An Evaluation of Quality Checklist Proposals – A participant-observer case study

Barbara A. Kitchenham
School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK
B.A.Kitchenham@cs.keele.ac.uk

O. Pearl Brereton
School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK
O.P.Brereton@cs.keele.ac.uk

David Budgen
Department of Computer Science, Durham University, Durham, Science Laboratories, South Road, Durham City, DH1 3LE, UK.
david.budgen@durham.ac.uk

Zhi Li
School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK
z.li@epsam.keele.ac.uk

Abstract Background: **A recent set of guidelines for software engineering systematic literature reviews (SLRs) includes a list of quality criteria obtained from the literature. The guidelines suggest that the list can be used to construct a tailored set of questions to evaluate the quality of primary studies. *Aim*: This paper aims to evaluate whether the list of quality criteria help researchers construct tailored quality checklists. *Method*: We undertook a participant-observer case study to investigate the list of quality criteria. The "case" in this study was the planning stage of a systematic literature review on unit testing. *Results*: The checklists in our SLR guidelines do not provide sufficient help with the construction of a quality checklist for a specific SLR either for novices or for experienced researchers. However, the checklists are reasonably complete and lead to the use of a common terminology for quality questions selected for a specific systematic literature review. *Conclusions*: The guidelines document should be amended to include a much shorter generic checklist. Researchers might find it useful to adopt a team-based process for quality checklist construction and provide suggestions for answering quality checklist questions.**

Keywords: case study, systematic literature review, primary study quality, quality checklists, guidelines evaluation

1. INTRODUCTION

This paper reports the results of one research question addressed by a case study undertaken by the Evidence-based Practice Informing Computing (EPIC) project to investigate the use of systematic literature reviews (SLR) in software engineering. The case study was defined in a case study protocol (Budgen, 2008). The specific research question addressed in this paper is:

> RQ10: How useful are the quality checklists provided in the latest version of the guidelines for performing systematic literature reviews.

The research question number was initially identified in the EPIC Scoping document (Brereton and Kitchenham, 2007). The guidelines in question are those developed by Kitchenham and Charters (2007). There are many recommendations for quality checklists (see for example, Crombie, 1996; Fink, 2005; Greenhalgh, 2000; Petticrew and Roberts, 2005). In their guidelines, Kitchenham and Charters compiled a collated list of checklist items (i.e. individual questions) for different types of study (controlled experiments, surveys, general empirical studies and qualitative studies). They pointed out that researchers needed to select and tailor checklist items into a set of items (i.e. a quality checklist) appropriate for their own study. However, they did not provide any advice on how this should be done.

The basic methodology used in the case study was a participant-observer study using Yin's approach to case study design (Yin, 2003). The "case" in this study was the planning stage of a systematic literature review (SLR) looking at unit testing. Following Yin's terminology, the case study protocol defined four propositions related to the research question:

- RQ10-P1 "The quality guidelines are of value to novice researchers."
- RQ10-P2 "The checklists in the guidelines are of value to experienced researchers."
- RQ10-P3 "Given the guidelines, different researchers will develop similar quality checklists for the same research question."
- RQ10-P4 "Different quality checklists will lead to different conclusions."

This paper addresses only the first three propositions, proposition 4 is stated for completeness. The method used to address these propositions was defined in the Case Study Protocol and is summarised in Section 2. The results of analysing the collected data are reported in Section 3 and the extent to which the results confirm or contradict our research propositions are discussed in Section 4. Our conclusions are reported in Section 5.

2. METHODOLOGY

This section describes the methodology we used to address our research question. The methodology was defined in our case study protocol prior to starting data collection (Budgen, 2008).

2.1 Case Study Roles

This study is being conducted entirely within the EPIC team, and hence some of the team members are required to perform specific roles as case study researchers as well as roles in the systematic literature review. In the SLR, the roles are:

Supervisors: Pearl Brereton and David Budgen
Researcher Assistants: Zhi Li (Keele) and Feng Bian (Durham)
Reviewers: Pearl Brereton, David Budgen, Barbara Kitchenham, Stephen Linkman, Mahmood Niazi and Mark Turner.

In the case study, David Budgen is the case study leader, and he and Pearl Brereton also act as *observers*, maintaining records of their supervisory activities. The other members of the case study team are: Barbara Kitchenham, Stephen Linkman, Mahmood Niazi, Michael Goldstein and Mark Turner. (Note, throughout this report, individual researchers performing specific roles are named, because anonymity is impossible when accurately reporting an observer-participant study.)

