University of Twente

Information System Group

Technical Report

An Experimental Evaluation of a Unified Checklist for Designing and Reporting Empirical Research in Software Engineering

Nelly Condori-Fernandez
Roel Wieringa
Maya Daneva
Bela Mutschler
Oscar Pastor

Technical Report nr: TR-CTIT-12-12
Date: 14-05-2012

University of Twente (The Netherlands)

An experimental evaluation of a unified checklist for designing and reporting empirical research in Software Engineering

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

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

Maya Daneva
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 Politécnica de Valencia
Valencia, Spain
opastor@dsic.upv.es

ABSTRACT

This article reports on an experimental evaluation of a unified checklist for case study and experimental research. The checklist aims at exhibiting the underlying, shared, structure of observational and experimental research, and is based on several published checklist in software engineering and other disciplines. We asked a small sample of subjects to apply it to two papers, one reporting on a case study and one reporting on an experiment. After this experience, we collected their perceptions of the perceived ease of use and usefulness of the checklist. The outcome reveals that respondents find some of the concepts used in the checklist hard to understand, thus difficult to be used by the subjects (researchers). However they also think a unified checklist like this would be useful for their empirical research. We discuss the findings and their limitations, and infer some implications for improving the checklist.

1. INTRODUCTION

In the past decades, several checklists for empirical research in software engineering (SE) have been proposed, mostly for experimental research [1][6][5] but also for case study research [17]. The checklists for experimental research are all slightly different from each other, and they all differ considerably from the checklist for case study research. This proliferation of checklists hides underlying similarities between the different forms of research (namely experiments and case studies) which, if they would be identified, would enhance our understanding of what is essential for any kind of research, as well as would improve the applicability of a unified checklist.

Recently, Wieringa proposed a unified checklist, based on a logical analysis of the empirical research cycle and a comparison of existing checklists inside and outside SE [23]. In this paper, we report on our empirical assessment of the understandability, perceived usefulness and ease of use of this unified checklist while researchers are reviewing reports on experimental and case study research. We present the unified checklist in Sect. 2 and our conceptual framework and research questions in Sect. 3. Sect. 4 discusses related work on checklists. Sect. 5 and Sect 6 are about our research design and its execution. Sect. 7 reports on the results. Sect. 8 concludes with our learnings and suggestions for further improvement of the unified checklist.

2. THE UNIFIED CHECKLIST

Wieringa's unified checklist proposed follows a rational problem solving cycle as shown in Figure 1. Figure 2 shows the questions of the unified checklist partitioned in five blocks according to the empirical cycle [23]. The unified checklist has several intended users, including researchers or practitioners. We next describe the first three blocks of checklist items from the point of view of the research designer, and the last two blocks from the point view of the reader of research reports.

During **research problem investigation** (phase I), the checklist asks whether there is a higher-level engineering cycle (U1), by which we mean an improvement goal, e.g. "improve the practice of requirements inspection", or "improve the efficiency of the testing process". The knowledge goal (U2), in the context of a higher-level engineering cycle [22], is to investigate an engineering problem (for example, to learn about the impact of personality characteristics on the performance of pairs), or to evaluate an implementation, or to validate a newly proposed design not transferred to practice yet. Formulating research questions requires a conceptual model of the phenomena (U3), in which concepts like programming pair and personality characteristic are defined.

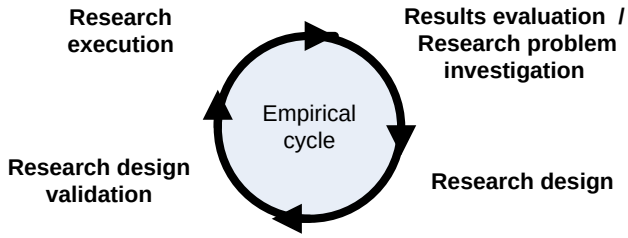

Figure 1 The empirical cycle, adapted from [22]

I. 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?
II. 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)
III. Research design validation
U12 Validity of unit of data collection?
    U12.1 External validity?
    U12.2 Ethics of treatment?
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?
    U14.3 Ethics of measurement?
U15 Validity of reasoning?
    15.1 Conclusion validity?
    15.2 Internal validity?
IV. 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?
V. 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 2 The unified checklist for empirical research

U4 asks whether the validity of this conceptual model is justified. This includes justification of the construct validity of any operationalizations of concepts. U5 asks what the unit of study (UoS) is, i.e., the UoS is the term we use for elements of the population. The population is the target of any generalization based on the findings of an empirical study. The (U6) research questions are questions about the UoS. Finally, U7 asks the researcher to consider what current knowledge is available about these questions.

Research design (phase II) starts with (U8) deciding on where data are to be collected (the unit of data collection(UoDC): from a *sample,* or from a *model* of an arbitrary UoS, or a few UoS's, to be called *cases*, to be studied in-depth). The researcher must decide (U8.1) on how the sample or case is selected, or how the model is constructed, respectively. This decision prompts the researcher to think about (U8.2) the structure of the UoDC: the required size of the sample and properties of its elements; or the desired structure of the model or case, respectively.

The choice of UoDC is directly linked to (U11) choices about the kind of reasoning to be applied, as well as to (U12) the validity of these choices. Suppose a researcher decides for a sample as the UoDC. In statistical reasoning about a sample, the sample must be large enough with respect to the expected effect size to be able to draw statistically significant conclusions about the population. By contrast, in case-based reasoning, the researcher reasons from observations about a model or a case to conclusions about arbitrary UoS's. For example, a real-world agile SE project may be treated as a typical case of any agile SE project with respect to the property of requirements management. In **research design validation** (phase III), the researcher must (U12.1) justify the representativeness of the UoDC with respect to the intended population. Finally, if the UoDC contains people, the researcher must (U12.2) justify that these are treated with respect, i.e. that they are aware of their participation as research subjects and consent with this.

