Can We Evaluate the Quality of Software Engineering Experiments?

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

Dag I.K. Sjøberg
Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, NO-0316 Oslo, Norway
Dag.Sjoberg@ifi.uio.no

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

David Budgen
School of Engineering and Computing Sciences, Durham University, Science Laboratories, Durham, DH1 3LE, UK
David.Budgen@durham.ac.uk

Tore Dybå
SINTEF, Trondheim and Department of Informatics, University of Oslo, Norway
Tore.Dyba@sintef.no

Martin Höst
Department of Computer Science, Lund University, SE-221 00 Lund, Sweden
Martin.Host@cs.lth.se

Dietmar Pfahl
Department of Informatics, University of Oslo, Norway and University of Calgary, Canada
dietmarp@ifi.uio.no

Per Runeson
Department of Computer Science, Lund University, SE-221 00 Lund, Sweden
Per.Runeson@cs.lth.se

ABSTRACT

Context: The authors wanted to assess whether the quality of published human-centric software engineering experiments was improving. This required a reliable means of assessing the quality of such experiments. **Aims**: The aims of the study were to confirm the usability of a quality evaluation checklist, determine how many reviewers were needed per paper that reports an experiment, and specify an appropriate process for evaluating quality. **Method**: With eight reviewers and four papers describing human-centric software engineering experiments, we used a quality checklist with nine questions. We conducted the study in two parts: the first was based on individual assessments and the second on collaborative evaluations. **Results**: The inter-rater reliability was poor for individual assessments but much better for joint evaluations. Four reviewers working in two pairs with discussion were more reliable than eight reviewers with no discussion. The sum of the nine criteria was more reliable than individual questions or a simple overall assessment. **Conclusions**: If quality evaluation is critical, more than two reviewers are required and a round of discussion is necessary. We advise using quality criteria and basing the final assessment on the sum of the aggregated criteria. The restricted number of papers used and the relatively extensive expertise of the reviewers limit our results. In addition, the results of the second part of the study could have been affected by removing a time restriction on the review as well as the consultation process.

ESEM'10, September 16-17, 2010, Bolzano-Bozen, Italy.

Categories and Subject Descriptors

D.2.0 [**Software Engineering**]: General

General Terms

Experimentation.

Keywords

Quality evaluation.

1. INTRODUCTION

We conducted the study reported herein because we wanted to investigate whether, given the increased number of guidelines and books on the topic, the standards of human-centric software engineering experiments had improved over the last decade. A prerequisite for such a study was to find a means of evaluating the quality of such experiments. We believed that as experienced researchers, we would have little difficulty in assessing the quality of human-centric experimental studies objectively. We were wrong. This paper describes our attempt to develop a procedure for evaluating the quality of software engineering experiments in terms of the number of assessors (i.e., judges)

needed to review each paper, the instrument and process by which quality can be assessed, and the process by which the assessments can be aggregated. Note that in this work we are in the same situation as ordinary reviewers of a journal or conference paper; we can only assess an experiment in terms of what can be inferred from the reporting of it in a paper.

In addition to our own interest, the issue of evaluating the quality of experiments is also of more general importance [22], particularly to Evidence-Based Software Engineering [14], [7]. This wider importance lies in the fact that the results of systematic literature reviews, which aggregate experimental evidence, have been shown to give different results if low-quality studies are omitted from the analysis. Low-quality studies, such as post hoc-correlation studies, sometimes favour a treatment, whereas high-quality studies, e.g., rigorously controlled field experiments, show no effect. This is the case for studies on the efficacy of homeopathy [21]. In the case of software engineering, Jørgensen and Moløkken-Østvold [8] point out that the original Chaos report should be omitted from any study investigating the incidence of project failures, because of the poor methodology used in the study.

From the viewpoint of undertaking systematic literature reviews, there have been several suggestions for quality checklists that can be used to evaluate the quality of empirical studies in software engineering. In particular, Dybå and Dingsøyr [5] developed a questionnaire that they used themselves in a study of agile methods [6] and that other researchers have adopted [2].