2.2 Data Collection

The case study addressing this research question, involved two RAs at different establishments independently undertaking the planning stage of a systematic literature review related to unit testing. As part of the first phase of the SLR, each RA needed to construct a quality checklist. Furthermore, phase 2 of the SLR required an integrated quality checklist to be used for the remaining phases of the SLR.

The case study protocol took advantage of the need for both independent quality checklists and an integrated quality checklist to investigate the value of quality checklist tables provided in the guidelines. The process we adopted was that the RAs and their supervisors prepared quality checklists independently (using the lists of quality questions provided in the guidelines) and completed a questionnaire about the process (the quality checklist evaluation form). Another member of the team (Kitchenham) also prepared a quality checklist without using the guidelines.

Subsequently, the case study team had a meeting to jointly construct an agreed quality checklist using the guidelines. The case study team meeting included other members of the EPIC team in addition to the RAs, Kitchenham, Brereton and Budgen, i.e. Michael Goldstein and Mahmood Niazi but did not include Mark Turner or Steve Linkman. This meeting reviewed the checklists and discussed which were most related to study quality. In general. the team found that:

- Some checklist items were more related to the quality of the report than quality of the study being reported. Researchers are usually advised to concentrate on the quality of the study methodology rather than how the study is reported. E.g. a clear statement of aims is good reporting practice, but even if it is missing, the aims can usually be inferred from a clear description of the result and analyses. However, in practice it is often difficult to disentangle these issues
- Some checklist items referred to similar issues but at different stages in the experimental process. E.g. “Do the study measures allow the questions to be answered?” is a design item and “Are the variables adequately measured” is a data collection issue.
- Some checklist items were related such that if one was not true the other could not be true, so in a sense one of the items was an indicator of the truth of the other. E.g. one item was “Are the variables used adequately measured?” and another was “Are the measures used fully defined?”. It would not be possible to answer the first question unless the second question was true.
- Given that the quality checklist is intended to provide an assessment of whether the results of the primary study are trustworthy, we should include an item to ask whether the study had performed a sensitivity or stability analysis to ensure that the results were not a result of one or two abnormal values.

Overall, the meeting aimed to select as few questions as possible and concentrated on those related to the validity of primary study results. The meeting identified a total of 11 questions. After further review of the selected questions, two questions were removed from the main list:

- The question "Has the method of randomization been defined" was moved to a quality data extraction procedure which says "If the randomization process has not been specified, we will contact the authors to ask for details".

- The question "Do the numbers add up across tables?" is not so much a quality criterion in its own right but an indicator of analysis problems such as that drop outs had not been properly handled, or that the analysis and design did not match. It was included as one of the factors to look for to assess whether the analysis was correct.

The full set of issues raised for each checklist item are presented in an EPIC technical report (Kitchenham et al., 2008).

In summary, the data used to address our research propositions are:

- For RQ10-P1 & RQ10-P3: The quality checklists prepared by the RAs during phase 1 of the SLR using the guidelines.
- For RQ10-P2 & RQ10-P3: Quality checklists prepared by each of the supervisors during phase 1 of the SLR using the guidelines.
- For RQ10-P1 & RQ10-P2 & RQ10-P3: A quality checklist prepared by another case study team member (acting as an expert) without using the guidelines
- For RQ10-P1 & RQ10-P2 & RQ10-P3: The quality checklist agreed for use by the SLR team and the RAs for phase 2 of the SLR.
- For RQ10-P1, RQ10-P2, & RQ10-P3: The quality checklist evaluation forms completed by the RAs and the supervisors.

2.3 Data Analysis

The data collection process resulted in six checklists:

1. C1 – the checklist prepared by research assistant 1.
2. C2 – the checklist prepared by research assistant 2.
3. C3– the checklist prepared by supervisor 1.
4. C4- the checklist prepared by supervisor 2.
5. C5 – the checklist prepared by the expert.
6. AC – The checklist agreed by the research team.

Each of the first five checklists was compared with the 6th checklist and the following metrics were calculated:

$$\text{Percentage correctness} = 100 \times CIi/TotAC \quad (1)$$

$$\text{Percentage completeness} = 100 \times CIi/TotCi \quad (2)$$

where i= 1,...,5.

