Experiences in Using Practitioner's Checklists to Evaluate the Industrial Relevance of Requirements Engineering Experiments

Maya Daneva, Klaas Sikkel
Universuty of Twente
The Netherlands
{m.daneva,k.sikkel}@utwente.nl

Nelly Condori-Fernandez
University of A Coruña
Spain
n.condori.fernandez@udc.es

Andrea Herrmann
Herrmann & Ehrlich
Germany
herrmann@herrmann-ehrlich.de

ABSTRACT

Background: A grand challenge for Requirement Engineering (RE) research is to help practitioners understand which RE methods work in what contexts and why. RE researchers recognize that for an RE method to be adopted in industry, RE practitioners should be able to evaluate the relevance of empirical studies to their practice. One possible approach to relevance evaluation is the set of perspective-based checklists proposed by Kitchenham et al. Specifically, the checklist from the practitioner's perspective seems to be a good candidate for evaluating the relevance of RE studies to RE practice. However, little is known about the applicability of this checklist to the RE field. Moreover, this checklist also requires a deeper analysis of its reliability. Aim: We propose a perspective-based checklist to the RE community that allows evaluating the relevance of experimental studies in RE from the practitioner's/consultant's viewpoint. Method: We followed an iterative design-science based approach in which we first analyzed the problems with a previously published checklist and then developed an operationalized proposal for a new checklist to counter these problems. We performed a reliability evaluation of this new checklist by having two practitioners apply the checklist on 24 papers that report experimental results on software requirements specifications' comprehensibility. Results: We report first-hand experiences of practitioners in evaluating the relevance of primary studies in RE, by using a perspective-based checklist. With respect to the reliability of the adjusted checklist, 9 of out 19 questions show an acceptable proportion of agreement (between two practitioners). Conclusions: Based on our experience, the contextualization and operationalization of a perspective-based checklist helps to make it more useful for the practitioners. However, to increase the reliability of the checklist, more reviewers and more discussion cycles are necessary.

*Produces the permission block, and copyright information
†The full version of the author's guide is available as acmart.pdf document
[1]It is a datatype.

WOODSTOCK'97, July 2016, El Paso, Texas USA

DOI: 10.1145/123_4

CCS CONCEPTS

• **Computer systems organization** → **Application-based systems**; • **Software**

KEYWORDS

Practitioner's checklist; reliability; validation; requirement specifications; experiments

1 INTRODUCTION

In the past 20 years, many calls have been made in the empirical software engineering (SE) community (e.g. [3, 6, 17, 22]) − and in the requirements engineering (RE) community in particular (e.g. [1, 4, 5, 23]), to increase the relevance and usefulness of empirical studies to practitioners and IT organizations. To empirical SE researchers, these calls cannot be translated simply into more reaching-out or "marketing" actions, in order to get practitioners value the lessons that researchers learn from their empirical work. The empirical SE community states (e.g. in [6, 22]) that relevance of empirical studies is contingent on the extent of practitioners involvement in the empirical research process and in the interpretation of the research outcomes. Results from empirical research in SE (and in the sub-area of RE, in particular) are considered more relevant if researchers and practitioners take each other's perspectives into account while creating and executing their empirical designs, and reporting their empirical results [17]. Drawing upon this understanding, a number of research methods (e.g. Grounded Theory, Focus Groups, Design Science, and Action Research) have been tried out by empirical SE researchers in order to actively engage practitioners in identifying industry-relevant research questions, problems worthwhile investigating, and evaluation criteria for the solution designs that researchers might come up with. Also, evidence-based SE methodologists [6] attempted formulating some reporting guidelines that empirical researchers could use in order to explicitly address questions about the practical implications of their empirical results. Moreover, such guidelines (in checklist format) are considered extremely useful in evaluating primary studies for inclusion in systematic literature review and mapping studies that pose research questions relevant to practice [6]. While the items in the checklist seem logical, little research however has been done so far, in order to evaluate the applicability of such a practitioners-perspective-oriented checklist to the various SE sub-fields, e.g. the area of RE. In this paper, we want to shed light into the extent to which this checklist is suitable for evaluating the relevance of empirical RE studies, to practitioners. As part of preparing this paper, we looked at the 40+ systematic reviews and mapping studies included in the editorial of the Journal of Systems and Software, on empirical RE [23] and the tertiary study of systematic reviews in RE of Bano et al. [24]. We found that no systematic review or mapping study among those cited in [23, 27], has used the practitioner's checklist of Kitchenham et al. [6] to evaluate the quality of the primary studies included in the respective secondary study. The fact that the applicability of the practitioner's checklist is not evaluated in real-life use in empirical RE clearly poses a problem: if we do not know what the practitioners' relevance of the primary studies is, how could we possibly judge the practitioners' relevance of the RE knowledge? The aim of this paper, therefore, is to make a first step towards the evaluation of the practitioner's checklist proposed in [6]. Specifically, we report on first-hand experiences of practitioners in evaluating the relevance of primary studies in RE from the practitioner's perspective. The experiences were gathered while carrying out a previously published mapping study [2] and a follow-up systematic literature review (SLR) on software requirements specification (SRS) techniques. To evaluate the relevance of the included primary studies from practitioners' perspective, we used the checklist-based approach in [6].

