisite for usability).

Utility is the property of usefulness for a purpose. Would using the checklist lead to better designs than not using the checklist? Would a report that can answer all the checklist questions be more understandable for the reader, be better in knowledge transfer from writer to reader, be more informative for the replicator, be more useful for the meta-analyst than a report that cannot answer all questions in the checklist? Would a report written using the checklists be regarded by experienced researchers, for independent reasons, as better than a report written without using the checklist? We expect the answers to these questions to be positive due to the way the checklist was constructed: From a fundamental view of the structure of research, and from an assembly of existing checklists into this structure. But we have no empirical evidence for this other than the perceived utility of these checklists in our own research: It improved our understanding of our own research. Future research of the use of these checklists by others is needed to answer these questions.

VII. Discussion and Further Work

The unified checklist can be used for both experimental research and for observational case study research, but for observational research, item (U9) should be skipped, because in observational research there is no treatment. All of the other questions are relevant in both kinds of research, although of course the answers to some of the design questions will be very different in sample-based research using statistical inference, and case-based research using reasoning by analogy.

The comparison with three other checklists shows that in particular experimental research can require some detailed decisions about sample selection and treatment assignment that are not covered by the unified checklist.

The comparison also shows that there is no attention to generalization in R&H's checklist(U25). There are however ways to generalize from case studies [25]. One requirement for generalization is that the conceptual model of UoS's in the intended population be clearly defined, and the comparison shows that this element is absent from R&H's checklist

too.

Further work on the unified checklist includes more comparisons and applications, including an application to methods for modelling and simulation and for action research. We are also planning an evaluation of the unified checklist as a checklist for reporting, using the approach of Kitchenham et al [11].

ACKNOWLEDGMENT

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

REFERENCES

[1] S. Pfleeger, "Experimental design and analysis in software engineering," *Annals of Software Engineering*, vol. 1, no. 1, pp. 219–253, 1995.

[2] 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.

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

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

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

[6] B. Kitchenham, "Procedures for performing systematic reviews," Keele University/National ICT Australia, Tech. Rep. TR/SE-0401/0400011T.1, 2004.

[7] 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.

[8] M. Höst and P. Runeson, "Checklists for software engineering case study research," in *Proceedings of the First International Symposium on Empirical Software Engineering and Measurement*. IEEE Computer Science Press, 2007, pp. 479–481.

[9] D. Moher, S. Hopewell, K. Schulz, V. Montori, P. Gøtzsche, P. Devereaux, D. Elbourne, M. Egger, and D. Altman, for the CONSORT Group, "CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trial," *British Medical Journal*, p. 340:c869, 2010.

[10] K. Schulz, D. Altman, and D. M. D, "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials," *Annals of Internal Medicine*, vol. 152, no. 11, pp. 1–7, 1 June 2010.

[11] 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.

[12] R. Wieringa, H. Heerkens, and B. Regnell, "How to write and read a scientific evaluation paper," in *16th IEEE International Requirements Engineering Conference, Atlanta, U.S.A.* IEEE Computer Society Press, 2009, pp. 361–364.

[13] 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.

[14] R. J. Wieringa, "Design science as nested problem solving," in *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, Philadelphia*. New York: ACM, 2009, pp. 1–12.

[15] N. Cross, *Engineering Design Methods: Strategies for Product Design. Second Edition.* Wiley, 1994.

[16] J. Eekels and N. Roozenburg, "A methodological comparison of the structures of scientific research and engineering design: their similarities and differences," *Design Studies*, vol. 12, no. 4, pp. 197–203, October 1991.

[17] N. Roozenburg and J. Eekels, *Product design: Fundamentals and Methods*. Wiley, 1995.

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

[19] D. Cooper and P. Schindler, *Business Research Methods, 8th edition*. Irwin/McGraw-Hill, 2003.

[20] J. Creswell, *Research design: Qualitative, quantitative, and mixed methods*. Sage, 2009.

[21] E. Babbie, *The Practice of Social Research*. Thomson Wadsworth, 2007, 11th Edition.

[22] C. Robson, *Real World Research*. Blackwell, 2002, second Edition.

[23] J. Miller, J. Daly, M. Wood, M. Roper, and A. Brooks, "Statistical power and its subcomponents — missing and misunderstood concepts in empirical software engineering research," *Informat6ion and Software Technology*, vol. 39, pp. 285–295, 1997.

[24] V. Kampenes, T. Dybå, J. Hannay, and D. Sjøberg, "A systematic review of effect size in software engineering experiments," *Information and Software Technology*, vol. 49, no. 11–12, pp. 1073–1086, November 2007.

[25] P. Seddon and R. Scheepers, "Other-settings generalizability in IS research," in *International Conference on Information Systems (ICIS)*, 2006, pp. 1141–1158.

[26] W. Shadish, T. Cook, and D. Campbell, *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, 2002.