CIi = the number of items in quality checklist i that are also in AC (i.e correct items).
TotAC = the number of items in the final agreed checklist.
TotCi= the total number of items in checklist Ci.

The best checklist would be one that maximised both the percentage completeness and the percentage correctness.

The Quality checklist evaluation forms were reviewed:

1. To investigate whether there are any commonly occurring problems.
2. To identify strategies for addressing common problems.
3. To identify any ways in which the guidelines can be changed to better support the development of appropriate quality checklists.

3. RESULTS

The checklists prepared by each participant are presented in Table 4 in Appendix 2, excluding the checklist prepared by one of the RAs who completely misunderstood the checklists provided in the guidelines and simply used all the available questions related to experiments. Table 1 summarises the overlap among the checklists. Table 2 shows the completeness and correctness scores for each checklist compared with the accepted checklist (AC).

The quality checklist evaluation form with all three responses is shown in Appendix 1. The RA who misunderstood the quality checklist task did not complete the quality checklist evaluation form. The agreed quality checklist contained only one item that was not included in the guidelines checklists (i.e. Was any sensitivity analysis performed?).

The following issues can be observed from the checklists produced by individuals (see Table 4):

1. When the guidelines are used they lead to similar checklists (see C1, C3, C4). In particular:
   - There is a strong overlap between the checklists based on the guidelines.
   - They use virtually the same terminology for checklist questions.
2. Without the guidelines there is much less similarity:
   - Even when they overlap, the questions are worded differently.
   - The C5 checklist items were sometimes at a lower level of granularity. For example:

a. C5 asked about specific validity issues i.e. nature of any seeded errors, representativeness of testing objects, student participants, in contrast C1 and C4 asked whether the sample was representative of the population to which the results would generalise.
b. C5 asked specific questions about how data analysis could go wrong i.e. invalid data preparation, fishing for results or lack of sensitivity analysis while C1, C3 and C4 asked whether the statistical methods were described.

Table 1: Overlapping Questions Among Checklists

	C2 (ZL)	C3 (DB)	C4 (OPB)	C5 (BAK)	AC	Total
C1	12	9	7	2	1	12
C2		24	18	6	8	50
C3			14	4	4	24
C4				4	5	18
C5					6	12
AC						9
Combined list						56

Table 2: Completeness And Correctness Values For The Checklists Compared With Accepted Checklist

Checklist	Correctness	Completeness
C1	1/9=0.11	1/12=0.083
C2	8/9=0.88	8/50=0.16
C3	4/9=0.44	4/24=0.17
C4	5/9=0.56	5/18=0.27
C5	6/9=0.67	6/12=0.5

In comparison with the team checklist (AC), C5 is closest. However, this may not be due to "expertise" on the part of Kitchenham (who compiled C5) but because the team meeting including two statisticians (Kitchenham and Goldstein) and concentrated strongly on discussions of the statistical validity of study results, because if primary studies are invalid any aggregation of them will also be invalid.

4. DISCUSSION

This section discusses the extent to which the results support or not our research propositions.

4.1 Proposition RQ10-P1

RQ10-P1 The quality guidelines are of value to novice researchers.
This proposition would be supported if the completeness and correctness values for the RAs are as good as or better than the completeness and correctness of the C5 checklist (i.e. the expert acting without the checklists in the guidelines).

Table 2 makes it clear that the RAs checklists were not as good as the expert's checklist (C5). Although RA2 had a high Correctness score, this was due to simply copying the entire checklist, and so he has a low Completeness score. **The proposition is therefore contradicted**.

It seems clear that the checklists by themselves are not helpful to novices. There needs to be an accompanying discussion of how to tailor the checklists to specific situations. The team meeting that developed the accepted guidelines identified the following issues:

- Restrict the number of checklist items. We suggest between 6 and 12. The fewer the items the more likely that they will be correctly evaluated, yet too few questions make it difficult to capture all the aspects of methodology quality.
- Consider what information is needed to answer the checklist question. Questions relating to issues that are not usually reported will require asking the authors for details. Questions that are very subjective will not give reliable answers.
- Concentrate on checklist items that relate to good reporting rather than good experimental procedures.
- Remove checklist items that overlap.
- Concentrate on checklist items related to whether the results are valid.
- Use some checklist items as a part of the advice for the evaluation of other checklists items.