We decided to use Dybå and Dingsøyr's checklist and undertake a pilot study to determine the sufficient number of researchers necessary to obtain a reliable assessment of the quality of software experiments. We initially looked at the reliability of individual assessments and were dismayed by the poor level of agreement. Subsequently, we investigated the effect of allowing evaluators to discuss their assessments and provide a joint evaluation. The purpose of this paper is to alert researchers in software engineering to the practical problems of assessing the quality of experiments and to offer some advice on the best way to conduct quality assessments. The results may also be of interest to the editors of conferences and journals who are attempting to improve the quality of reviews.

The study we report in this paper addressed the following issues:

- How many judges are needed to obtain a reliable assessment of the quality of human-centric software engineering experiments?
- What is the best way to aggregate quality assessments from different judges; in particular, is a round of discussion better than using a simple median?
- Is using a quality checklist better than performing a simple overall assessment?

Note that this is an investigatory study, not a formal experiment; hence, we do not present formal hypotheses.

2. RELATED RESEARCH

Weller [23] produced an extensive review of studies that investigate peer review, covering 1,439 studies published between period 1945 and 1997. Bornmann [3] has written a review concentrating on research from 2000 up to the beginning of 2009 concerning three important issues for the peer review process: reliability (i.e., inter-rater agreement), fairness, and predictive validity.

These reviews show that there is a considerable body of literature on the topic of peer review. However, the majority of studies have looked at peer review of journal or conference papers (see, for example, [24], [20]) or the extent to which reviewers agree on whether to accept or reject research grant applications or research fellowships (see, for example, [16]).

Generally, researchers have found that reliability is poor. Bornmann [3] reports the results from 16 studies for which the Kappa or Intraclass correlations (ICC) "generally fall in the range from 0.2 to 0.4". He also refers to a meta-analysis currently under review that included 48 studies and found overall agreements of approximately 0.23 for ICC, 0.34 for the Pearson product moment correlation and 0.17 for Kappa [4]. Values of Kappa between 0 and 0.2 indicate only slight agreement. The only paper we found in the field of information science [24] also reported low levels of reliability in two conferences: one conference had kappa = -0.04, the other had kappa = 0.30.

Neff et al. [17] modelled the peer-review process, focussing on the editors' prescreening of submitted manuscripts and the number of referees used. Their model suggests that with respect to the number of reviewers, "the frequency of wrongful acceptance and wrongful rejection can be optimized at about eight referees". Looking at research proposals, Marsh et al. [16] refer to a study in which "it would require at least six assessors per proposal to achieve more acceptable reliability estimates of 0.71 (project) and 0.82 (researcher)".

However, in our case, we are not interested solely in a decision regarding acceptance or rejection, as is normal for journal papers and research proposals; we are interested in whether the use of a checklist leads to greater reliability. Several researchers have suggested that reliability can be improved by the use of checklists [19], [18]. Reporting on experiences of evaluating abstracts over a 4-year period, Poolman et al. [18] reported ICC values between 0.68 and 0.96 with only two of 13 values being less than 0.8 with between six and eight reviewers when the assessment was made on an aggregate of the individual criteria. Rowe et al. [19] reported a study on the acceptance of abstracts using a quality checklist. They found that changes to the guidelines for using the checklist that were made in response to criticism increased the reliability of the aggregate score from ICC = 0.36 to ICC = 0.49 with three reviewers. They noted that reviewers agreed less well on the individual criteria than on the sum of individual criteria and less well on subjective criteria than on objective criteria.

In the context of the criteria for quality that are used in systematic literature reviews, Kitchenham et al. [13] report the outcome of two different strategies they used to assess the quality of systematic literature reviews in software engineering using the DARE method, which is based on four criteria. They suggest that a process they referred to as "consensus and minority report" is more reliable than the median of three independent assessments or an assessment made based on two independent assessments and a discussion. The "consensus and minority report process" involved three researchers making individual assessments, followed by two researchers coming to a joint assessment and then comparing their joint assessment with the third review.