In the next sections, we provide a detailed account of our experiences: Section 2 presents the concept of quality of primary studies and summarizes the practitioner's checklist in [6]. Section 3 reports on our research approach to the application of this checklist and our results (problems we experienced and our modification of the checklist to counter these problems). Section 4 concludes the paper.

2 BACKGROUND

Evidence-based SE (EBSE) aims to improve decision-making related to software development and maintenance by integrating current best evidence[1] from research with practical experience and human values [9]. Although this "best evidence" is obtained from carrying out SLRs, more often than not there is no assessment of the practitioners' relevance of the primary studies to be aggregated. One reason for this is that there is not yet any consensus on the definition of "industry relevance as a quality attribute of empirical research" [9]. In the Cochrane handbook [13], quality is defined as the extent to which the study minimizes bias and maximizes internal and external validity. However, [13] suggests that higher quality studies are only possible if we use constructs that we understand well and are able to communicate precisely. Thus, construct validity is also a particularly important quality issue to be considered. Furthermore, Dyba et al. [10] identified a set of criteria related to rigorousness, credibility, and relevance of empirical studies to assess the quality of primary studies in SE. Furthermore, according to [11], the quality of the experimental study is also based on how it is reported, which means that the also depends on the extent to which it assists readers in (a) understanding how an experiment is conducted, (b) finding the information they are looking for, and (c) assessing the validity of the results [12].

As we can only assess an experimental study in terms of what can be inferred from what is reported, we used the assessment method proposed in [6], which was inspired by checklist-based reviews from multiple perspectives (researcher, practitioner, meta-analyst, replicator, reviewer and author). The practitioner's checklist includes 22 items formulated as questions concerning the information required for the practitioner's perspective. 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. In [6], each question is rated using a dichotomous scale ("1" when a question can be answered as "yes" and "0" when the answer is "no"). In addition, a set of rationale statements is also listed for each question.

Drawing on the work in [6], this paper describes our experience in using the practitioner's checklist [6] to evaluate a set of experiments on comprehensibility of requirements specifications.

3 APPLICATION OF THE CHECKLIST

The overall process of applying and evaluating the checklist is presented in Figure 1. It follows Wieringa's iterative design science based approach [17,18] to carrying out problem analysis and building up solution proposals. We chose this approach because of its fit to our research context, and also because it fosters active learning through experience, which could possibly lead to a more refined guidelines formulation (such as in [21]). The overall design science approach has been defined as one that helps "create things that serve human purposes or that help improve existing things to serve human purposes better" [18]. It uses the general structure of a rational problem-solving process [19]: analyze the current situation and current change goals, propose possible changes to meet those goals, evaluate possible changes and select one, apply the change and then start all over again. In the context of our research, this process was applied as follows: (1) we applied the existing checklist to a number of RE studies; (2) we identified difficulties with the checklist; (3) we suggested an improved checklist and (4) we applied the improved checklist to see if it works better in the area of RE.

The process included two practitioners. The first has 10 years of RE experience in industry and the second has 13 years. Both are active members of the RE community in their areas of experience. The practitioners got as input the 24 primary studies that were selected of out 46 studies reported in the mapping study [2], which was designed to identify what quality aspects of SRS are empirically evaluated by RE-researchers, in which context, and by using which research method. These 24 studies present experimental results focused on the comprehensibility of SRS

[1] Evidence is defined as a synthesis of the best quality scientific studies on a specific topic or research question **Error! Reference source not found.**.

only. The references of these studies are in the Appendix of this paper.

The two practitioners worked in two different locations, used the checklist independently from each other, and had no communication between them. When they exchanged their evaluation forms, it turned out that each one had a different interpretation of the checklist items and a different way to make a judgment on whether the item applied or not to each of the 24 primary studies. None of the authors of this paper expected such a huge variation in these respects and collectively they felt a need to modify the checklist so that it is fit for serving its purpose (namely to help obtain a transparent and reproducible evaluation). One of the practitioners (Daneva) worked with one researcher (Sikkel) on an on-going basis to come up with incremental improvements of the checklist. The second practitioner (Herrmann) served as a reviewer of each increment and ‘tested’ it to see whether it made sense or not.

When all authors reached a consensus on the items to be included in the checklist and their interpretations, the practitioners used the newly adjusted checklist and again independently from each other evaluated the quality of the 24 studies. The practitioners compared their results and whenever they disagreed, a discussion took place to resolve disagreements.

