Practical Relevance of Experiments in Comprehensibility of Requirements Specifications

Nelly Condori-Fernández, Maya Daneva, Klaas Sikkel
University of Twente
Drienerlolaan 5, 7522 NB Enschede, The Netherlands
{n.condorifernandez, m.daneva, ksikkel}@utwente.nl

Andrea Herrmann
Technical University Braunschweig
Mühlenfordstr. 23
Braunschweig, Germany
AndreaHerrmann3@gmx.de

Abstract—Recently, the Requirements Engineering (RE) community has become increasingly aware of the importance of carrying out industry-relevant research. Researchers and practitioners should be able to evaluate the relevance of their empirical research to increase the likely adoption of RE methods in software industry. It is in this perspective that we evaluate the 24 experimental studies on comprehensibility of software requirements specifications to determine their practical value. To that end a checklist based on Kitchenham’s approach was operationalized from a practitioner’s perspective and an analysis with respect to the main factors that affecting on comprehensibility was carried out. Although 100% of the papers reviewed reported statistically significant results, and 96% of them take examples from a real-life project. 80% of the papers do not scale to real life, 54% of the papers do not specify the context in which the results are expected to be useful. We also found that there is a lack of underlying theory in the formulation of comprehensibility questions.

Keywords—practitioner’s checklist; comprehensibility; requirements specifications; experiments

I. INTRODUCTION

The Evidence Based Software Engineering (EBSE) paradigm aims to improve decision-making related to software development and maintenance by integrating current best evidence from research with practical experience and human values [8]. Evidence is defined as a synthesis of the best quality of empirical studies on a specific topic or research question. However, an important problem acknowledged by the supporters of EBSE is that there is not yet any consensus on the definition of “quality”.

In the Cochrane handbook [12], the best known reference in EBSE, quality is defined as the extent to which the study minimizes bias and maximizes internal and external validity. However, we consider that construct validity is also a particularly important quality issue to be considered, since higher quality studies are only possible if we use constructs that we understand well and are able to communicate precisely.

Dybå et al. [9] identified three key aspects to assess the quality of primary studies in SE: (i) rigorousness (Has a thorough and appropriate approach been applied to key research methods in the study?), (ii) credibility (Are the findings well-presented and meaningful?), and (iii) relevance (How useful are the findings to the software industry and the research community?).

In our study on comprehensibility of software requirement specifications (SRS), we focus on one of the criteria identified by Dybå et al., namely the relevance. We analyze 24 studies that report on experimental results on comprehensibility of SRS. The 24 sources were identified in a previous systematic mapping study [2], in which comprehensibility was identified as the quality attribute that is studied most by RE researchers.

A secondary purpose of the paper, in addition to the result of the study on comprehensibility, is to contribute to the EBSE methodology. We used a practitioner’s perspective-based checklist [17], which was inspired by checklist-based review from multiple perspectives proposed by Kitchenham et al. [5].

Comprehensibility, also called understandability, can be defined as the degree to which information contained in a SRS can be easily understood by a reader of that SRS [22]. It is an essential quality attribute of SRS. For example Krogstie [18] considers comprehension as the only goal of pragmatic quality1 and the comprehensibility (understandability) as an important factor in achieving the empirical quality.

Comprehensibility is the quality attribute that has most frequently been empirically studied, yet Shima et al. [19] warn that it is difficult to evaluate it in practice. Comprehension is an internal human process, which requires a research that is more multidisciplinary in nature (e.g. by using cognitive theories). Motivated by Shima’s affirmation, we asked ourselves how the empirical evaluations on comprehensibility have been carried out so far. If we evaluate how relevant the empirical results on comprehensibility, published until 2008 are for the software industry, then this contribution will allow us to know what kinds of empirical studies could be useful to the software industry and the research community.

The remainder of this paper is organized as follows: Section II. discusses an analysis of the 24 studies with respect to the main factors that affecting on comprehensibility. Section III summarizes the practitioner’s checklist used and the reliability analysis that we carried out. Section IV. reports on our results for each of the checklist questions analyzed. Finally conclusions and further work are presented.

1 It relates to the effect that the model has on the participants and the world.
II. EXPERIMENTS REPORTED ON COMPREHENSIBILITY OF REQUIREMENTS SPECIFICATIONS

An important feature of a good RE process is its ability to facilitate effective communication of requirements among different stakeholders, who, more often than not, have only limited background in computing [1]. Thus, requirements notations and processes must help RE professionals to maintain a delicate balance between the suitability of requirements descriptions for a non-computing audience and the level of precision needed for the downstream developers.