For experimental research, there will be a treatment to be applied to the UoDC (U9). The treatment must be (U9.1) specified, and (U9.2) assigned to the elements of the sample, and the application of the treatment to the sample must be (U9.3) planned. (U9.4) Instruments may be needed to apply the treatment, such as an instruction sheet to instruct subjects. The external validity of (U13.2) the treatment as well as of (U13.1) the instrument, must be justified. For example, are the programming tasks given to pairs in the experiment representative, with respect to the properties of interest, to the programming tasks performed by pairs in the intended population? If there are human subjects, (U12.2) is the application of the treatment fair and is no harm done?

In all kinds of empirical research, (U10) measurements will be taken. This will require (U10.2) instruments, such as questionnaires, and (U10.1) the taking of measurements must be planned. The validity of the (U14.2) instruments and (U14.1) measurement procedures must be justified. For example, does the questionnaire indeed measure the intended constructs, and is there no influence of the measurement procedure on the constructs measured? If the UoDC contains people, then (U14.3) the researcher must consider ethical questions such as confidentiality and privacy of data.

We now switch to the perspective of a reader of a research report. To understand a report, a reader would need answers to all applicable questions so far. Only a few questions are not applicable in some cases: There is (U1) no higher-level improvement goal in curiosity-driven research, no (U9) treatment in observational case study research, and in research that does not involve people, there are no ethical issues.

During **research execution** (phase IV), events worth reporting about may happen. For example, (U16.1) what sample or case was actually acquired and how was this constructed or selected? (U16.2) Are there properties of the sample or case that are important for the reader to be aware of to understand the results, but have not been planned for in the research design? (U16.3) Did anything happen to the sample or case that the reader should know about, such as drop out of subject or changes in the structure of the case? Were (U17) the treatment applied and (U18) measurements taken as planned, or

are there events worth reporting about? (U19) Are the data available to the reader, possibly anonymized? (U20) Is there a trace from data to the place and time of measurement?

For **results evaluation** (phase V), the checklist asks if (U21) fragments of data are presented in the report, and (U22) how these data were transformed into observations. For example, in quantitative data, outliers may be removed and scales be transformed. In qualitative data, interviews may be coded into keywords. (U23) To *explain* observations means to postulate possible causes of the observations. For example, if an experiment did not show impact of emotional instability of a programmer on the performance of programmer pairs, one possible explanation is that the programmers have learned to manage their emotional instability, and another possible explanation is that the experimental programming tasks were simple enough to avoid stress. A report should list these possible explanations and assess their plausibility. This should lead to (U24) answers to the research questions, and the plausibility of (U25) possible generalization of these answers to the population should be considered. No observation, explanation or generalization is infallible, and a report should (U26) discuss their limitations. Finally, referring back to the first two checklist items, (U27) a report should summarize the increment of knowledge acquired and (U28) the contribution to any improvement goals that can be derived from this.

3. THE RESEARCH PROBLEM

We use a design science approach in which we study an artifact in its context of use [24]. Our artifact is the unified checklist, and its context of use is a research setting in which a researcher is using it to do the empirical research (e.g. designing), write reports about it and read reports (e.g. for evaluating purpose). We have three questions about this UoS:

RQ1. Is the checklist understandable by the researcher?

RQ2. Is the checklist perceived by the researcher as easy to use when reviewing a paper?

RQ3. Is the checklist perceived as useful in designing and reporting experiments and case studies?

The motivation for RQ1 is that understandability by the user is a minimum requirement for usability of the checklist. RQ2 asks for the self-perception of users about ease of use and RQ3 asks for perceived usefulness.

Drawing on perspectives on understandability in cognitive studies based on [1], for the purpose of this research, we define *understandability* (RQ1) as the degree to which information contained in a checklist can be easily understood by a reader. It constitutes of two aspects: the effort required to read (efficiency) and the correct interpretation of the checklist questions (effectiveness). The understandability effectiveness (understandability level (UL)) is measured by the harmonic mean of the recall and precision measures, which we adapted from [16]: UL=2*Precision*Recall/(Precision + Recall)

Precision (p) is the number of correct answers divided by the number of all answered questions by the reader (correctly or incorrectly). **Recall (r)** is the number of correct answers divided by the number of questions that should have been answered. The best UL score is 1 and the worst score is 0.

Although in this experimental assessment, the understandability efficiency (effort required) was not measured objectively (e.g. units of hours required to answers the questions), it was measured based on the user's perception (RQ2).

In order to evaluate the perceived ease of use (RQ2) and usefulness (RQ3), some items from the Method Evaluation Model [13] were adapted:

Perceived Ease of Use (PEOU): the extent to which a person believes that using a particular checklist would be free of effort.

Perceived Usefulness (PU): the extent to which a person believes that a particular checklist will be effective in achieving intended objectives.

We measured both variables on a 5-point Likert scale. In our context to evaluate this perception, our objective identified is to provide guidance for obtaining better designs and reports of cases studies and experiments.

4. RELATED WORK

The benefits of using checklists are well accepted and described in many papers, e.g. in [20]: the chance of forgetting to check something important is reduced and the impact of the halo effect is cut down as a checklist forces evaluators to take into account all relevant dimensions of merit.

For empirical research, there also exist checklists: a well-known one, for example, is provided by Jedlitschka & Pfahl [5], another - by the CONSORT group [12], [19]. A checklist for observational case studies is provided by Runeson & Höst [17]. A detailed comparison of these three checklists with Wieringa's unified checklist is given in [23].

Generally, the validation of evaluation checklists themselves is only addressed by few authors. Martz [11], as one rare example, presents a validation study of the *Organizational Effectiveness Evaluation Checklist* (OEC) [10], a checklist for evaluating organizational business and industry settings. In [9], Kitchenham et. al assess the quality of published human-centric SE experiments. The goal of their study was thereby not only to evaluate published human-centric SE experiments based on a quality evaluation checklist, but also to confirm the usability of their checklist and to specify an appropriate process for evaluating the quality of SE experiments.

Kitchenham et. al [8] also investigate quality checklist proposals for systematic literature reviews in SE. They conclude that available checklists do not provide sufficient help neither to novices nor to experienced researchers. In turn, existing checklists only provide some value as a starting point for constructing a unified evaluation checklist for systematic literature reviews.