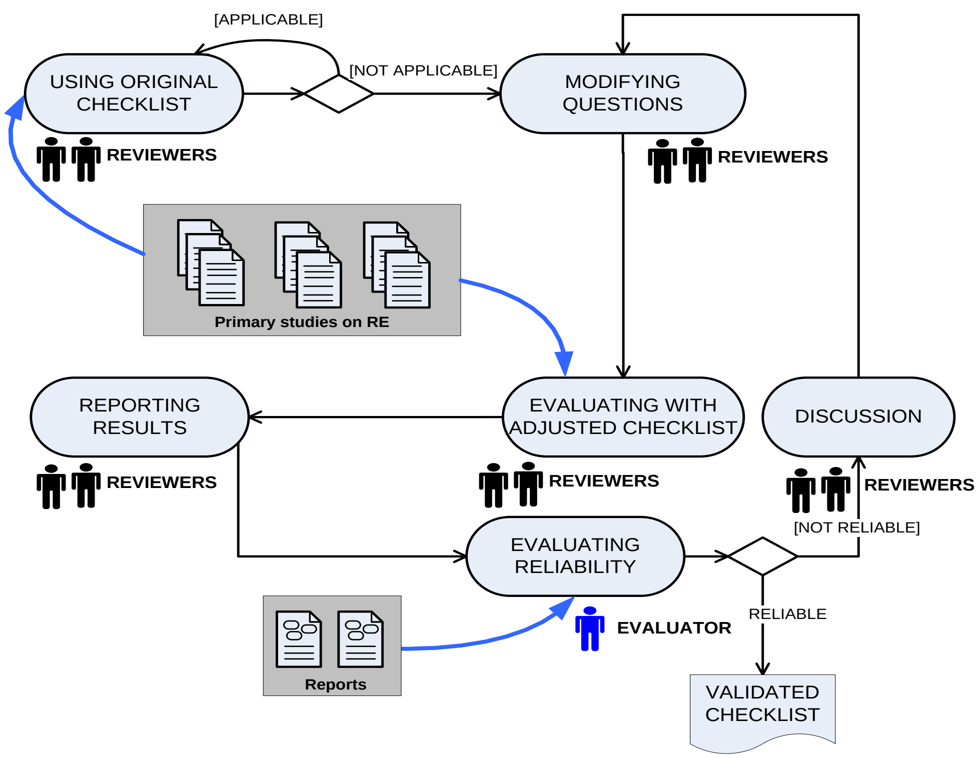

Figure 1: The process of checklist modification and evaluation.

3.1 Using the Original Checklist

An Excel form was created to include the original 22 items of the practitioner’s checklist in [6]. The form was used according to a predefined procedure. First, the practitioners became familiar with the checklist and verified whether it was (i) clear to understand and (ii) intuitive to follow. As these preliminary conditions were satisfied, the practitioners deemed unnecessary a face-to-face conversation to make sure they had a common understanding of the checklist. Instead, they decided to get together only if they experienced problems in using the checklist. Second, they performed their evaluations, by highlighting in each study the parts of the text that the practitioner deemed indicative for the presence or the absence of the information required to answer each question in the checklist. Each answer was based on the practitioner’s judgment. The practitioners also kept a log of problems and their reflections on these problems along with the ratings they could assign to a checklist item based on the information they found in the papers. Below we summarize the common observations (also problems) of the practitioners regarding the use of the checklist items:

- A consistent observation was that items P8-P19 in [6] could not discriminate between studies of various quality. These criteria mix the quality of the reporting with the quality of the analysis.
- Seventy-five percent of the items seemed to need revision for at least one of the following reasons: (i) an item implies a meaning that is clearly understood in SE

(e.g. using fully automated tools), but is unclear in RE; (ii) an item asks for a feature in the studies that is common for all RE studies (e.g. P14: "Is the experiment based on concrete examples of use/application or only theoretical models?"); (iii) an item seems to ask for a study's feature, which might be very important in other SE-sub-areas, while is unimportant in RE and therefore RE-authors have no incentives to add it (e.g. P17: "Does the paper make it clear who is funding the experiment and whether they have any vested interests?" We make the note that none of the 24 studies compared SRS based on commercial RE tools; so, no study was sponsored by a vendor of RE-tools.