Numerous SRS techniques have been proposed with the purpose of improving the quality of requirements documents and consequently increasing user satisfaction with the final product [2],[16].

This paper focuses on the evaluation of 24 articles reporting experiments on comprehensibility of SRS, which were selected out of 46 articles identified by means of a Mapping Study realized by [2]. The comprehensibility according to this mapping study was one of the aspects most studied in the RE literature from 1976–2008. The references of these studies are in the Appendix of this paper.

A description of these 24 studies in terms of the main factors that affect the comprehensibility of requirements specifications is presented in Table I. These factors are: type of task, language expertise, domain expertise, and problem size [15].

**Type of task:** The different tasks that readers perform with a representation are facilitated or hindered to varied degrees. For example, comprehensibility for information search or for information retention can lead to different evaluation results.

**Language expertise:** The user’s previous expertise with the modeling language under study.

**Domain expertise:** The user’s previous expertise with the domain being modeled.

**Problem size:** Size of the specification. Different modeling languages scale up with variable degrees of success.

We list also the objects of study, the object being analyzed in each of the 24 empirical studies. This can be a language, method, technique (guidelines, diagrams, etc.), or tool for SRS (See Table I).

<table>
<thead>
<tr>
<th>ID</th>
<th>Object of Study</th>
<th>Type of Task</th>
<th>Language expertise</th>
<th>Domain expertise</th>
<th>Problem size</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Stereotypes of UML sequence diagrams</td>
<td>Answering questions concerning the semantic of the diagrams, and describing the functionality of diagrams.</td>
<td>Students had knowledge about the use of stereotypes in general. They were taught about the specific stereotypes used</td>
<td>Not indicated</td>
<td>Not indicated</td>
</tr>
<tr>
<td>S2</td>
<td>UML statecharts and/or sequence diagrams</td>
<td>Answering questions concerned several aspects of the system requirements.</td>
<td>Students had received classes on the modelling of OO software systems using UML</td>
<td>Familiarized with the domain of both systems used</td>
<td>Not indicated</td>
</tr>
<tr>
<td>S3</td>
<td>Black-box (SCR) and White-box (UML) requirements specifications</td>
<td>Answering questions about the semantics of specifications</td>
<td>Training session of 3 hours</td>
<td>Familiarized with the domain</td>
<td>Black box specification: 12 pages White box specification: 20 pages</td>
</tr>
<tr>
<td>S4</td>
<td>Use Case Models</td>
<td>Answering questions about the Use case read</td>
<td>A half-day seminar on use cases.</td>
<td>Not indicated</td>
<td>Not indicated</td>
</tr>
<tr>
<td>S5</td>
<td>Multimedia requirements specification</td>
<td>Answering questions about the semantics of specifications</td>
<td>Training period is not indicated</td>
<td>Not indicated</td>
<td>Not indicated</td>
</tr>
<tr>
<td>S6</td>
<td>Object oriented models and process oriented models</td>
<td>Answering questions based on the model provided.</td>
<td>Students had some prior experience with process-oriented modeling;</td>
<td>Not indicated</td>
<td>The problem size is not indicated</td>
</tr>
<tr>
<td>S7</td>
<td>Use Case Models</td>
<td>Answering questions about the functionality specified in the use case model with different guidelines.</td>
<td>None of the students were familiarized with the guidelines. Training session of two hours for basic use case modeling, and one hour for the use of guidelines.</td>
<td>Few subjects with application domain experience.</td>
<td>Use case model consisted of 7 actors and 9 use cases.</td>
</tr>
<tr>
<td>S8</td>
<td>Data flow diagrams and textual specification.</td>
<td>Identification of the functions that are performed by the system and answering questions about the problem-domain scenarios.</td>
<td>Programmers were not familiarized with data flow diagram</td>
<td>Familiarized with the problem domain</td>
<td>“Toy problems”</td>
</tr>
<tr>
<td>S9</td>
<td>saUML sequence diagrams</td>
<td>Answering questions formulated with two levels of abstraction: comprehensibility and application.</td>
<td>Students were mostly Java juniors with basic concurrency concepts.</td>
<td>Students receive a brief introduction to the readers/writers problem.</td>
<td>Not indicated</td>
</tr>
<tr>
<td>S10</td>
<td>UML sequence,</td>
<td>Answering questions about</td>
<td>Students received a course</td>
<td>Not indicated</td>
<td>Not indicated</td>
</tr>
</tbody>
</table>
As shown in Figure 1, UML diagrams were the most studied (e.g. use case, sequence, collaboration, state, object diagrams) by the researchers. 17% of the studies focused on the analysis of the effect of stereotypes on comprehensibility of requirements specifications. Comparisons between UML and non-formal languages has received slightly more attention than comparisons with formal languages.
With respect to the type of task performed, 87.5% of the studies externalized the comprehensibility by means of questions formulated about the semantic of specifications. However, quality of these questions is not discussed by the majority of the researchers. Only four studies (S1, S9, S12, S14) considered an underlying theory in the formulation of comprehensibility questions. For example, in S9, according to Bloom’s taxonomy[20], the questions were formulated at two levels of abstraction: comprehensibility and application.