Review of the Quality Checklist evaluation form completed by RA1, suggests that overall he felt the checklists were useful and complete. He suggested more discussion of statistical and experimental terms would be beneficial.

4.2 Proposition RQ-P2

RQ10-P2 The checklists in the guidelines are of value to experienced researchers.
This proposition will be supported if the Completeness and Correctness values for the supervisors' checklists are as good as or better than the Completeness and Correctness of the C5 checklist (i.e. the expert acting without the checklists in the guidelines).

The Correctness of the C5 checklist was better than that of the supervisors' checklists but only by one or two questions. The Completeness of the C5 checklist was much better than the supervisors' completeness. **The proposition is therefore contradicted**.

The main problem with the supervisors' checklists was the number of unnecessary checklist items. This would be improved by:

- Suggesting some limits to quality checklist size in the guidelines
- Making sure that the checklist concentrates on whether results are trustworthy rather than whether they are well-reported.

The supervisors had mixed views about the checklists and some concern that the checklists would be difficult for non-statisticians or researchers inexperienced in empirical studies. One supervisor noted that the testing SLR is primarily a mapping SLR and thought the quality checklists were misleading for that type of study. Both supervisors reinforce the point made by the RA, that some statistical expertise is required to use the guidelines effectively and some terminology needs to be included in the guidelines.

4.3 Proposition RQ10-P3

RQ10-P3 Given the guidelines different researchers will develop similar quality checklists for the same research question.
This proposition would be supported if the supervisors' checklists are more similar to each other than they are to the expert's checklist based on the pair wise completeness and correctness percentage metrics.

Overall Table 1 and Table 2 make it clear that there was considerable overlap between the checklists prepared by the two supervisors and that there was much less overlap with that of the expert (C5). Also, from the actual checklists (Kitchenham and Charters, 2007), it was clear that use of the guidelines encouraged a common terminology for quality criteria. **Thus, our proposition is supported.**

4.4 Implications for Quality Checklists

The exercise we undertook to construct a team-based quality checklist for our testing SLR suggested that the number of generic checklist items could be much reduced. Based on our team discussions, it would seem that the quality checklist shown in Table 3 would both reduce redundancies among items and remove items related to reporting quality, leading to a much simpler generic checklist.

However, any generic checklist still needs to be refined/amended in the light of the specific requirements of the SLR. We found the process of constructing quality checklists individually and then discussing the similarities and differences in order to produce a team checklist was very useful. We would recommend such a process to other researchers.

We would also recommend considering issues related to answering the questions asked by the checklist items as shown in column 3 of Table 3. The quality checklists are intended to be more objective and auditable than a simple subjective assessment of study quality but this implies that there needs to be a means of ensuring different researchers evaluate each checklist item in a consistent manner. Establishing additional criteria to help answer the checklist item question is a starting point, but it is also important to prototype the quality evaluation process in order to assess its consistency. That is, the quality checklists should be trialled on a subset of primary studies and the results compared and any disagreements discussed to refine both the quality checklists and the quality evaluation process.

Table 3: Generic Quality Checklist for Quality of Experiments

Id	Checklist item	Answering the question posed by the checklist item
1	Are the study measures valid?	These are indicators of valid measures: 1. The measures are plausible measures of the construct they are meant to represent. 2. The measures are direct measures of well-defined concepts. 3. The measurement scales are respected (e.g. categorical measures are not treated as ordinal or interval)? 4. The data collection process is defined and appropriate. If information about the variables, measures, or data collection process is not reported, you will need to request information from the authors of the study.
2	Were experimental units appropriately allocated to treatments given the number of units and the overall experimental design?	If the process was not defined, the authors must be contacted and asked about the randomization process. Appropriate random allocation is different for experiments where units are exposed to one treatment and experiments were units are exposed to several treatments. 1. For experiments where experimental units are exposed to only one treatment, units should be assigned randomly to treatment allowing for any blocking factors, and (usually) an equal number of units per experimental condition.