Table 1: Modified Checklist

Item	Question	Rationale/ Consultants need
Q1	Is the claim supported by believable evidence?	To be sure that any claims are supported by evidence
Q2	Is it claimed that the results are industry-relevant?	To clearly see whether the conclusions/results have practical relevance and why the authors think so
Q3	How can the results be used in practice?	Guidance by the authors on how the results would be used in industry
Q4	Is the result/claim useful/ relevant for a specific context?	To know 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?	To get explicit information on the implications of the authors' work for practice
Q6	Are the results of the paper viable in the light of existing research topics and trends?	To know how the current work in the paper relates to current research trends.
Q7	Is the application type specified?	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.)
Q8	Do the authors show that the results scale to real life?	To be sure that the results scale to real life
Q9	Is the experiment based on concrete examples of use/application or only theoretical models?	To be sure that the results have a clear practical application.
	QUESTIONS for 'What-is-Better' Papers:	
Q10	Can a technique be used as-is?	To know whether the research is complete or the approach needs further research to become practically viable
Q11	Is the availability of required support environment clear?	To know if any required tool support is available and under what conditions
Q12	Are any technology pre-requisites specified?	To know about any technological prerequisites that might limit the applicability of the results
Q13	Are the experience or training costs required by RE staff and/or clients defined?	To know the training/experience requirements implicit in the approach
Q14	Are any risks associated with adoption defined?	To know the possible risks associated with adoption of the technique, if any
Q15	Does the paper discuss existing technologies, in particular the technologies it supersedes and the technologies it builds on?	To be sure that the experiment involves comparisons of appropriate technologies. They need to know that a new approach is better than other equivalent approaches
Q16	Is the new approach, technique, or technology well described?	To be sure that they understand the new approach/technique/technology well enough to be able to adopt it
Q17	Is the paper explicit about the possible biases?	To be sure that the experiment is as objective as possible
Q18	Does the paper make it clear what commitment is required to adopt the technology?	To know whether adoption of an approach/technology requires a complete and radical process change or can be introduced incrementally.
Q19	Are Technology Transfer issues discussed?	To know what the objections to a new technology are likely to be, and whether there are any clear motivators or de-motivators.

- We found it impossible for the 24 papers to meet even 50% of the items in the checklist. When one of the practitioners became aware of this, she contacted two authors (K. Cox and M. Ali Babar) of the checklist paper [6] to ask for suggestions on what could possibly represent a good cut-over percentage (this is the ratio of the number of met checklist items and the total number of items). The overall response was that 100%.
- One item was deemed irrelevant by the practitioners (P11: "Is the expense involved in adopting the approach (namely, the SRS approach) defined?"). This is because the research questions in the RE studies were of nature that precludes a discussion on the cost of adoption.
- One item was deemed 'a default feature of any good study in RE' (P22: Is there any discussion of required further research?). This item was found identical with

the item that asks about the possible implications of the study results for the future.

- The scale of 'yes'/'no' appeared to be insufficient to reflect the information content that a practitioner implied in her ratings. It turned out that a detailed operationalization was needed at checklist-item-level, so that a well-reasoned 'yes'/'no' backed up with facts, can come out.
- The 24 studies were of two types: (1) experiments to compare two SRS techniques to know which one is better in a given context (we call them 'what-is-better' papers); and (2) experiments that investigated the factors that affect the understandability of SRS. Eleven out of the original 22 items in the checklist turned out 'not applicable' for the studies of type (2). Finally, the practitioners discussed their observations with a senior researcher and a collective decision was made to adjust the checklist items in a way that seems to make more sense in the quality evaluation context of the RE studies. The modification and its justification is presented in the next section.

3.2 Modification and Justification

To act upon the observations presented in Sect III.A, we applied the following modifications to the original checklist:

(1) We excluded those questions that seemed irrelevant for the studies and were deemed 'a default feature' that is subsumed in another item.

(2) Each of those items that we deemed relevant was contextualized by defining its meaning in terms that a RE practitioner can understand and relate to. This was needed to reduce ambiguity and variation of interpretations.

(3) For each item, we also defined a scale, which is 'yes, 'partially', and 'no'. These fine-grained evaluation points were to allow for variation in quality of the analysis in a paper. The point 'partially' means, that (i) a study has no explicit answer to a question of the checklist but implicitly states it, or (ii) that the authors deemed the question irrelevant and mention a good reason for this.

(4) We restructured the checklist items in a way that allows the evaluator to clearly see those items that apply to studies which investigate SRS understandability factors (and do not compare RE-techniques to find which one is better). In Table 1, items Q1-Q9 refer to this study type. Because these studies seek the understanding of a SRS in specific contexts, items Q10-Q19 are irrelevant for them. In contrast, when evaluating the quality of 'what-is-better' papers, it is appropriate to use all the items in the new checklist.

3.3 Evaluation of the Reliability of the Modified Questionnaire

Once the practitioners used the adjusted checklist, their results were compared for the purpose of determining the proportion of agreement between them (the practitioners). This analysis was done by the third author, who did not participate in the adjusting process of the practitioner's checklist. This proportion of agreement was measured for each of the nineteen questions of the checklist. We make the note that in this study we followed the recommendation of Gwet [20] that in case of two raters, the rating data are reported using the distribution of the reviewed artifacts (the RE studies in our case) by rater and response category as shown in Table 2.