The following three factors are very strongly related to the external validity\(^2\).

With respect to language expertise, only 1 of out 24 studies involved professionals experienced with UML and familiarized with the usage of stereotypes (S24). The rest of studies had to conduct a training process to teach the modeling language under study. The training period observed in these studies was very varied. It oscillates from 45 minutes until 5 weeks as maximum. Furthermore, only 1 study (S8) did not require that the subjects (Programmers) were familiarized with the modelling language (data flow diagram).

With respect to domain expertise, 50% of the studies do not mention any level of the subjects’ familiarization with the problem domain. S7 mentioned that few subjects had some experience with the domain under study, which were randomly assigned to the treatments with the purpose of blocking the influence of this variable (domain) on comprehensibility. Only S11 analyzed this variable for evaluating the difference in diagrams comprehension for the Management Information Systems (MIS) and Real Time domains.

Considering the combination of these two factors (language and domain expertise) we have three main potential types of subjects, summarized as follow:

- A subject has less knowledge on the modeling language than knowledge of the domain (e.g. clients);
- A subject has an equal knowledge on the modeling language and knowledge of the domain (e.g. requirements engineers); and
- A subject has more knowledge on the modeling language than knowledge of the domain (e.g. analysts).

Although the 24 studies involved students as subjects, the majority of them evaluated the SRS comprehensibility from “analyst” perspective. Training sessions were more focused on the modelling language than the problem domain.

With respect to the problem size, we note a diversity of units of measure to express the problem size (e.g. number of elements of model, number of lines, number of pages). The majority of these studies (62%) recognized the lack of external validity due to use of “toy problems”. Only 1 study (S17) affirmed that the problem size was “representative” of a real-life case. However, the “representative” term was not expressed in a quantitative way. On the other hand 38% of studies described the problem but its size is not mentioned.

III. CHECKLIST-BASED EVALUATION ON RELEVANCE OF EXPERIMENTAL STUDIES ON COMPREHENSIBILITY

The review process included two practitioners, the second and the forth authors of this paper. Prior to becoming university researchers, they worked as RE consultants. The first has 10 years of RE experience in industry and the second 8 years. Both are active members of the RE community in their areas of experience. The two practitioners worked in two different locations, used a checklist independently from each other, and had no communication between them. Next, the checklist used to evaluate the relevance of these 24 studies is described.

We make the note that using two practitioners only is not enough to obtain generalizable results, and that including more practitioners is important. However, we thought that the learning experience in the evaluation can still be valuable because of its potential to open up new questions for discussion and reflection on our future work.

A. Practitioner’s checklist

Based on the approach of Kitchenham [5], a checklist was designed to be used in the RE community. It consists of 19 items formulated as questions concerning the information required for the practitioner’s perspective [17]. A practitioner is someone who provides summary information for use in industry and wants to know whether the results in a paper are likely to be of value to his/her company or clients. Each question is rated using a 3-point ordinal scale (“2” when a question can be answered as “yes”, “1” when a question can be answered “partially”, and “0” when the answer is “no”).

We had two raters who answered the 19 questions in the checklist while reading all 24 papers. Calculating the inter-rater reliability [13] of the questions of the checklist, 9 of out 19 questions showed an “acceptable agreement” (\(\kappa_{w} > 0.21\)) [17]. Based on this we discarded 10 questions because the agreement level among the answers of the reviewers was too low.