3. MATERIALS AND METHODS

This section discusses the checklist we used and the way in which the study was organised.

3.1 Quality Checklist Construction

One of us (DP) produced a revised set of criteria for determining quality that was based primarily on Dybå and Dingsøyr's checklist [5] but included some content from Kitchenham et al.'s checklist [12] and introduced ordinal-scale responses to the individual questions. This checklist was reviewed and revised by five of us (BAK, DS, TD, PR, DP) at a meeting in Oslo on 22 Feb 2009. Those of us that did not attend the Oslo meeting (MH, PB, DB) reviewed the quality checklist to assess:

- Whether the current checklist coincided with their subjective opinion of paper quality.
- Whether they understood each top-level question and the associated more detailed questions.
- Whether they felt confident that they would be able to answer the question.
- Whether there were any specific ambiguities, errors, or omissions.

After some discussion, the checklist was further refined. The final version of the checklist is shown in Table 1. Each question is answered on a 4-point scale where:

- "4 = Fully" means all questions listed in the "consider" column can be answered with "yes"
- "3 = Mostly" means the majority of all (but not all) questions listed in the "consider" column can be answered with "yes"
- "2 = Somewhat" means some (but the minority) of the questions listed in the "consider" column can be answered with "yes"
- "1 = Not at all" means none of the questions listed in the "consider" column can be answered with "yes"

However, we also recognized that the sub questions are not guaranteed to be complete and other issues may influence the answer for a specific study.

3.2 Quality Checklist Validation – Part I

After the final version of the checklist was agreed, we undertook the first part of the study in which we assessed the checklist for usability and consistency. We selected four papers from the set of human-centric experiments found by Kampenes [9]. The papers were A: [15]; B: [11]; C: [1]; D: [10]. All team members evaluated each paper independently, using the criteria for determining quality that are presented in Table 1, noting:

1. The answers to each quality question for each paper.
2. The time taken to evaluate each paper. We agreed to try to restrict ourselves to about 30 minutes per paper. This time limit was suggested by TD as a result of his experience using his checklist.
3. Any difficulties that arose using the quality checklist.
4. Whether the checklist-based evaluation of quality was consistent with their general view of the quality of each paper.
5. A subjective assessment of the overall quality of the papers, based on a 5-point ordinal scale: excellent (5), very good (4), acceptable (3), poor (2), and unacceptable (1). This variable was used to assess whether a simple overall assessment is as good as an assessment based on a number of different criteria.

To ensure that the papers were assessed in a different order (so that the analysis of how long it takes to evaluate the quality checklist would not be confounded with the learning process or the specific papers), the researchers were assigned at random to four different orders (1: A,B,C,D; 2: D,A,B,C; 3: C,D,B,A; 4: B,C,D,A), such that two researchers were assigned to each order.

The results of part I were intended to assess:

1. The reliability of the checklist items in terms of inter-rater agreement.
2. Whether the checklist appears to give a reasonable evaluation of paper quality.
3. Whether four independent reviewers are sufficient to obtain reliable results.
4. How much time each researcher would be likely to need for the full study.

The results suggested that we achieved a rather poor inter-rater reliability even with four judges, so we undertook part II of the study to investigate whether allowing judges to discuss their assessments would improve the reliability.

3.3 Quality Checklist Validation – Part II

In part II, we reread each of the papers individually, revised our initial assessments and added a rationale for each revised assessment. We did not place any limit on the time to be spent rereading each paper. After we had reviewed the papers again, we worked in pairs to make a joint evaluation. Allocation to pairs was not done at random, but was done in such a manner that each pair was different for each paper. As for part I, we answered each of the nine questions and gave an overall assessment of the paper.