Table 2: Observed Agreement by Category

Question	C1	C2	C3
Q1			1
Q2	0.7		0.4
Q3	0.43		0.76
Q4	0.7		0.45
Q5	0.81		0.5
Q6	0.67		0.84
Q7	0.25	0.25	0.43
Q8	0.87		
Q9			0.96
Q10			0.4
Q11	0.28		0.3
Q12	0.27		0.33
Q13	0.39		0.33
Q14	0.45		0.17
Q15	0.18		0.23
Q16	0.24		0.25
Q17			0.43
Q18	0.33		0.33
Q19	0.36		

The two reviewers agreed on a total of 24 primary studies within each of the three categories C1, C2, and C3. The categories C1 – C3, stand for the points on the three point scale, respectively, meaning 'no', 'partially', and 'yes'. Table 2 shows the proportion of agreement observed between the two reviewers within each of the three categories (C1, C2, C3) for the 19 questions of the checklist.

By carrying out a composite assessment, 42 % of total of questions formulated (Q1, Q2, Q3, Q4, Q5, Q6, Q8, and Q9) show having a better proportion of agreement (>=50%) between two reviewers (See Table 3).

From the results that we obtained, we calculate the Kappa coefficient, which is widely used to measure inter-observer variability **Error! Reference source not found.**. As ordinal scale questions were formulated, we apply a **weighted Kappa**, which penalizes disagreements in terms of their seriousness, whereas unweighted kappa treats all disagreements equally. A number of methods of weighting are available **Error! Reference source not found.**, but quadratic weighting is most common. For this reason, we used a Kappa with quadratic weighting in this study. The equation for κ_w is:

$$\kappa_w = (p_{o(w)} - p_{e(w)}) / (1 - p_{e(w)})$$

where $p_{o(w)}$ is the weighted observed proportional agreement between the two observers, and $p_{e(w)}$ is the weighted proportional agreement expected just by chance.

Table 3: Proportions of Agreement (Composite Assessment).

Question	Maximun possible	Chance expected	Observed	.95 CI of Observed	
				Lower	*Upper*
Q1	1	1	1	0.8342	1
Q2	1	0.5424	0.76	0.5448	0.8984
Q3	0.84	0.5808	0.76	0.5448	0.8984
Q4	0.76	0.4912	0.76	0.5448	0.8984
Q5	0.92	0.6032	0.8	0.587	0.9239
Q6	0.8261	0.6087	0.8261	0.6045	0.9428
Q7	0.76	0.3312	0.48	0.2834	0.6825
Q8	0.76	0.3728	0.68	0.4645	0.8427
Q9	0.96	0.96	0.96	0.7768	0.9979
Q10	0.4	0.4	0.4	0.2181	0.6111
Q11	0.28	0.1008	0.2	0.0761	0.413
Q12	0.36	0.2352	0.32	0.1573	0.5355
Q13	0.44	0.3568	0.4	0.2181	0.6111
Q14	0.44	0.3136	0.4	0.2181	0.6111
Q15	0.44	0.216	0.24	0.1016	0.4552
Q16	0.44	0.2144	0.28	0.1287	0.496
Q17	0.44	0.3696	0.4	0.2181	0.6111
Q18	0.36	0.2992	0.36	0.1871	0.5738
Q19	0.44	0.4048	0.36	0.1871	0.5738

In order to interpret the strength of agreement for the κ_w values obtained, Landis and Koch **Error! Reference source not found.** proposed the following scale: $\kappa_w \leq 0$ means poor, κ_w varying between 0.1 and 0.2 mean slight, κ_w falling between 0.21 and 0.40 means fair, between 0.41 and 0.60 - moderate, between 0.61 and 0.80 - substantial, and between 0.81 and 1.0 - almost perfect. According to this interpretation scale [15], we take a Kappa value $\kappa_w > 0.21$ to indicate an acceptable agreement for the purpose of our study. The κ_w values in our study are presented in Table 4. Therein, we observe that questions Q11-Q18 show a poor agreement. For questions Q10 and Q19, the κ_w value was not calculated because the observed concordance is equal or smaller than mean-chance expected concordance (see Table 3). The κ_w coefficient could not be calculated for the Questions Q1 and Q9 either. Therefore, we found that 9 of out 19 questions show an acceptable agreement level.

Table 4: Kappa with Quadratic Weighting.