5. RESEARCH DESIGN

5.1 Experimental subjects

No element of the population of interest (researchers using this checklist in their real-life research settings) exists yet, so we are going to simulate them. We planned to ask up to 18 researchers in three research groups (in Twente, Valencia and Weingarten) to apply the checklist to four papers, two reporting on experiments and two reporting on case studies. The 18 subjects selected by convenience, were asked to participate by the three research leaders (Mutschler, Pastor and Wieringa) of the respective research groups. 10 subjects voluntarily accepted to participate in the experiment (5 from Twente, 2 from Valencia and 3 from Weingarten).

The sample is homogeneous in the sense that they do scientific research in the same discipline and some of them publish at the same conferences. However, there may be a significant difference in research experience (2 senior researchers, 1 Post doc, 7 PhD Students), which may impact the results. Also, all the subjects from Twente had been exposed to the engineering cycle approach [22]. This may influence the understandability of the checklist. In a demographic questionnaire, the respondents were asked to provide their number of years of experience in research, their exposure to other checklists for research (See Figure 3), and their experience in conducting case studies or experiments.

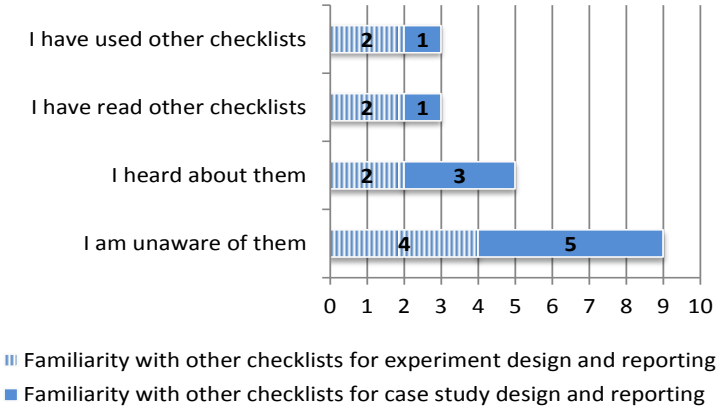

Figure 3. Familiarization level with other checklists

Considering the familiarization level with the engineering cycle, two groups of subjects were considered in this experiment (familiarized or not familiarized).

5.2 Treatment

A treatment is something that experimenters administer to experimental units. In our experiment two empirical papers (one reporting on an experiment [18] and one reporting on a case study [4]) were assigned to the subjects. The papers selected are of high quality, which minimizes the possibility that problems to understand the paper would prevent any use of the unified checklist at all. Fig. 4 illustrates how checklist questions were presented to the subjects. If the respondent clicks on "Defined" for this question, this triggers the text window to pop up, so that the respondent could optionally summarize what parts of the paper's text were used to answer the question. In what follows, we will refer to the closed list of possible answers to a question as the *closed* part of the answer, and to text entered in an optional text box as the *open* part of the answer. The open questions are introduced with the purpose of verifying whether the respondent's reason for answering the closed part of the question is correct.

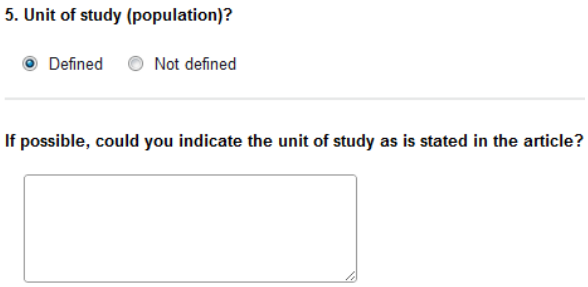

Figure 4. Screenshot of one checklist question.

We sent the treatments to the participants by email with instructions to review the two papers using the checklist with its respective reference guide (explanations for each item of the checklist). We asked each subject to review the two papers (within subjects-design) in a particular order, which could differ for different subjects. Each paper was reviewed on a different day between 13 and 19 Feb 2012. The duration of each review was not limited.

Subjects were also asked to fill in a questionnaire about their perception of the usefulness and ease of use after applying the checklist to the two papers. Responses were anonymous as participants generated a random identifier from a web site[1] and used this to identify all forms that they filled in (one checklist per paper and a questionnaire). The questionnaire used a 5-point Likert scale (5=Strongly Agree and 1=Strongly Disagree). Perceived Ease of Use (PEOU) was measured using five items on the survey (Questions 1, 2, 3, 6, and 10); Perceived Usefulness (PU) was measured using two items on the survey (Questions 4 and 5). By using the Cronbach's alpha the generic value obtained for the checklist was 0.78, indicating that the items on the survey are reliable, as claimed by Nunally [14]. The Cronbach's alpha value for PEOU was 0.754, and PU was 0.795.

The checklist and questionnaire were implemented with the SurveyGizmo tool[2]. A link to the web version of the instructions, checklist, reference guide, and questionnaire can be found at: http://users.dsic.upv.es/~nelly/checklist-eval.html

The experimental procedure and implementation of the instrumentation was managed by Condori-Fernandez. To test the treatment procedure, we did two pilot studies. First, Daneva and Wieringa applied the procedure on the original planned set of four papers to see if the instructions and guidelines were unambiguous. Second, two master's students and two senior researchers from UT tested the experiment material. A positive feedback was received from the second pilot, which allowed improving the reference guide and adjust the initial treatment plan (four papers in two different days) by reducing the set of papers to two.

Our experimental context is rather hypothetical because we do not expect anyone to use the checklist in full merely to understand a paper; more likely, parts of it may be used. However, full application of the checklist may be useful when replicating earlier research; and a shorter version of the checklist may be useful when evaluating a paper. We do not consider these other possible uses of the checklist in this experiment

We make the note that subjects had to spend considerable time on the experiment. While this does not cause harm, it is a sacrifice and might cause decreased reliability of subjects' answers towards the end of the checklist. To stimulate their participation, we conducted a lottery with three first prizes among the subjects who completed all forms.