		2. For experiments where experimental units are exposed to several different treatments, the order of treatments should be randomized for experimental units.
3	Were all treatment groups (including any control groups) treated equivalently during the preparation for and conduct of the experiment?	No if: 1. Some treatment groups were given less training than others. 2. Training was of better quality for some treatment groups than others (e.g. experimenters were experts in one method but not another). 3. Experimenters gave more attention to some groups than others. 4. Some treatment groups were recognised as more important or prestigious than others. 5. Some treatment groups expected expertise not available to the subjects.
4	If results were not statistically significant, does a power analysis confirm that the number of experimental units was sufficient to detect a moderate effect?	If no power analysis was performed, the SLR researchers should undertake a power analysis themselves (if the information needed for such an analysis is unavailable, the study authors must be contacted).
5	Could lack of blinding introduce bias?	Yes if either of the following is true: 1. Experimenters knew which subject was in which experimental group during the experiment. 2. Outcome assessment (i.e. any marking/evaluation of experimental outcomes) made it clear which group a subject was assigned to.
6	Did the statistical analysis cater appropriately for any design/conduct problems?	No if any of the following are true: 1. No necessary adjustment was made for drop-outs or unequal sizes of treatment group. 2. No adjustment was made for possible selection bias when random allocation was not possible (e.g. covariance analysis). 3. No appropriate adjustment was made for missing values.
7	Were the statistical analyses used in the study appropriate for the questions of relevance to the SLR?	The analysis might be incorrect if: 1. Blocking effects included in the design were not considered in the analysis. 2. The design is unbalanced (e.g. there are different numbers of experimental units in each experimental condition). 3. Data analysis was (dangerously) inconsistent with measurement scales (e.g. categorical data treated as numerical). 4. Complex designs involving blocks and treatments were analyzed using simplistic non-parametric statistics. 5. Number do not add up across different tables Note if the answer to question 6 is “No” the answer to this question must also be “No”.
8	Is there evidence of multiple statistical testing or large numbers of post hoc analysis?	Consider: 1. Whether a large number of correlations have been tested for significance. 2. The number of tests is large compared with the number of experimental units. 3. Significant results are reported only for a number of subsets of the data – particularly if the subsets appear to have been decided post-hoc. 4. Significant results are found after removing or including statistical outliers that are not justified as outliers by other factors.
9	Have the results been confirmed by a sensitivity or stability analysis?	If the authors have not reported any such analysis it is probably safe to assume that none has been performed.

4.5 Limitations of the study

There are two major limitations to this study:

1. We are evaluating our own work, so might introduce bias into the evaluation.
2. The results assume that the accepted quality checklist is itself correct.

With respect to the first concern, this can never be completely discounted. In particular, Kitchenham created the quality checklists presented in the guidelines. In this case, there was a gap of over a year between compiling the list of quality criteria questions and this study, which would reduce any influence constructing the compiled list of quality criteria questions had on constructing the testing study quality checklist. Kitchenham also contributed to the discussion of the agreed quality checklist, probably influencing its emphasis on statistical validity. However, we have tried to ensure that we used the observer-participant case study methodology as rigorously as possible. In particular we specified our propositions, data collection, analysis and interpretation procedures prior to data collection.

With respect to the correctness of the accepted quality checklist (AC), again we cannot be sure it is the best possible checklist. In fact, trials of using the checklist have identified two issues:

1. The answers should probably be an ordinal scale (e.g. not at all, a little, somewhat, mostly, fully) rather than a simple Yes/No.
2. The questions are appropriate for human-based controlled experiments but not technology-based studies that are probably better described as "benchmarking" studies.

5 CONCLUSIONS

It is clear that the guidelines checklists do not provide sufficient help either to novices or to experienced researchers to allow them to construct appropriate quality checklists for a specific systematic literature reviews.

However, the Quality Checklist Evaluation forms completed by the RA and supervisors suggest there is some value in having the checklists in the guidelines as a starting point for constructing a quality evaluation checklist for an SLR. In addition, the use of the guidelines checklists tends to impose some similarity in terminology. Furthermore the guidelines checklists are reasonably complete. Only one checklist item in the accepted quality checklist (AC) was not to be found in the guidelines checklist. In addition, all the participants who completed the quality check list evaluation form agreed that the guidelines check lists were satisfactory or excellent with respect to completeness.

The revised generic checklist items shown in Table 3 may be a better starting point for constructing checklists for human-based experiments and quasi-experiments (although not for technology-centred benchmarking studies). In addition, the guidelines need to be supplemented with more information about statistical and experimental terminology and advice on how to tailor them such as:

- Ensuring that questions selected can be answered and providing advice to make the answers as objective as possible.
- Considering ordinal scale answers not just Yes/No answers.
- Aiming to limit the number of checklists items. We suggest between 6 and 12 items would be manageable.
- Working as a team to discuss, refine and agree appropriate quality checklist items.