\(^2\) It refers to the approximate truth of conclusions to generalize the results of our experiment to industrial practice [21].
Table II shows the 9 questions that were used in this study, as well as the respective rationale for each question.

### TABLE II. QUESTIONS OF THE PRACTITIONER’ CHECKLIST

<table>
<thead>
<tr>
<th>Item</th>
<th>Question</th>
<th>Rationale/ Consultants need</th>
<th>Inter-rater Reliability [13]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Is the claim supported by believable evidence?</td>
<td>To be sure that any claims are supported by evidence</td>
<td>1</td>
</tr>
<tr>
<td>Q2</td>
<td>Is it claimed that the results are industry-relevant?</td>
<td>To clearly see whether the conclusions/results have practical relevance and why the authors think so</td>
<td>0.476</td>
</tr>
<tr>
<td>Q3</td>
<td>How can the results be used in practice?</td>
<td>Guidance by the authors on how the results would be used in industry</td>
<td>0.427</td>
</tr>
<tr>
<td>Q4</td>
<td>Is the result/claim useful/ relevant for a specific context?</td>
<td>To know the context in which the results are expected to be useful</td>
<td>0.474</td>
</tr>
<tr>
<td>Q5</td>
<td>Is it explicitly stated what the implications of the results/conclusion for practice are?</td>
<td>To get explicit information on the implications of the authors' work for practice</td>
<td>0.560</td>
</tr>
<tr>
<td>Q6</td>
<td>Are the results of the paper viable in the light of existing research topics and trends?</td>
<td>To know how the current work in the paper relates to current research trends.</td>
<td>0.660</td>
</tr>
<tr>
<td>Q7</td>
<td>Is the application type specified?</td>
<td>To know what type of applications the results apply to. In particular whether the results are specific to particular types of application (e.g. finance, or command and control etc.)</td>
<td>0.220</td>
</tr>
<tr>
<td>Q8</td>
<td>Do the authors show that the results scale to real life?</td>
<td>To be sure that the results scale to real life</td>
<td>0.514</td>
</tr>
<tr>
<td>Q9</td>
<td>Is the experiment based on concrete examples of use/application or only theoretical models?</td>
<td>To be sure that the results have a clear practical application.</td>
<td>0.9</td>
</tr>
</tbody>
</table>

**B. Validity evaluation of the results obtained in using the checklist**

Conclusion validity. Although our evaluation is based on the questions in the checklist that had an acceptable level of inter-rater reliability, there is a threat that the number of questions used cannot be enough for evaluating the relevance of the empirical studies from practitioner perspective. To reduce this threat our evaluation was complemented with the analysis of each one of the 24 studies in terms of the variables that affecting the comprehensibility (problem size, type of task, user expertise respect to the domain and modeling language) which are strongly related with the external validity.

Construct validity. Evaluator expectancies can bias the results of a study unconsciously, more even if we include feedback of two practitioners only. However, including more evaluators – by this we mean engaging more RE-practitioners from companies, was infeasible. We did have access to practitioners interested in the study, however there were resource constraints in their organizations that companies that impeded their participation.

We know that bringing more practitioners as evaluators it’s better. However, we also note that our evaluations (see the next Section) brought some strong conclusions (e.g. that 80% of the reported studies do not scale up to real life, see Q8 and Figure 5 on the next page). Therefore, given the strength of the conclusions, in this case we think that more precision would not add much more value because bringing in a 3rd and 4th evaluator would not change much the results.

**IV. RESULTS OBTAINED FROM CHECKLIST**

Next, we discuss each one of the questions that had an acceptable agreement for the 24 empirical studies reported.

**Q1. Is the claim supported by believable evidence?**

This question asks whether the reader can buy into it. If pieces of evidence build upon each other, the reader is likely to deem the evidence worthy of being accepted as true or reasonable. In our findings, the evaluators had 100% agreement regarding believability of the papers that were reviewed.

**Q2. Is it claimed that the results are industry-relevant?**

This question checks (i) whether the research was motivated by some real-life problem experienced in industry, e.g. in software projects, in carrying our certain tasks by professional requirements engineers, (ii) whether the conclusions relate to what was claimed in the motivation regarding the practical problems, and (iii) what analytical argument the authors use to convince the readers that their conclusion have relevance. We found that 58% of the studies, from a practitioner’ viewpoint, reported results not relevant for industry (See Figure 2)

**Q3. How can the results be used in practice?**

This question checks whether the authors clearly indicate how the results can be used in a real-life project. We found that 62% of the 24 articles propose a plan on how can be the results used in practice. (See Figure 2).
Q4. Is the result/claim useful/relevant for a specific context?