Question	Observed Kappa	Standard Error
Q1	1	-
Q2	0.476	0.189
Q3	0.427	0.165
Q4	0.474	0.158
Q5	0.560	0.190
Q6	0.660	0.172
Q7	0.220	0.141
Q8	0.514	0.229
Q9	0.9	-
Q10	-	-
Q11	0.105	0.085
Q12	0.111	0.069
Q13	0.194	0.144
Q14	0.126	0.077
Q15	0.066	0.131
Q16	0.179	0.114
Q17	0.095	0.049
Q18	0.191	0.106
Q19	-	-

3.4 Reflections and Lessons Learned

The practitioners used the modified checklist (Table 1). Their evaluations of the 24 studies indicate that: (1) their ratings on 9 questions fully agreed with respect to all 24 studies. (2) there were disagreements on 10 items, with respect to the 24 studies. However these disagreements were resolved in a discussion and did not trigger further modification of the checklist. It is therefore our view that the checklist in Table 1 seems one possible instrument that could be considered in the evaluation of the quality of primary RE studies. However, to increase the reliability of the checklist, more reviewers are required and more discussion cycles are necessary. Our plan is to involve at least two more practitioners in order to improve the reliability of the practitioner's checklist proposed.

In our reflection, we acknowledge that the availability of RE-context-specific and operationalized definitions for the checklist items is a critical factor for the checklist's applicability in the RE field. In our view, if a team of researchers is to do their evaluations well, then sizeable amount of time should be allocated in a research project for contextualization and operationalization of the checklist items. We must note however, that we do not claim to have a complete list of checklist problems (that would cost a research team some time to resolve). Regarding the number of known problems, we are conscious that the included feedback of such a small number of practitioners has the following limitation: it might be the case that we could have got a much better indicative of checklist problems if we would have involved a large population of RE practitioners. However, this was infeasible at the time of carrying out this research due to resource constraints. We also think that including two experienced RE practitioners who currently are also university researchers offsets this limitation, as the practitioners are sensitive to the research method being used and know what kind of text to watch for in a primary study and how to rate the evidence reported. We also wanted to compare our experiences to those of other researchers/practitioners that used the checklists. However, we could not find any publication which was explicitly focused on this specific checklist. We, therefore, invite other researchers from the empirical SE and RE communities to get involved in SLRs, use the checklist and share their experiences.

Once the practitioners used the adjusted checklist, their results were compared for the purpose of determining the proportion of agreement between them (the practitioners). This analysis was done by the third author, who did not participate in the adjusting process of the practitioner's checklist. This proportion of agreement was measured for each of the nineteen questions of the checklist. We make the note that in this study we followed the recommendation of Gwet [20] that in case of two raters, the rating data are reported using the distribution of the reviewed artifacts (the RE studies in our case) by rater and response category as shown in Table 2.

4 CONCLUSIONS

This paper presents experiences and lessons learnt in using the practitioner's checklist that was published in [6]. We make the note that while [6] provides the checklist, it gives very little support on how to 'operationalize' the checklist items, and how to aggregate the ratings of the checked items into an overall rating. (e.g. a practical question encountered by the two practitioners was whether or not all items should be present in order to conclude that a primary study is of sufficient quality, or one can decide on a cut-over percentage). We proposed a justified modification of the checklist (see Table 1) that fits the RE research context. We used it to evaluate 24 studies on the comprehensibility of SRS. Our evaluation of the inter-rater reliability of the checklist indicates that 9 out of the 19 questions in the checklist obtained an acceptable level of agreement.

Based on our experience, we learned three important lessons: (1) Evaluating the quality of studies in RE takes more time than one might think; (2) For a checklist to be useful, it should be contextualized, be operationalized, and keep separate the items referring to the quality of reporting from those referring to the quality of analysis; and (3) Evaluating the reliability of the checklist by means of the kappa coefficient and by using at least two reviewers is an important point to indicate follow-up steps towards increasing the checklist's reliability. (In our research, we plan to engage at least two more practitioners and replicate the study).

We think it's worth sharing the modified checklist with other researchers who undertake SLRs in RE, as they would not have to start from scratch in their efforts to contextualize and operationalize the checklist in [6]. However, we, by no means, claim that Table 1 is complete, e.g. while we left out those items that do not pertain to RE, we did not investigate what aspects that do pertain to RE should be brought in. Understanding what represents a complete checklist forms our second line for immediate future research

A APPENDIX: STUDIES INCLUDED

The rules about hierarchical headings discussed above for the body of the article are di.erent in the appendices. In the appendix environment, the command section is used to indicate the start of each Appendix, with alphabetic order designation (i.e., the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure within an Appendix, start with subsection as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

S1. Genero M., Cruz-Lemus J., Caivano D., Abrahão S., Insfrán E., Carsí J., Assessing the Influence of Stereotypes on the Comprehension of UML Sequence Diagrams: A Controlled Experiment. Model Driven Engineering Languages and Systems. 11th International Conference, MoDELS 2008, Toulouse, France, September 28 - October 3, 2008. Proceedings, pp 280-294.