Table 1. Quality Checklist

#	Question	Things to consider
Category: Questions on Aims		
1.	Do the authors clearly state the aims of the research?	*Do the authors state research questions, e.g., related to time-to-market, cost, product quality, process quality, developer productivity, and developer skills?* *Do the authors state hypotheses and their underlying theories?*
Category: Questions on Design, Data Collection, and Data Analysis		
2.	Do the authors describe the sample and experimental units (=experimental materials and participants as individuals or teams)?	*Do the authors explain how experimental units were defined and selected?* *Do the authors state to what degree the experimental units are representative?* *Do the authors explain why the experimental units they selected were the most appropriate for providing insight into the type of knowledge sought by the experiment?* *Do the authors report the sample size?*
3.	Do the authors describe the design of the experiment?	*Do the authors clearly describe the chosen design (blocking, within or between subject design, do treatments have levels)?* *Do the authors define/describe all treatments and all controls?*
4.	Do the authors describe the data collection procedures and define the measures?	*Are all measures clearly defined (e.g., scale, unit, counting rules)?* *Is the form of the data clear (e.g., tape recording, video material, notes, etc.)?* *Are quality control methods used to ensure consistency, completeness and accuracy of collected data?* *Do the authors report drop-outs?*
5.	Do the authors define the data analysis procedures?	*Do authors justify their choice / describe the procedures / provide references to descriptions of the procedures?* *Do the authors report significance levels and effect sizes?* *If outliers are mentioned and excluded from the analysis, is this justified?* *Do the authors report or give references to raw data and/or descriptive statistics?*
6.	Do the authors discuss potential experimenter bias?	*Were the authors the developers of some or all of the treatments? If yes, do the authors discuss the implications anywhere in the paper? (If the authors developed the treatments (or parts of them) without discussing the implications, the answer to question 6 is "not at all".)* *Was there random allocation to treatments?* *Was training and conduct equivalent for all treatment groups?* *Was there allocation concealment, i.e., did the researchers know to what treatment each subject was assigned?*
7.	Do the authors discuss the limitations of their study?	*Do the authors discuss external validity with respect to subjects, materials, and tasks?* *If the study was a quasi-experiment, do the authors discuss the design components that were used to address any study weaknesses?* *If the study used novel measures, is the construct validity of the measures discussed?*
Category: Questions on Study Outcome		
8.	Do the authors state the findings clearly?	*Do the authors present results clearly?* *Do the authors present conclusions clearly?* *Are the conclusions warranted by the results and are the connections between the results and conclusions presented clearly?* *Do the authors discuss their conclusions in relation to the original research questions?* *Are limitations of the study discussed explicitly?*
9.	Is there evidence that the E/QE can be used by other researchers / practitioners?	*Do the authors discuss whether or how the findings can be transferred to other populations, or consider other ways in which the research can be used?* *To what extent do authors interpret results in the context of other studies / the existing body of knowledge?*

4. METHODS OF DATA ANALYSIS

There is no well-specified way of assessing the reliability of *k* judges evaluating *n* target objects in *m* dimensions where each dimension is an ordinal-scale subjective variable taking values 1 to 4. In this paper, we report the results of using the Kappa statistic and an *ad hoc* statistic, *Diff1*.

4.1 The Kappa Statistic

The Kappa measure assumes a nominal scale evaluation variable, usually a single variable (i.e., a variable of the type accept/reject, yes/no) although it is possible to have more than binary categories. The basic formula for Kappa applies to two judges as follows:

$$\text{Kappa} = (\text{PO}-\text{PC})/(1-\text{PC})$$

Where PO = proportion of the target values that are the same for two judges

PC = the probability that an assessment would have been the same by chance.

In our case, we have nine criteria to be assessed by each judge on each paper on a 4-point scale (i.e., 1, 2, 3, 4). Thus, PO is the number of agreements divided by nine and PC = 0.25 (i.e., we would expect 2.25 agreements by chance).

However, Kappa ignores the extent to which judges *almost* agree. Furthermore, if we want to see the effect of averaging the assessments for two judges to see the effect of combining evaluations, we will probably cause Kappa to have a reduced value when we compare paired evaluations because we will have values such as 1.5, 2.5, and 3.5 leading to seven separate "categories" such that a value of 3.5 and a value of 4 will be considered a disagreement although they are close when considered as ordinal-scale measures. The usual method of assessing Kappa is to use the interpretation scale shown in Table 2. However, a statistical test can be based on the empirical distribution of data that conforms to the null hypothesis, i.e., a set of evaluations of nine 4-point ordinal-scale criteria made at random (see Section 4.3.2).