5.3 Measurement

To measure the understandability (RQ1), Daneva and Wieringa produced a reference model with the correct answers for the two papers. They first applied the checklist independently from each other to the papers, and then compared their answers to resolve any differences. This led to a decision about a reference answer to each open and closed question. Using this reference model, the correct answers to closed questions were rated as 1, and the incorrect answers as 0. The answers to open questions were rated as 2 (correct), 1 (partially correct), and 0 (incorrect). Our measurement procedure allowed subjects to fill in the checklists and questionnaire in their own time within a week, and if two respondents learned from each other that they were participating in the same experiment, this might have impacted the data. But this would have shown up as similarity or identity of textual answers, and we have found no evidence of such influence in the data.

[1] http://www.brenz.net/services/randomID.asp

[2] https://appv3.sgizmo.com

6. EXECUTION

All subjects returned completed checklists for the two papers reviewed (five in Twente, three in Weingarten, two in Valencia). However, one of them did not complete the questionnaire. The data have been collected automatically in anonymous form and cannot be traced by us to the respondents. Once the data was collected in an excel file, Condori-Fernandez, Daneva and Wieringa rated the open questions by using the reference answers. The closed questions were rated automatically by means of a formula that was defined for each question. During this coding process, our reference answers were adjusted a few times. We have kept a trace of these changes. The data, including these changes, are available from the authors on request.

7. ANALYSIS AND INTERPRETATION

7.1 Understandability

The precision and recall for each block of the checklist was calculated. We provide some possible explanations of our observations, and provide tentative answers to our research questions. We label each of our observations. We first mention two observations from the pilot studies.

O1. We already made the note that a piece of information, asked about in a checklist question, can be more or less present, depending on the interpretation of the rater (the one who applies the checklist). For example, the two papers [18],[4] did state research questions but did not explicitly return to them. However they did provide the information to answer the research questions. How shall question U24 be answered: Answers to research questions "Described" or "Not Described"? The answer depends on the desire of the rater for explicitness. So even though our scales are nominal, the rater's judgment is sometimes a point on a continuous scale.

O2. When applying the checklist, we felt a normative pressure. When the checklist asked whether an item X was described in the paper, it was easy to slip into the interpretation that X *had to be* present. But there is no such intention of the checklist. For example, some papers are curiosity-driven and there is no higher-level engineering cycle (U1).

We now turn to the analysis of the answers by the respondents. We summarize many observations in terms of vague quantifiers such as "most respondents" or "many respondents", because we feel that giving exact numbers here would convey a false sense of exactness. As will become apparent in the sequel, our observations are rough and approximate.

7.1.1 Research problem investigation (PI)

O3. Most respondents answered the closed questions U1 and U2 about higher-level improvement goals, and about knowledge goal of the paper, correctly. However, the open part of the answers showed that some respondents mentioned a knowledge goal, such as to learn about the impact of neuroticism on pair programming, as an improvement goal.

O4. The case study paper [4] defines the concept of global SE (GSE) and provides several definitions of trust, but does not compare the validity of these definitions for the goal of this paper (studying ways to manage trust in GSE). Most respondents correctly identified the conceptual model described in the case study paper but half of them said that validation is present in the paper, rather than absent.

The experiment paper [18] defines the concept of pair programming and gives literature references for the concept of personality characteristic, and assumes that the operationalization of personality characteristics in measurable properties, taken from personality psychology, is valid. Most respondents could not find this conceptual model and gave a variety of incorrect answers to the question whether the paper described conceptual model validity. One explanation of this mixed performance is that the concept of conceptual model validity is not familiar to the respondents. Another explanation is that in the experiment paper, relatively little space is spent on the description of the conceptual model.

O5. The population of the case study [4] is the set of all GSE projects, and the population of the experiment [18] is the set of all academic SE students. Most respondents correctly stated that the population is described in the two papers but when explaining their answer in the open part of their answer to question U5, we found a variety of incorrect answers. Some thought that the set of interviews was the case study population. Some thought the sample of pairs investigated in the experiment was the population. These answers indicate that the concept of a population is not well-understood by all respondents.

One possible explanation of this observation is that the authors of the case study paper freely indicate that they want to generalize to all GSE projects; hence, it is easy to recognize that the population consists of all GSE projects. But the authors of the experiment paper carefully avoid generalizations and do not use the concept of population. Following the Goal Question Metric (GQM) method, they call the set of undergraduate computer science and SE students the *context* ([18], section 4.1).

O6. Research questions were clearly stated in the papers, and current knowledge was clearly summarized. Respondents correctly identified this in the papers.

Figure 5 shows the recall and precision of the seven questions formulated in this block. According to these two scatterplots, for the seven closed problem investigation questions (CPI) and their respective open questions (OPI), the subjects who were familiarized with the engineering cycle approach (five subjects from UT) showed a better understanding of these questions than other respondents (UPV, and Weingarten), who had more difficulties in reviewing the experiment paper. Only one respondent of this group (of five subjects) had the highest precision and recall.

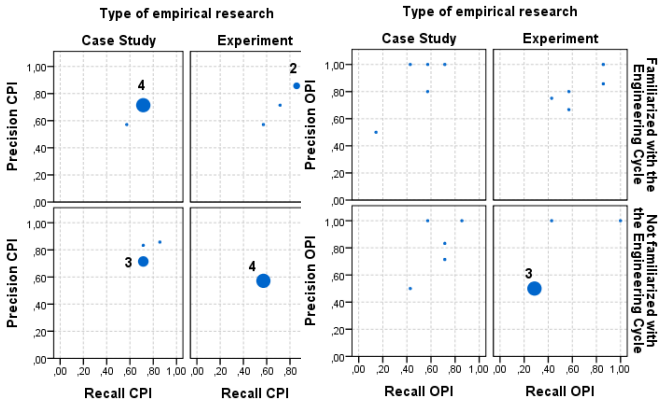

Figure 5. Binned scatterplots for the recall and precision of CPI and OPI questions.