REFERENCES.

[1] Brereton, O. P., Kitchenham, B.A. (2007) The Scope of EPIC Case Studies. EPIC technical Report EPIC-2007-01.
[2] D. Budgen. (2008) Supporting Novices undertaking Systematic Literature Reviews. EPIC Case Study Protocol No: CS001/07, May.
[3] Crombie, I.K. (1996) The Pocket Guide to Appraisal, BMJ Books, 1996.
[4] Fink, A. (2005) Conducting Research Literature Reviews. From the Internet to Paper, Sage Publication, Inc., 2005.
[5] Greenhalgh, Trisha. (2000) How to read a paper: The Basics of Evidence-Based Medicine. BMJ Books.
[6] Kitchenham B.A. and Charters S.M. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering, Version 2.3. EBSE Technical Report EBSE-2007-01, Keele University and Durham University, July 2007.
[7] Kitchenham B.A., Brereton, O. P., Budgen, D. and Li, Z. (2008) Results from Case Study 1 – Quality Checklists, EPIC Technical Report, EBSE 2008-009. http://www.dur.ac.uk/ebse/ebse-dev/tr.php.
[8] Pettigrew, M. and Roberts, H. (2005) Systematic Reviews in the Social Sciences: A Practical Guide, Blackwell Publishing, 2005.
[9] Yin, Robert K. (2003) Case Study Research: Design and Methods, 3rd Edition, Sage Publications.

Acknowledgements
This study was funded by the UK Engineering and Physical Sciences Research Council project EP/E046983/1. We thank Mahmood Niazi and Michael Goldstein for their help constructing the team-based quality checklist.

APPENDIX 1 QUALITY CHECKLIST EVALUATION FORM AND RESPONSES

1. Overall did the checklists provided in the guidelines help you create your quality checklist?:
   A great deal / Somewhat / Not at all

Responses
   DB: Somewhat
   OPB: A great deal
   ZL: A great deal

2. How do you judge the checklist in terms of the ease of adapting them to a specific SLR
   Very Simple /Reasonably simple / Somewhat difficult/ Very difficult

Responses
   DB: Somewhat difficult
   OPB: Reasonably simple
   JL: Reasonably simple

Do you have any suggestion to improve the adaptability/usability of the checklists?
DB: Hard to fit on to what is essentially a mapping study, maybe we need a separate table for mapping studies.
OPB: Barbara's suggestions for heading were very helpful[1]*.*
ZL: Some of the "wording" in the checklists could have been made clearer, e.g., "Was the study designed with these questions in mind?" – it was not very clear to me what do "these questions" refer to, therefore, I left it out from my list. The checklists assume that "population" are always people, e.g., "who was included?" and "who was excluded?". Can the "population" in SE be anything other than people? Would it be possible for the word "population" to be replaced by "subject"? For example, in unit testing, the number of faults detected can be used to indicate the effectiveness of a particular method. In this case, it is the testing methods or techniques that are being studied, rather than people.

3. How do you judge the applicability of the checklist items relative to the requirements of the Testing SLR
Excellent/ Satisfactory / Unsatisfactory
Response:
DB: Unsatisfactory
OPB: Satisfactory
ZL: Excellent

Do you have any suggestions to improve applicability?
DB: As above, much of the detail is misleading for a mapping study.
ZL: None. Perhaps more useful suggestions can be given as the SLR continues. I still feel that common criteria for comparing unit testing methods should be established, and added to one of the questions in the quality checklist.

4. How do you judge the completeness of the checklists relative to the requirements of the Testing SLR:
Excellent/ Satisfactory / Unsatisfactory
Responses
DB: Satisfactory
OPB: Excellent
ZL: Excellent

Do you have any suggestions to improve completeness?

5. Do you think that the checklist items meaningful for non-statisticians?
Yes / Somewhat / No
Responses
DB: Somewhat
OPB: Somewhat
ZL: Yes, most of them are.

Do you have any suggestions to improve the checklists that would help non-statisticians?
OPB: Perhaps link each check to relevant type of validity. I should have referred to Fink's book when I did this.
ZL: I find that some of the words used in the checklists have particular and important meanings, e.g., "population", "treatment", "blinding", "randomisation" in the design section, and many in the analysis section. Perhaps some simple explanations could have been added as texts before or after the table (or as footnotes).
6. Any other comments?
DB: Mainly a point that I have made before, that we tend to assume that users of the guidelines know about primary studies in some detail—maybe we need to make it more explicit that this knowledge is needed.
OPB: I lack statistics expertise and am left feeling that my list needs checking by an expert.