Another major problem with the use of the Kappa metric is that it is commonly used to assess the reliability of two judges who are assessing multiple targets, *not* two judges who are assessing multiple criteria that pertain to the same target, which is what we were doing in our study.

Table 2. Interpretation scale for Kappa

Kappa value	Interpretations
<0	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

4.2 Diff1

As an alternative to Kappa and to balance Kappa's inherent problems with ordinal-scale assessments, we constructed a new metric, which we refer to as the Diff1 statistic, which is the number of times that two judges differ by more than one point on the nine criteria, i.e., a value of 1 for Diff1 meant there was only one occasion out of a possible nine when a pair of assessments differed by more than one point.

Diff1 is of particular relevance to situations where there are several criteria per target and the criteria are numerically equivalent (in our case, 4-point ordinal-scale measures). However, it does not consider multiple judges.

Diff1 does not have a statistical test, but it is possible to obtain the null distribution of the statistic empirically. Diff1 can be used for assessments that include values such as 1.5, 2.5, and 3.5, i.e., for assessments that are aggregated by averaging. Its main disadvantage is that it is a coarse statistic when the number of points on the ordinal scale is small (as it is in our case), so it may not be possible to construct confidence limits on the values obtained at a specifically required alpha level.

4.3 Establishing Baseline Distributions of the Test Statistics

This section identifies the empirical distribution of Diff1 and Kappa for the case of nine criteria assuming that assessments of each criterion were random. We use the empirical distribution to identify whether it is likely that the agreement between judges was better than we would expect if all judgements were made at random.

4.3.1 Diff1

The overall distribution of the Diff1 metric for random evaluations is summarized in Table 3. For Diff1, we were interested in low values of the statistic for our evaluation data and take a value of 0 or 1 to indicate an acceptable agreement.

Table 3. Distribution of Diff1 for random assessments

Statistic	Values
Observations	1000
Mean	3.349
Std. Dev	1.4620
1% Percentile	0
5% Percentile	1
10% Percentile	2
25% Percentile	2
50% Percentile	3

4.3.2 Kappa

The distribution of the Kappa statistic for pairs of random evaluations is summarized in Table 4. In this case, we take a Kappa value $\kappa>0.26$ to indicate acceptable agreement.

Table 4. Distribution of Kappa for random pairs of assessments of nine variables

Statistic	Values
Observations	1000
Mean	-0.00163
Std. Dev	0.1929
75% Percentile	0.1111
90% Percentile	0.2593
95% Percentile	0.2593
99% Percentile	0.5555

5. DATA ANALYSIS AND RESULTS

This section presents the analysis of the data. First, we present Kappa and Diff1 values for each paper separately. Then we discuss the effect of aggregating results.

5.1 Results for Individual Papers

For part I of the study, we calculated the reliability statistics for each of the 28 possible ways of pairing the eight judges. Table 5 summarises the results, which indicate that all papers show good agreement among judges with respect to Diff1 but only paper C shows good agreement with respect to Kappa. Equivalent results for part II of the study where pairs of judges provided a joint assessment are shown in Table 6. In this case, there were four joint evaluations and, therefore, six ways in which the joint evaluations could be compared. These results for the joint evaluations show good agreement for both Diff1 and Kappa with the exception of paper B for the Kappa results.

5.2 Composite Assessments

The median assessment of the eight individual assessments and median of the four joint assessments are shown in Table 7. It shows the results for each question, the sum of median assessments and the subjective overall assessment for the paper (an additional question scored on a five point ordinal scale). The agreement is remarkable for all four papers and for all questions. However, we observe that overall assessment suggests papers B and D are of equivalent quality, whereas the sum of the nine quality questions suggests that that paper B is better than paper D.