This question refers to the extent to which authors characterize the context where their proposals should work best. Context details are instrumental for practitioners to make a judgement about whether or not the proposal is relevant to their specific work settings. As shown in Figure 3, 54% of the articles do not specify the context in which the results are expected to be useful.

Q5. Is it explicitly stated what the implications of the results/conclusion for practice are?

This question checks whether or not the authors include a section that discusses the implications of their research for practice. We found that 67% of the articles reviewed did not address the implications of their results for practice (See Figure 3).

Q6. Are the results of the paper viable in the light of existing research topics and trends?

This question is to judge how the results are positioned with respect to other findings known in literature on the same topic (comprehensibility of SRS) with respect to the same artifact (e.g. an UML model, a statechart). Usually, the answers to this question are in the Related Work section, or in the Discussion section of a paper. As shown in Figure 4, 62% of the articles discussed their contribution with respect other findings on similar topic.

Q7. Is the application type specified?

This question checks whether or not the authors give enough details about the application type and the implications of this for their research. As shown in Figure 4, 50% of the articles specified partially the type of application to be used.

Q8. Do the authors show that the results scale to real life?

This question checks whether or not the authors discuss the assumptions they make about using their proposal in real life projects. We found that 80% of the article do not scale up to real life (See Figure 5).

Q9. Is the experiment based on concrete examples of use/application or only theoretical models?

This question asks whether the authors take examples from a real-life project as a starting point in their empirical work or they deliberately take a simplified example used in text books, or in their own teaching (theoretical models). As shown in Figure 5, all articles except one take examples from a real-life project.

V. CONCLUSIONS AND FUTURE WORK

In this paper we report on a preliminary evaluation of the relevance of results in empirical studies on comprehensibility of requirements specifications. This evaluation was carried out using a checklist that was adapted from the Kitchenham approach [5].

With respect to our primary research goal, we found that a majority of experiments reported have significant limitations with respect to the artifacts and subjects utilized, which does not facilitate the application of research results to industry (generalizability of results). This clearly indicates the need for more research to evaluate SRS techniques in real-life settings.

On the other hand, we found that 88% of the studies externalized the comprehensibility by means of questions about the semantic of specifications. However, there is a lack of underlying theory in the formulation of these comprehensibility questions. This means that we could, and should, bring the insights of other research areas to our empirical evaluations on
comprehensibility (e.g. cognitive psychology). This lack of underlying theory affects unfavourably the construct validity, which is also related to generalizing the results of the experiment.

With respect to our secondary goal, we found that while the checklist is in no way perfect, we found it to fit the purpose of our study and yet be inexpensive to use. This encourages us using in follow up research, for example, in evaluating the relevance of empirical studies on comprehensibility of SRS published between after 2008. This will both complement the results we have now regarding the studies themselves and regarding the checklist.

We also welcome RE experiments with more relevance to reality. If we as a community generate and continually collect experience reports that relate to the challenges and the possible solutions of RE practitioners in their day to day work, our knowledge about how to leverage our research output to software project or process success would increase.

We plan to carry out a second and much larger evaluation of the relevance of empirical studies in RE, once the checklist is validated. To do this validation, we are going to adjust the rest of questions of the checklist [17], and to evaluate the effectiveness of the checklist by means of independent researchers/practitioners. For this purpose, we consider those practitioners from companies in the Netherlands who already participate in industry-university projects with the University of Twente. We also wanted to compare our experiences with other researchers/practitioners that could possibly use the checklists. We therefore invite other researchers from the empirical SE and RE communities to share ideas about how we could improve both our checklist and our checklist-based evaluation approach.

VI. ACKNOWLEDGEMENT

This work was in part funded by the EU Marie Curie Fellowship Grant 50911302 PIEF-2010.

APPENDIX : STUDIES INCLUDED


S23. Lange Ch., Chaudron M., Effects of Defects in UML Models – An Experimental Investigation. ICSE’06, May 20–28, 2006, Shanghai, China Copyright 2006 ACM.

REFERENCES