S2. Gravino C., Scanniello G. and Tortora, G. An Empirical Investigation on Dynamic Modeling in Requirements Engineering. Model Driven Engineering Languages and Systems Conference, MODELS 2008, pp. 615-629

S3. Kamsties E., Antje von Knethenb, Reussner R.. A controlled experiment to evaluate how styles affect the understandability of requirements specifications, Information and Software Technology Journal, 45 (2003) 955–965.

S4. Overmyer S., A Methodology for Constructing User-Oriented Requirements Specifications for Large-Scale Systems Using Electronic Hypermedia. Requirements Engineering (1999) 4:1–18, Springer Verlag.

S5. Agarwal R., Prabuddha D, and Sinha A.. Comprehending Object and Process Models: An Empirical Study, IEEE Transactions on Software Engineering, 25(4):541-556, August 1999.

S6. Anda B., Sjøberg D., and Jørgensen M., Quality and Understandability of Use Case Models. ECOOP 2001, Object-Oriented Programming, LNCS, Springer Berlin, 402-428

S7 Cioch F., Measuring Software Misinterpretation, Journal Systems Software, Elsevier, 1991, 14:85-95

S8. Xie S., Kraemer E., and Stirewalt R., Empirical Evaluation of a UML Sequence Diagram with Adornments to Support Understanding of Thread Interactions. 15th IEEE International Conference on Program Comprehension (ICPC'07), 2007

S9. Otero M., Dolado J., An Initial Experimental Assessment of the Dynamic Modelling in UML, Empirical Software Engineering Journal, 7, 27–47, 2002. Kluwer Academic Publishers.

S10. Glezer Ch., Nahmani E., Shoval P., Experimental Comparison of Sequence and Collaboration Diagrams in Different

Application Domains, Proceedings of the Workshop on Evaluating Modeling Methods for Systems Analysis and Design, (EMMSAD'05), Portugal, 2005, pp. 463-476.

S11. Rozilawati R., Snook C., Poppleton M., Comprehensibility of UML-based Formal Model – A Series of Controlled Experiments, WEASELTech'07, November 5, 2007, Atlanta Georgia, USA, Copyright ACM

S12. Finney K., Fenton N., Fedorec A., Effects of structure on the comprehensibility of formal specifications, IEE Software. 146(4): 193-202, August 1999

S13. Agarwal R.; Sinha A.; Tanniru M. Cognitive fit in requirements modeling: A study of object and process method, Journal of Management Information Systems; Fall 1996. pg. 137

S14. Glezer Ch., Last M., Nachmany E., Shoval P.. Quality and comprehension of UML interaction diagrams-an experimental comparison, Information and Software Technology 47 (2005) 675–692.

S15. Kuzniarz L., Staron M., Wohlin C., An Empirical Study on Using Stereotypes to Improve Understanding of UML Models, Proceedings of the 12th IEEE International Workshop on Program Comprehension (IWPC'04)

S16. Cruz-Lemus J., Genero M., Morasca S., and Piattini M., Using Practitioners for Assessing the Understandability of UML Statechart Diagrams with Composite States, J.-L. Hainaut et al. (Eds.): ER Workshops 2007, LNCS 4802, pp. 213–222, 2007. Springer-Verlag Berlin Heidelberg 2007.

S17. Cruz-Lemus J., Genero M., Manso E., and Piattini M. Evaluating the Effect of Composite States on the Understandability of UML Statechart Diagrams. L. Briand and C. Williams (Eds.): MoDELS 2005, LNCS 3713, pp. 113-125, 2005. Springer-Verlag Berlin Heidelberg 2005.

S18. Otero M., Dolado J., An empirical comparison of the dynamic modeling in OML and UML. The Journal of Systems and Software 77 (2005) 91–102.

S19. Peleg M., Dori D., The Model Multiplicity Problem: Experimenting with Real-Time Specification Methods. IEEE Transactions on Software Engineering, 26(8)742-759, August 2000.

S20. Swan, T. Barker, C. Britton, M. Kutar, "An empirical study of factors that affect user performance when using UML interaction diagrams," ISESE, pp.10 pp., 2005 International Symposium on Empirical Software Engineering, 2005.

S21. Kuzniarz L., Byckling P., Gerdt P., Sajaniemi J., Increasing Comprehensibility of Object Models: Making the Roles of Attributes Explicit in UML Diagrams. Nord. J. Comput. 13(3): 149-161 (2006)

S22. Cruz-Lemus J., Genero M., Piattini M., and Toval A., An Empirical Study of the Nesting Level of Composite States Within UML Statechart Diagrams, J. Akoka et al. (Eds.): ER Workshops 2005, LNCS 3770, pp. 12 – 22, 2005.

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.