However, we wanted to know whether we can assess the quality of papers with fewer than eight reviewers per paper and whether or not allowing judges to have a round of discussion is useful. These issues cannot be addressed with a formal statistical analysis, but we present the effect of various strategies for aggregating assessments in Table 8. This analysis is restricted to the sum of the nine questions.

The strategies used in Table 8 are:

- Median of eight independent evaluations.
- The average of any two independent evaluations. There are 28 possible ways of aggregating two evaluations from eight.
- The average of any four independent evaluations. There are a total of 120 different ways in which four assessments can be aggregated. We selected 30 such combinations at random and found the median for each question.
- Median of four paired evaluations.
- The average of any two-paired evaluations. There are six possible ways of aggregating four evaluations. We found the average of each pair. The sum was the total of the average value for questions 1 to 9.

In each case, the sum was calculated as the total of the aggregated values for questions 1 to 9, and the overall assessment.

Table 8 shows the minimum, maximum, and range of values for each option. The results show:

- In three of the four cases, pair-wise aggregation of individual assessments was less reliable than evaluations based on two judges with a round of discussion.
- In two of the four cases, aggregation based on four individual judges was less reliable than evaluations based on two judges with a round of discussion. In one case, the median of four individuals was better, and in the final case, the results were the same. This suggests four individual assessors are broadly equivalent to two assessors who discuss their assessments.
- In three of the four cases, aggregation based on the average of two consensus evaluations was better than any other aggregation strategy. In the fourth case, it was as good as the consensus evaluation...

Table 5. Summary of evaluation results part I

Measure	Paper A		Paper B		Paper C		Paper D	
	Diff1	Kappa	Diff1	Kappa	Diff1	Kappa	Diff1	Kappa
Number acceptable agreements of 28 (Diff1≤1, κ>0.259)	19	10	26	9	23	21	22	7
Average	1.179	0.339	0.571	0.238	0.75	0.444	1.00	0.228
Std Dev.	1.249	.2035	0.742	0.244	0.7515	0.169	1.018	0.203

Table 6. Summary of evaluation results part II

Measure	Paper A		Paper B		Paper C		Paper D	
	Diff1	Kappa	Diff1	Kappa	Diff1	Kappa	Diff1	Kappa
Number acceptable agreements of 6 (Diff1≤1, κ>0.259)	6	6	6	2	5	5	6	6
Average	0.33	0.65	0.667	0.235	0.5	0.481	0	0.630
Std Dev.	0.52	0.153	0.516	0.218	0.837	0.182	0	0.156

Table 7. Median assessments for eight individual judges and four pairs of judges

Paper	Judges	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Sum of Questions	Subjective Overall Assessment
A	8 individuals	2.5	2	2	2	2	1	1	2	2	16.5	2
A	4 pairs	2	2	2.5	2	1.5	1	1	2	2	16	2
A	**Diff**	**0.5**	**0**	**-0.5**	**0**	**0.5**	**0**	**0**	**0**	**0**	**0.5**	**0**
B	8 individuals	4	3	4	3.5	3.5	3	3	3	2.5	29.5	4
B	4 pairs	4	3	4	3.5	3.5	2.5	3	3.5	2.5	29.5	4
B	**Diff**	**0**	**0**	**0**	**0**	**0**	**0.5**	**0**	**-0.5**	**0**	**0**	**0**
C	8 individuals	4	3	4	3.5	4	3	3.5	4	2.5	31.5	4.5
C	4 pairs	4	3	4	3.5	4	3	3.5	4	2.5	31.5	4.5
C	**Diff**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
D	8 individuals	4	3	3	3	3	2	3	3	3	27	4
D	4 pairs	4	3	3	3	3	2	3	3	2	26	4
D	**Diff**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**0**

Table 8. Effect of various aggregation strategies on the sum of the criteria

Evaluation source	Paper A			Paper B			Paper C			Paper D		
	Min	Max	Rng	Min	Max	Rng	Min	Max	Rng	Min	Max	Rng
Eight independent evaluations	12	21	9	26	34	8	29	36	7	24	32	8
28 pair-wise averages	12.5	21	8.5	27	32	5	29	33.5	4.5	25	29	4
30 random combinations of four (median)	14	20	6	27.5	31	3.5	29.5	33	3.5	26	30	4
Four joint evaluations	15	19	4	26	32	6	30	32	2	24	28	4
Six combinations of joint evaluations (average)	15	17.5	2.5	27.5	31	3.5	30	32	2	24.5	27.	2.5

6. DISCUSSION AND CONCLUSIONS

The inter-rater agreement statistics confirm that:

- the reliability obtained for individual assessments was relatively poor, and
- the reliability obtained by pairs of judges with a round of discussion was generally quite good.