We cannot generalize these answers to other uses of the checklist, with other responses or other papers. But these observations do indicate that the checklist needs to improve its explanation of the difference between improvement goals and knowledge goals, of the role of a conceptual model and what it means to motivate the validity of the model for the research purpose at hand, and of the concept of a population. More generally, in view of observations O1, and O2, an improved checklist should make clear whether or not checklist items ask for mandatory information.

7.1.2 Research design

An overview of the precision and recall of the nine questions of the research design block (RD) is shown in Figure 6. The five subjects familiarized with the approach had a good precision (>0,80) answering the open and closed questions but only for the case study paper. On the other hand four out of five subjects of the second group (not familiarized) had also a good recall and precision (>0,70) answering the closed questions only for the experiment paper. The precision of the respective open questions was acceptable (> 0,60).

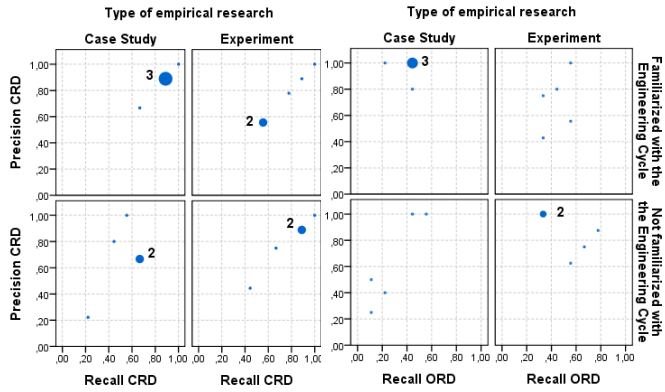

Figure 6. Binned scatterplots for the recall and precision of CRD and ORD questions.

By analyzing the questions of this block, we observed:

O7. The correct answer about the UoDC (U8) was a sample for the experiment paper and a set of cases for case study paper. However, most respondents stated that the UoDC in both papers is a model. In the reference guide of the checklist provided to them, we define a model as "an entity not in the population that is used as a representative for the units in the population." In the experiment paper the use of students as models of professional software engineers is described. Most respondents from the first group (who are experienced conducting only case studies)misunderstood the concept of a model. This may be related to their misunderstanding of the concept of a population.

O8. Most respondents found the information provided by the two papers about sample size and case structure

O9. For the experiment paper, all respondents agree that a treatment is specified, but their textual answers showed that they used a different concept of treatment as that provided in the reference guide of the checklist provided to

them(treatment was defined as an *intervention* or input the UoDC was exposed to). This is, however, ambiguous. In the experiment paper, the students had to fill out an on-line personality test to assess their level of neuroticism, and the experiment used blocking on the neuroticism variable to measure the effect of this variable on academic performance of the pair. Each pair was assigned programming tasks and academic performance of the students was measured. According to the reference guide, the treatment would be only the programming task. But a common view in the methodological literature is to define the treatment as a level of the independent variable [25]. Most of the respondents (7 out of 10) called the independent variable as 'the treatment', some of them identified the personality test as 'the treatment', and none of them identified the programming task as the treatment. Clearly, the explanation for the question U9.1 needs to be clarified.

O10. The sampling procedure was found in the experiment paper by most respondents, but very few respondents mentioned the treatment assignment procedure in their textual answers. This is explained by the previous observation: if the treatment is a personality characteristic, then this is not assigned but given to the experimenter.

O11. For the experiment paper, the question about the treatment plan (U9.3) was missed by most respondents, again a consequence of O9.

O12. In the experiment paper, nothing relevant needed to be reported about treatment instruments (U 9.4). About half of the respondents agreed that treatment instruments was not relevant for this experiment.

O13. About the question U10.1, measurement procedures were described explicitly in the case study paper, a7 out of 8 respondents were able to find this (2 respondents did not answer the open question). However, this question in the experiment paper was partially answered by the majority of the respondents. Some of them identified only the measurement of the independent variable (the personality test), others identified only the measurement of the dependent variable (measurement of academic performance). One possible explanation is that the experiment has a more complex structure than the case study, which makes information in the experiment paper harder to find for most respondents.

O14. For both papers, most respondents correctly identified measurement instruments (U10.2).

O15. Most respondents correctly identified the kind of reasoning (U11), i.e. grounded reasoning in the case study paper, and various statistical tests in the experiment paper.

To summarize, the concept of UoDC was not well-understood by most respondents, which may be related to the problem of understanding what a population is. However, they could find the information in the papers about the structure of the UoDC. There is an important misunderstanding about the concept of a treatment, which by some respondents was identified with the concept of level of an independent variable. Also, the treatment plan, treatment instruments, and measurement procedures were not easy to identify more for respondents without experience in conducting experiments. Thus, the understandability effectiveness for this set of questions seems to be influenced by the researcher experience in conducting a type of empirical research.

7.1.3 Research design validation

The validation questions in the checklist belong to several decisions in research design as identified in the checklist. Figure 7 shows that the recall and precision of the questions of this block. Although these values are acceptable (0,60) for the 3 out of 5 subjects familiarized with the engineering cycle, by reviewing the experiment paper, both measures are the lowest values answering their respective open questions (0,00). This confirms that for all subjects the 10 questions of this block were very hard to understand. So, the reference guide of the checklist requires to be improved, by providing some definitions, giving examples, etc.

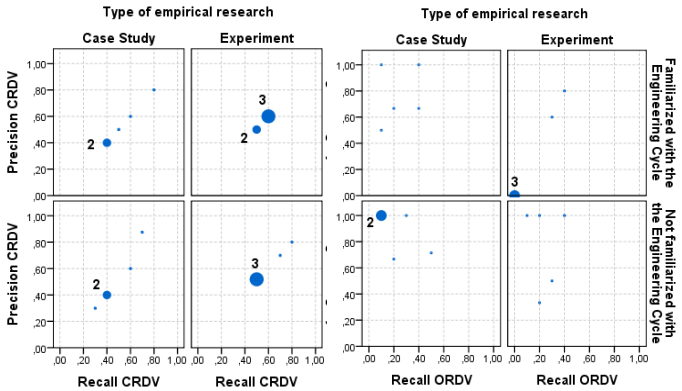

Figure 7. Binned scatterplots for the recall and precision of CRDV and ORDV questions.