APPENDIX 2 QUALITY CHECKLISTS CONSTRUCTED DURING THE CASE STUDY

Table 4: Quality Checklists

C1 RA1 (Li) Checklist items	C3 DB (Budgen) Checklist items	C4 OPB (Brereton) Checklist items	C5 BAK (Kitchenham) Checklist items	AC Agreed Checklist items
Are the aims clearly stated?	Are the aims stated clearly?	Are the aims clearly stated?		
Do the study measures allow the questions to be answered?	Do the study measures allow the question to be answered?	Do study measures allow the questions to be answered?		
	Were the testing tasks randomly allocated?	Were treatments randomly allocated?	Is this a true experiment or a quasi experiment?	Was the allocation to treatment suitable given the experimental design and the number of subjects?

[1] *This involved adding columns to identify what part of the process the guideline related to and how it could be assessed.*

		Is the sample representative of the population to which the results will generalise?		
			Were seeded errors realistic?	
			Did the experimenters use a toy program/document as the testing object?	
	Was a control group used?			
	If a control was used, were the participants similar to the treatment group in terms of their testing skills and experience?			
Was the sample size justified?				
	Were the forms of testing clearly defined?	Is the technology (i.e. the testing approach) clearly defined?		
Are the variables used in the study adequately measured (i.e., are the variables likely to be valid and reliable)?	Are the variables used adequately measured?	Are variables used in the studies adequately measured (i.e. likely to be valid and reliable)?	Is the dependent variable valid?	Are the study measures valid?
Are the measures used in the study fully defined?	Are the measures used fully defined?	Are measures used in the studies fully defined?		
	Were any problems reported with the conduct of the study?			
Are the data collection methods adequately described?	Are the data collection methods adequately described?	Are data collection methods adequately described?		
	If two groups were being compared, were they treated similarly in the study?	If two groups are being compared, were they treated similarly within the study?		Were testers in each treatment group treated equivalently?
Are the study participants or observational units adequately described? For example, SE experience, type (student, practitioner, consultant), nationality, task experience and other relevant variables	Are the characteristics of the participants relevant to testing adequately described	Are the actual study participants or observational units adequately described (e.g. characteristics of units tested and testers)		
	Is the raw testing data described?	Were the basic data adequately described?		
			Were data preparation activities appropriate?	
			Was this a student-based experiment?	
				Has the analysis and interpretation accounted properly for any drop outs or missing values?
Are the statistical methods described?	Are the statistical methods described?	Are the statistical methods described?		
		Are the statistical	Did the method of	Is the analysis consistent

		methods justified?	data analysis conform with the experimental design?	with the experimental design?
	Are any scoring systems described?			
	Are potential confounding factors adequately controlled for purpose of analysis?			
Do the numbers add up across different tables and subgroups?	Do the numbers add up across tables?		Were related analysis results consistent?	
	If there were any differences between groups was any attempt made to control for these in the analysis	If different groups were different at the start of the study or treated differently during the study, was any attempt made to control these differences, either statistically or by matching?	If the experiment is a quasi-experiment, are additional design elements added to address possible bias	If no random allocation was used, was the design and or the analysis adjusted to cater for non-random allocation.
		If yes, was it successful?		
If yes, was statistical significance assessed?				
	If statistical tests were used to determine values, are the p values reported?			
			Was there evidence of fishing for results?	Were there an excessive number of statistical tests performed?
			Was any sensitivity analysis performed? Yes/No	Was sensitivity or stability analysis performed?
		Are the study questions answered?		
			Were the conclusions justified by the results?	
	What do the main findings mean in terms of unit testing?	What do the main findings mean?		
	Are negative findings presented?			
Is practical significance described?	If statistical tests were used, are the practical consequences of the outcomes discussed?			
				If no significant effect was detected, was the sample size large enough to have detected a reasonable effect size?
	How do results compare with previous results?			
How do the results add to the literature?				
	Are the consequences of any problems with the validity/reliability of measures explained?	Do researchers explain the consequences of any problems with validity/reliability of their measures?		