Our results suggest that good agreement can be achieved with eight judges, whether or not there is any discussion among the judges. However, we also achieved almost as good a consensus by using four judges, with the judges working in pairs to arrive at two independent consensus evaluations that were then averaged. The present study did not allow us to determine whether three judges are sufficient if there are two rounds of consensus making, as proposed by Kitchenham et al. [13]. The results also show that using the sum of the criteria to rank papers was better than using a simple 5-point scale assessment. In particular, the overall assessment was unable to distinguish between papers B and D, whereas the sum of the individual criteria made it clear that paper B outscored paper D.

The main limitation of the study is the number of targets. With only four papers we cannot be sure how well our results will generalise to our target of all human-based software engineering experiments. For example, the papers do not constitute an homogenous sample. In particular, paper A is rather different from the other papers because the human-based experiment presented in the paper was only a small part of a wider evaluation exercise. Overall, paper A was good, but the human-based experiment was weak. Further, even if the sample were homogenous, it might not be representative. Other limitations are that we, as a group of researchers, have extensive experience of empirical software engineering, so our results may be better than those that would have been obtained by a random selection of researchers. In addition, in part II of the study, we not only had a period of discussion among pairs of judges, but we also reviewed each paper for a second time without a time restriction. Thus, the more favourable results with respect to reliability may be due not only to the discussion, but also to the additional time spent on reviewing the paper.

Finally, an ultimate goal of the research community is to conduct experiments of high quality. To reach this goal, we must be able to evaluate the quality of experiments. To achieve (as much as

possible) consensus on what is high quality, one needs to agree on a conceptual definition of quality as well as a set of operational quality criteria. Using a checklist is one way of implementing a set of operational quality criteria. The checklist that we used was a modified version of a checklist that had already been developed and used by others. Using another checklist with another set of criteria for determining quality, might have given other results.

Generally our results are consistent with other reports. For example, as suggested by Neff et al. [17], we found that the aggregated results from eight reviewers gave very reliable results.

In the related papers that we were able to find, we found no discussion of the use of criteria to determine the quality of studies that are to be used in meta-analysis. This is an important issue and the reason we undertook our study in the first place. Our results suggest strongly that a discussion among judges is needed when two judges are being used. This indication is in agreement with the advice given in most guidelines on how to conduct systematic reviews. On the other hand, our results suggest that if the quality of papers is a critical part of the review, two judges and a discussion might not be sufficient to obtain a reliable assessment of the quality of the studies. Our study showed that four judges acting in pairs obtained very high levels of reliability.

REFERENCES

[1] Abrahão, S. and Poels, G. 2007. Experimental evaluation of an object-oriented function. Information and Software Technology, 49(4), pp 366–380.

[2] Afzal, W., Torkar, R. and Feldt, R. 2009. A systematic review of search-based testing for non-functional system properties. Information and Software Technology, 51(6), pp 959–976.

[3] Bornmann, L. 2011. Scientific Peer Review. Annual Review of Information Science and Technology, 45, in press.

[4] Bornmann, L., Mutz, R. and Daniel, D.-D. A reliability-generalization study of journal peer reviews – a multi-level analysis of inter-rater reliability and its determinants, submitted.

[5] Dybå, T. and Dingsøyr, T. 2008a. Strength of Evidence in Systematic Reviews in Software Engineering, ESEM'2008, pp 178–187.

[6] Dybå, T. and Dingsøyr, T. 2008b. Empirical studies of agile software development: A systematic review. Information and Software Technology, 50(9-10), pp 833–859.