O16. Considering that the case study paper discusses the representativeness of the studies cases with respect to the intended population (concerning the properties of geographical distribution and the diversity of the nationalities

involved), 9 out of 10 respondents could correctly answer the question U12.1. The experiment paper did not discuss the representativeness of the sampled students, possibly because no generalization is intended in this paper. However, half of our respondents by considering the size of the sample as enough to be able to draw conclusions about the population.

O17. The ethics of UoDC acquisition (U12.2) is about awareness and consent of participants. This was explicitly described in the experiment paper, and our respondents could identify it. It was not mentioned in the case study paper, and again our respondents were divided over "Described" and "Not described".

O18. We already observed that nothing relevant needed to be reported about treatment instruments in the experiment, and so there is no issue of treatment instrument validity. (An example of such an issue would be a confounding effect created by the instrument.) Our respondents correctly identified absence of discussion of treatment instrument validity (U13.1).

O19. External validity of the treatment, according to our reference answer for the question U13.2, is whether the programming tasks given to the pairs were representative of programming tasks assigned in the population of interest. Although this was not discussed in the paper [18], most respondents counted this as "Discussed". This can be explained by the fact that most of them did not identify the programming tasks as a treatment (O9).

O20. Validity of measurement procedures (U14.1) is the absence of influence of the measurement procedure on the measured value of a variable. This is discussed in the case study paper. Only 4 out of 10 respondents could find correctly this information in the paper. In the experiment paper, possible influence of measurement procedures on the measured values are not discussed, possibly because the authors thought it obvious that there is no such influence. 6 of our respondents answered correctly this question.

O21. The ethics of measurement is about confidentiality of data (U14.3). This is discussed in the case study paper. In the experiment paper, it is implicitly indicated by referring to an ethics board approval. Only 3 out of 10 respondents could find the information in the case study paper (7 of them answered this question incorrectly). However, although 6 of our respondents answering the closed question for the experiment paper, none of them answered the open question.

O22. Conclusion validity (U15.1) is about the validity of the reasoning from raw data to observations [21]. It is discussed in the experiment paper but not in the case study paper. Our respondents were divided in their answers, indicating a lack of shared understanding of what conclusion validity is. The textual answers given by some respondents indicated that they identified conclusion validity with internal validity, discussed next.

O23. Internal validity (U15.2) is about the validity of reasoning from claims about observed (co)variation to claims about what has caused this (co)variation [21]. The case study paper makes claims about the effect of some practices to manage trust in GSE projects; the experiment papers makes claims about the effect of neuroticism on academic performance in pair programming. Both papers justify these claims. Although 8 of our respondents answered correctly this closed question for the experiment paper, only four subjects could find this justification (3 out of 8 subjects did not justify their answer). With respect to the case study paper, none of the 10 subjects answered correctly this question.

Overall, looking at the number of times respondents answered the open questions of the checklist, they answered the questions about external validity of UoDC, internal validity and conclusion validity, but other validity questions, concerning the validity of instruments, procedures, and external validity of treatment were scarcely answered. Our explanation is that they are not familiar with these other validity issues. This is not surprising, as validity of instruments and procedures is rarely discussed in the literature. Respondents could identify the information about internal validity in the papers but some respondents confused conclusion validity with internal validity. The reference guide of the checklist should be improved by clarifying these concepts. Overall, about half of the respondents could find the information about ethics in the papers. Here too, clarification of the checklist is needed.

7.1.4 Research execution

Seven questions in the checklist follow the events during research execution, from acquisition of the UoDC through the application of the treatment to measurement, storing the data, and making the data available to others. If the research was executed exactly as planned, the author could state this explicitly. If something happened that the reader must know, the author should describe this.

O24. The two papers [18], [4] provide some information about what happened during the research, but few of our respondents could find this information. This is confirmed in Figure 8, a very low recall for the open questions (OEX) was obtained for the majority of all respondents.

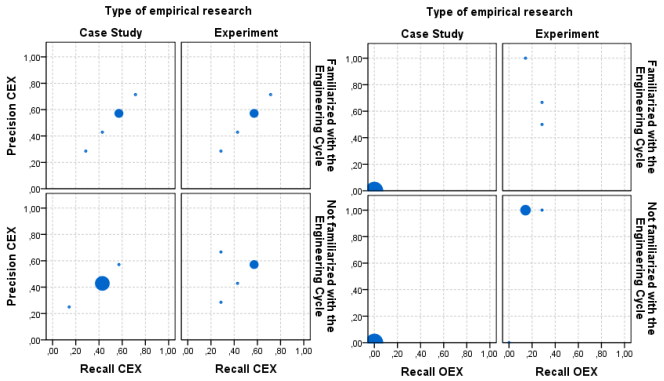

Figure 8. Binned scatterplots for the recall and precision of CEX and OEX questions.

7.1.5 Results evaluation

O25. Data from the case study would be anonymized interview fragments, data from the experiment would be anonymized individual scores on the personality tests or academic tests. This data is not included in the papers; rather the papers provide summaries in the form of recommendations given by interviewed project managers, and descriptive statistics of scores. Our respondents could find this information, but as we have seen earlier, some respondents did not correctly identify absence of information.

O26. Explanations are given to some extent in the case study paper [4]. For example, one recommendation given by interviewed project managers in the case study is to plan regular face-to-face or video meetings in advance. The paper describes the effects of this as predictability and ensured frequency of meetings. This goes some way into the direction of explaining the effect of planned regular meetings on trust, but it does not put it in a wider framework of theories about trust.