S24. Staron M., Kuzniarz L., Wohlin C., Empirical assessment of using stereotypes to improve comprehension of UML models: A set of experiments, Journal of Systems and Software, 2006, pp. 727-742.

REFERENCES

[1] Cheng, B., Atlee, J. Research Directions in Requirements Engineering, Future of Software Engineering (FOSE) - ICSE 2007, IEEE CS Press, pp. 285-303.

[2] Condori Fernandez, N., Daneva, M., Sikkel, K., Wieringa, R., Dieste, O., Pastor, O., A Systematic Mapping Study on Empirical Evaluation of Software Requirements Specifications Techniques. In: ESEM2009. IEEE CS Press, pp. 503-505.

[3] Carver, J.C., Dieste, O., Kraft, N.A. et al. How Practitioners Perceive the Relevance of ESEM Research. 10th ESEM 2016.

[4] Kaindl, H., Brinkkemper, S., Bubenko J., et al. Requirements Engineering and Technology Transfer: Obstacles, Incentives and Improvement Agenda. REJ 7(3):113– 123, 2002.

[5] Franch, X., Méndez Fernández, D., Oriol, M., Vogelsang, A., Heldal, R., Knauss, E., Travassos, G., Carver, J.C..Dieste, O., Zimmermann, T.: How do Practitioners Perceive the Relevance of Requirements Engineering Research? An Ongoing Study. RE 2017: 382-387

[6] Kitchenham, B.A., Al-Kilidar H., Babar, M.A., M. Berry, Cox, K., Keung L., Kurniawati, F. Staples M., Zhang H., Zhu L, Evaluating guidelines for reporting empirical software engineering studies. Empirical Software Engineering 13(1), 2008, pp 97-121.

[7] Kitchenham, B., Procedures for Undertaking Systematic Reviews, Joint Technical Report, Computer Science Department, Keele University (TR/SE-0401) and National ICT Australia Ltd. (0400011T.1), 2004.

[8] Racheva, Z., Daneva, M., Sikkel, K., Value Creation by Agile Projects: Methodology or Mystery? In: Int. Conf. on Product-Focused Software Process Improvement, Springer, pp. 141-155.

[9] Kitchenham B.A., T. Dybå, and M. Jørgensen, "Evidence-Based Software Engineering," Proc. 26th Int'l Conf. Software Eng, IEEE CS Press, 2004, pp. 273–281.

[10] Dybå, T. and Dingsøyr, T. (2008) Empirical Studies of Agile Software Development: A Systematic Review, Information and Software Technology, 50(9-10): 833–859.

[11] Sjøberg, D.I.K., Dybå, T., and Jørgensen, M. (2007) The Future of Empirical Methods in Software Engineering Research, FOSE'07, IEEE CS Press, pp. 358–378.

[12] Jedlitschka, A., Ciolkowski, M. & Pfahl, D. (2008), Reporting experiments in software engineering, in F. Shull, J. Singer & D. Sjøberg, eds, 'Guide to Advanced Empirical Software Engineering', Springer, London, chapter 8.

[13] Higgins, J.P.T., S. Green, Cochrane Handbook for Systematic Reviews of Interventions, Wiley, 2008.

[14] Sim J., Wright C., The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements, Physical Therapy Journal, March 2005 vol. 85 no. 3 257-268

[15] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics.1977 ;33:159–174

[16] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull.1968 ;70:213–220

[17] Wieringa, R.J. (2010) Relevance and problem choice in design science. In: Global Perspectives on Design Science Research (DESRIST). 5th International Conference, 4-5 June, 2010, St. Gallen. pp. 61-76. LNCS 6105.

[18] Wieringa, R.J. (2009) Design Science as Nested Problem Solving. In: Proceedings of the 4th Int. Conference on Design Science Research in Information Systems and Technology, Philadelphia. pp. 1-12. ACM.

[19] A.T. March and G.F. Smith. Design and natural science research on information technology. Decision Support Systems, 15(4):251-266, December 1995.

[20] Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement, British Journal of Mathematical and Statistical Psychology, 61, 2008, pp 29–48.

[21] Kuhrmann, M., Méndez Fernández, D., Daneva, M., On the pragmatic design of literature studies in software engineering: an experience-based guideline. Empirical Software Engineering 22(6): 2852-2891 (2017)

[22] Wieringa, R., Daneva, M. Six strategies for generalizing software engineering theories. Sci. Comput. Program. 101: 136-152 (2015)

[23] Daneva, M., Damian, D., Marchetto, A., Pastor, O., Empirical research methodologies and studies in Requirements Engineering: How far did we come? Journal of Systems and Software 95: 1-9 (2014)

[24] Bano, M., Zowghi, D., Ikram, N., Systematic reviews in requirements engineering: A tertiary study. EmpiRE 2014: 9-16