[7] Dybå, T., Kitchenham, B.A. and Jørgensen, M. 2005. Evidence-based Software Engineering for Practitioners, IEEE Software, 22 (1), pp 58–65.

[8] Jørgensen, M. and Moløkken-Østvold, K. 2000. Impact of effort estimates on software project work? A review of the 1994 Chaos report. Information and Software Technology, 48(4), pp 297–301.

[9] Kampenes, V.B. 2007. Quality of Design Analysis and Reporting of Software Engineering Experiments. A Systematic Review. PhD Thesis, Dept. Informatics, University of Oslo.

[10] Karahasanović, A. Levine, A.K. and Thomas, R. 2007. Comprehension strategies and difficulties in maintaining object-oriented systems: An explorative study. Journal of Systems and Software, 80, pp 1541–1559.

[11] Karlsson, L., Thelin, T., Regnell, B., Berander, P. and Wohlin, C. 2007. Pair-wise comparisons versus planning game partitioning – experts on requirements prioritisation techniques. Empirical Software Engineering, 12, pp 3–33.

[12] Kitchenham, B.A., Brereton, O.P., Budgen, D. and Li, Z. 2009. An evaluation of quality checklist proposals – A participant observer case study. EASE'09, BCS eWic.

[13] Kitchenham, B., Brereton, P., Turner, M., Niazi, M., Linkman, S., Pretorius, R. and Budgen, D. Refining the systematic literature review process – Two observer-participant case studies, accepted for publication in Empirical Software Engineering.

[14] Kitchenham, B.A., Dybå, T. and Jørgensen, M. 2004 Evidence-based Software Engineering. Proceedings of the 26th International Conference on Software Engineering, (ICSE '04), IEEE Computer Society, Washington DC, USA, pp 273–281.

[15] Liu, H. and Tan, H.B.K. 2008. Testing input validation in Web applications through automated model recovery. Journal of Systems and Software, 81, pp 222–233

[16] Marsh, H.W., Jayasinghe, U.W. and Bond, N.W. 2008. Improving the Peer-Review Process for Grant Application. Reliability, Validity, Bias and Generalizability. American Psychologist, 63(3), pp 160–168.

[17] Neff, B.D. and Olden, J.D. 2006. Is Peer Review a Game of Chance. BioScience, 56(4), pp 333–340.

[18] Poolman, R.W., Keijser, L.C., de Waal Malefijt, M.C., Blankevoort, L., Farrokhyar, F., Bhandari, M. 2007. Reviewer agreement in scoring 419 abstracts for scientific orthopedics meetings. Acta Orthopaedica, 78(2) pp 278–284.

[19] Rowe, B.H., Strome, T.L. Spooner, C., Blitz, S., Grafstein, E. and Worster, A. 2006. Reviewer agreement trends from four years of electronic submissions of conference abstract. BMC Medical Research Methodology, 6(14).

[20] Schultz, D.M. 2009. Are three heads better than two? How the number of reviewers and editor behaviour affect the rejection rate. Scientometrics, Springer. doi:10.1007/s11192-009-0084-0.

[21] Shang, A., Huwwiler-Müntener, K., Nartney, L., Jüni, P., Dörig, S., Pwesner, D. and Egger, M. 2005. Are the clinical effects of homeopathy placebo effects? Comparative study of placebo-controlled trials of homeopathy and allopathy. Lancet, 366 (9487), pp 726–732.

[22] Sjøberg, D.I.K., Dybå T. and Jørgensen. M. 2007. The Future of Empirical Methods in Software Engineering Research, In: Future of Software Engineering (FOSE '07), ed. by Briand L. and Wolf A., pp. 358–378, IEEE-CS Press.

[23] Weller, A.C. 2002. Editorial Peer Review: Its Strengths and Weaknesses. Medford, NJ, USA. Information Today, Inc.

[24] Wood, M., Roberts, M. and Howell, B. 2004. The reliability of peer reviews of papers on information systems. Journal of Information Science, 30(1), pp 2–11.