The experiment did not show an effect of neuroticism, but gives no possible explanations of this. One possible explanation is that the programming tasks were simple enough to avoid stress in the programming pairs; another possible explanation is that the level of neuroticism in the students as reported in the paper is not extremely high. A third possible explanation is that students made their decision to participate in the experiment based on their ability to handle their own and others' emotional instability. These explanations are hypothetical, and could be used to motivate further research. A repeated analysis of the experiment paper revealed to us that in the final paragraph, where future work is discussed, one possible fourth explanation is given: Coupling with a low-neuroticism partner could help a student with high neuroticism cope with his or her neuroticism.

Thus, in both papers, some explanations are present but not very explicitly, and alternative explanations are not explored. Most respondents said the experiment paper provided explanations, but they were divided over whether the case study paper provided explanations. Very few textual explanations were given and of those, only one was correct. We conclude that the checklist should clarify the concept of explanation and its place in a paper.

O27. Both papers answered their research questions, and many respondents could find these answers in the papers.

O28. The case study paper generalized its findings to all GSE projects, but the justification for this remains implicit in the earlier indication of external validity of the cases (O16). The experiment paper avoided explicit generalization. However, the experiment paper tested hypotheses, and in statistical hypothesis-testing, hypotheses are statements about a population. Most of our respondents could find the generalizations in the case study paper, but most said that generalizations were absent in the experiment paper.

O29. Limitations are, for example, the extent to which researchers have not been able to counter validity threats, and may also include important events during execution of the research, such as drop out of subjects, mistakes in measurements etc. Limitations are discussed explicitly in both papers. Our respondents could find this information.

O30. The case study paper does not explicitly identify contributions to the knowledge goal, but the experiment paper does. Our respondents correctly identified this absence resp. presence of information about contribution to knowledge goals in the papers.

O31. The case study gives its observations in the form of recommendations given by project managers, and so explicitly identifies a contribution to improvement goals. The experiment paper identified improvement of pair programming as its improvement goal ([18], section 4.1) but did not explicitly indicate its contribution to this goal. This may be because the results of the paper are negative, or preliminary. Our respondents' answers were divided, indicating that not all of them could identify the contributions listed in the case study paper, or ascertain the absence of such a description in the

experiment paper. At this point, one possible explanation is that the respondents were tired. As we will see in the next section, they complained that it took them too long to finish the checklist.

Overall, we conclude that not all respondents were aware of the distinction between observations and data. This should be explained better in the reference guide of our checklist. Moreover, according to the recall and precision values obtained for the questions of this block, we found that the respondents familiarized with the engineering cycle had a better understanding level than the respondents not familiarized with this approach.

We also found that explanations and generalizations have gotten very little space in both papers. The case study paper does generalize from its finding, but does not justify this explicitly.

7.2 Perceived ease of use and usefulness

To answer RQ2 and RQ3, the data collected from the questionnaire was analyzed. First, the scores of each subject were averaged over the different items that are relevant for a construct. We obtained two mean values (PEOU, PU) for each subject. Figure 9 illustrates the relationship between the mean values of these two variables.

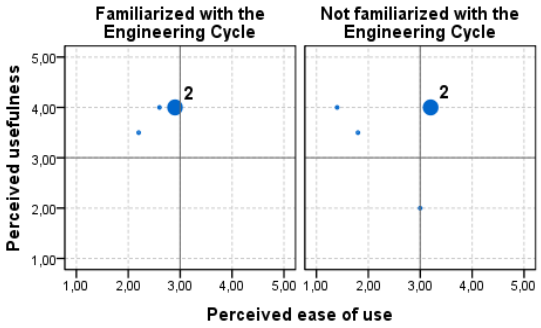

Figure 9. Scatterplots for the PU and PEOU by group.

Note that eight out of nine subjects (researchers familiarized or not with the engineering cycle) perceived the checklist as useful for reporting and designing (>3.00). However seven out of them perceived the checklist as difficult to use (<=3.00). They perceived the questions as not easy to understand and that the checklist took them too long to finish it. Moreover, 8 out of 9 respondents considering that a preliminary training would be necessary to use the checklist more efficiently. Table 1 shows descriptive statistics for these variables.

Table 1. Descriptive statistic for perception variables.

Statistic	PEOU	PU
Mean	2,58	3,67
Standard deviation	0,64	0,66

On the other hand, by analyzing the relationship of these variables (PEOU and PU) with the understandability level (UL[3]) obtained, when researchers reviewed the case study (Figure 10, left) and experiment paper (Figure 10, right) using the checklist, we observed that the only subject, who perceived the checklist as "not useful" (PU=2,00), had a good understanding of the checklist questions, but only for the experiment paper (0,76). According to our demographic questionnaire, this subject (a post doc researcher) answered to have experience conducting only experiments and using other checklists.

[3] It was calculated by the harmonic mean of the precision and recall measures of the 41 closed questions of the checklist.

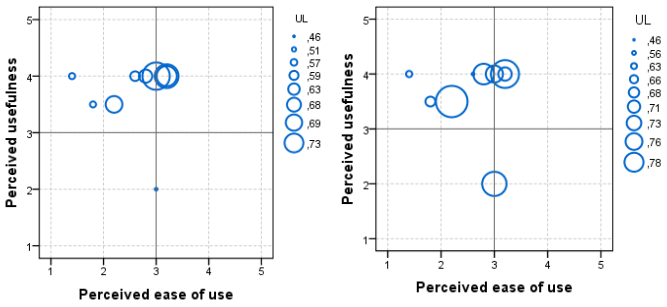

Figure 10. Bubble plots for the PU, PEOU and UL, reviewing the case study (left) and experiment paper (right).

We found these results very interesting for further refining of the checklist to improve its ease of use; more even knowing that researchers with low "empirical research experience" appreciate its usefulness for reporting and designing experiments or case studies.

8. DISCUSSION AND CONCLUSIONS

We have already discussed the limitations of our research design in section 5. During execution, an important limitation that became apparent is the length of time it took respondents to apply the checklists. The number of answers to the open part of the checklist questions rapidly decreases when reading from the first to the last questions of the checklist. We interpret this as increasing fatigue.

As indicated earlier, the size of our sample is small and our observations are approximate. The composition of the sample from PhD students in SE gives some basis for a weak generalization, namely that in *some* replications of this experiments with PhD students, we will make similar observations. As a knowledge result, this is very weak, but as a suggestion for improvement of the checklist, this is strong enough.

We here list our most important improvement suggestions for the checklist:

- The checklist should be reduced in size.
- Its use for different purposes (research design, research execution, research reporting, research evaluation, research replication) should be clearly indicated. Not all questions are relevant for all purposes.
- Optional questions should be clearly marked as such.
- In the questions about research problem investigation, the concept of population should be clarified. This should be related to an improvement in the explanation of UoDC and of the relation between UoDC and population.
- Validity concepts as currently used in the checklist are much more detailed than what is customary in the literature. Validity of a conceptual model, measurement procedure, treatment instrument, are not familiar. This needs to be better explained in the next version of the checklist.
- The concept of a treatment needs to be clarified. We speculated that this can be done by being more precise about the structure of the experimental set up.
- The checklist items about research execution can be reduced to a list of optional items.
- The distinction between data and observations was not clear to all our respondents and this must be clarified in the next version of the checklist.
- Explanation and generalization are not very prominent in the papers. This is consistent with earlier observations about the small role of theory in SE experiments [2], [24]. We conclude that at least the checklist needs to clarify the nature and importance of these items.

With the purpose of providing a checklist that are understandable, easy-to-use, and useful to design experiments or case studies, we are currently planning a second evaluation with an improved checklist version. The subjects will again be senior and novice researchers.

Finally, we emphasize that a unified checklist can be viewed as a shared ontology of empirical SE concepts. As our experimental evaluation shows, there is no universal agreement about basic concepts of empirical SE, and continuous debate and improvement of a unified checklist can help to provide such a commitment to a shared ontology. We hope that can contribute to the improvement of the quality of research designs and research reports in the empirical SE community.

9. ACKNOWLEDGMENTS

This work has been developed with the support of Intra European Marie Curie Fellowship Grant 50911302 PIEF-2010. We also acknowledge to all the participants of this experiment.

10. REFERENCES

[1] Anderson, J.R. (2010): Cognitive Psychology and its Implications. 7th ed., New York, Worth.

[2] Hannay J. E., Sjoberg D. I. K, and Dyba T. A Systematic Review of Theory Use in Software Engineering Experiments. IEEE Trans. Softw. Eng. 33(2) 2007), pp. 87-107.

[3] Host M. and Runeson P. Checklists for Software Engineering Case Study Research. IEEE ESEM'07 pp. 479-481.

[4] Jalali S., Gencel C., and Šmite D. Trust dynamics in global software engineering. ACM-IEEE ESEM '10. Article 23.

[5] Jedlitschka A., Pfahl D., "Reporting guidelines for controlled experiments in software engineering," IEEE ISESE 2005, pp.94-104.

[6] Kitchenham B. A., Pfleeger Sh. L., Pickard L. M., Jones P. W., Hoaglin D. C., El Emam K., and Rosenberg J.. Preliminary guidelines for empirical research in software engineering. IEEE Trans. Softw. Eng. 28(8), 2002, pp. 721-734.

[7] Kitchenham B., Al-Khilidar H., Ali Babar M., Berry M., Cox K., Keung J., Kurniawati F., Staples M., Zhang H., and Zhu L. Evaluating guidelines for reporting empirical software engineering studies. Empirical Softw. Eng 13(1), 2008), pp. 97-121

[8] Kitchenham B., Brereton O., Budgen D., Li Z."An Evaluation of Quality Checklist Proposals - A participant-observer case study". EASE 2010.

[9] Kitchenham B., Sjoeberg D., Brereton O. P., Budgen D., Dyba T., Höst M., Pfahl D., Runeson P., Can we evaluate the quality of software engineering experiments?. In: Proc. of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, 2010.

[10] Martz W., "Evaluating organizational effectiveness". Dissertation Abstracts International, 69 (07), Publication No. ATT3323530, 2008.

[11] Martz W., "Validating an evaluation checklist using a mixed method design". in: Evaluation and Program Planning, Volume 33, pp. 215-222, 2010.

[12] Moher D., Hopewell S., Schulz K., Montori V., Gtzsche P., Devereaux P., Elbourne D., Egger M., and Altman D., "CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trial". BMJ, 2010pp. 340-869.

[13] Moody D.L., The Method Evaluation Model: a theoretical model for validating information systems design methods. 11th ECIS 2003, pp. 16-21.

[14] Nunally J.., Psychometric Theory, McGraw-Hill, 2nd ed., New York, NY1978.

[15] Pfleeger S.L.. Experimental design and analysis in software engineering. Annals of Soft Eng, 1(1) 1995, pp. 219-253

[16] Raghavan V., Bollmann P., and Gwang S. J.. A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Inf. Syst. 7(3) 1989.

[17] Runeson P. and Höst M. 2009. Guidelines for conducting and reporting case study research in software engineering. Empirical Softw. Engg. 14(2), 2009, pp. 131-164.

[18] Salleh N., Mendes E., Grundy J., and Giles St. Burch J. 2010. The effects of neuroticism on pair programming: an empirical study in the higher education context. ACM-IEEE ESEM 2010. Article 22, 10 pages.

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

[20] Scriven M., Checklists. In: S. Mathison (Ed.), Encyclopedia of evaluation, Sage, Thousand Oaks, 2005. pp. 53-59,

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

[22] Wieringa R. J. Design science as nested problem solving. ACM 4th DESRIST, 2009 , pp. 1–12.

[23] Wieringa, R.J. (2012) A Unified Checklist for Observational and Experimental Research in Software Engineering (Version 1). TR-CTIT-12-07, CTIT, UT, Enschede. ISSN 1381-3625

[24] Wieringa, R.J. and Daneva, M. and Condori-Fernández, N. (2011) The Structure of Design Theories, and an Analysis of their Use in Software Engineering Experiments. IEEE ESEM 2011, pp. 295-304.

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