Information and Software Technology 54 (2012) 804–819

Contents lists available at SciVerse ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Three empirical studies on the agreement of reviewers about the quality of software engineering experiments

Barbara Ann Kitchenham [a,*], Dag I.K. Sjøberg [b], Tore Dybå [b,c], Dietmar Pfahl [b,d], Pearl Brereton [a], David Budgen [e], Martin Höst [d], Per Runeson [d]

[a] School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK
[b] Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, NO-0316 Oslo, Norway
[c] SINTEF, P.O. Box 4760 Sluppen, Trondheim, Norway
[d] Department of Computer Science, Lund University, SE-221 00 Lund, Sweden
[e] School of Engineering and Computing Sciences, Durham University, Science Laboratories, Durham DH1 3LE, UK

ARTICLE INFO

Article history:
Available online 7 December 2011

Keywords:
Quality evaluation
Empirical studies
Human-intensive experiments
Experimentation
Software engineering

ABSTRACT

Context: During systematic literature reviews it is necessary to assess the quality of empirical papers. Current guidelines suggest that two researchers should independently apply a quality checklist and any disagreements must be resolved. However, there is little empirical evidence concerning the effectiveness of these guidelines.
Aims: This paper investigates the three techniques that can be used to improve the reliability (i.e. the consensus among reviewers) of quality assessments, specifically, the number of reviewers, the use of a set of evaluation criteria and consultation among reviewers. We undertook a series of studies to investigate these factors.
Method: Two studies involved four research papers and eight reviewers using a quality checklist with nine questions. The first study was based on individual assessments, the second study on joint assessments with a period of inter-rater discussion. A third more formal randomised block experiment involved 48 reviewers assessing two of the papers used previously in teams of one, two and three persons to assess the impact of discussion among teams of different size using the evaluations of the "teams" of one person as a control.
Results: For the first two studies, the inter-rater reliability was poor for individual assessments, but better for joint evaluations. However, the results of the third study contradicted the results of Study 2. Inter-rater reliability was poor for all groups but worse for teams of two or three than for individuals.
Conclusions: When performing quality assessments for systematic literature reviews, we recommend using three independent reviewers and adopting the median assessment. A quality checklist seems useful but it is difficult to ensure that the checklist is both appropriate and understood by reviewers. Furthermore, future experiments should ensure participants are given more time to understand the quality checklist and to evaluate the research papers.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

As part of a long-term project to assess trends in the quality of human-intensive software engineering experiments and quasi-experiments, we are interested in how reliable assessments of the quality of research papers are in the field of software engineering. Although our interest arose from a specific situation, the quality of empirical studies is an important issue in its own right, since an assessment of quality is required when performing systematic literature reviews aimed at aggregating empirical results by meta-analysis or tabulation.

In the following sections we provide some context for our paper by discussing:

- why quality evaluation in the context of systematic reviews is important by providing examples of problems that can arise when quality is ignored;
- what the current recommendations are for performing quality evaluations;
- the checklist we based our evaluation criteria on and the reasons for choosing it;
- the goals of the studies described in this paper;
- the structure of the paper.

* Corresponding author. Tel.: +44 1782 733979; fax: +44 1782 734268.
E-mail addresses: B.A.Kitchenham@cs.keele.ac.uk (B.A. Kitchenham), Dag.Sjoberg@ifi.uio.no (D.I.K. Sjøberg), Tore.Dyba@sintef.no (T. Dybå), Dietmar.Pfahl@cs.lth.se (D. Pfahl), O.P.Brereton@cs.keele.ac.uk (P. Brereton), David.Budgen@durham.ac.uk (D. Budgen), Martin.Host@cs.lth.se (M. Höst), Per.Runeson@cs.lth.se (P. Runeson).

0950-5849/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2011.11.008

1.1. The importance of quality evaluation

Quality evaluation is recommended because systematic literature reviews in the medical domain have been shown to give different results if low-quality studies are omitted from the analysis. A systematic review of 159 systematic reviews in medicine found that "in the majority of meta-analyses exclusion of trials with inadequate or unclear concealment[1] and trials without double-blinding led to a change towards less beneficial treatment effect, which was often substantial" [12]. In a recent systematic review of homoeopathy, including low-quality studies, such as simplistic quasi-experiments (i.e. asking whether someone feels better after taking the treatment with no control group) suggested that homoeopathy performs well, whereas high-quality studies, e.g., rigorously controlled field experiments with blinding and controls show no significant effect [38]. In addition, observational studies suggested that beta carotene and vitamin A protect against lung cancer, and that vitamin E protects against heart disease. However, in both cases subsequent high-quality randomised controlled trials found different results. In the case of protection against lung cancer, the use of beta carotene and vitamin A actually appeared harmful [33]. In the case of vitamin E, it simply appeared to have no affect on heart disease [42]. In the case of software engineering, Jørgensen and Moløkken-Østvold [16] point out that the original Chaos Report looking at the rate of software failures used an extremely poor methodology. This implies that it should be omitted from any systematic review of the rate of software failure. Although there are examples from medical studies where observational studies and randomised controlled trials actually agree, the extent to which we can expect agreement is unknown [14], so a systematic review, or a meta-analysis based on a systematic review, needs to look for consistency or inconsistency among results from studies of different quality.

1.2. Current procedures for quality evaluation

The general advice for quality assessment for systematic literature reviews is to use two reviewers, a quality checklist and a mechanism to address disagreements among reviewers [34]. As a preliminary to our planned study of quality trends in empirical software engineering studies, we undertook a pilot study that we thought would confirm that we could obtain reliable assessments of quality using a checklist. Since we were all experienced researchers, we believed that we would have little difficulty in assessing the quality of human-intensive experimental studies objectively; it transpired that we were wrong. As a result, we undertook two further studies to investigate how best to organise the evaluation of the quality of human-intensive software engineering experiments. This paper describes our attempt to develop a procedure for quality evaluation in terms of the number of assessors (often referred to as *judges*) needed to review each paper, the process by which quality can be assessed (i.e. whether or not a period of discussion among judges is necessary), and the process by which the assessments can be aggregated (i.e. whether assessments prepared jointly by judges are better than simple arithmetic aggregation of independent assessments).

1.3. Using checklists for quality evaluation

From the viewpoint of undertaking systematic literature reviews in software engineering, there have been several suggestions for constructing quality checklists that can be used to evaluate the quality of empirical studies in software engineering. In particular, Dybå and Dingsøyr [9] developed a questionnaire that they used in their study of agile methods [10] and that other researchers have since adopted e.g. [2,3,6].

Since Dybå and Dingsøyr's checklist had been published and used by several different researchers performing systematic reviews, we decided to use it as the basis of our checklist and to undertake a pilot study to determine the number of judges sufficient or necessary to obtain a reliable assessment of the quality of software experiments. Initially, we thought we were validating our quality checklist and identifying the optimum number of judges, however, when we looked at the reliability of individual assessments, we were dismayed by the poor level of agreement. Subsequently, we investigated the effect of allowing judges to discuss their assessments and provide a joint evaluation. Finally, in a third study we further investigated the impact of discussions among judges by comparing the assessments made by individuals with assessments made by teams of two or three persons.

1.4. Goals

The purpose of this paper is to alert researchers in software engineering to the practical problems of assessing the quality of experiments in the context of systematic literature reviews and to offer some advice on the best way to conduct such assessments. The results may also be of interest to the editors of conferences and journals who are attempting to improve the quality of reviews or the reviewing process.

The studies we report in this paper addressed the following research questions:

- RQ1: How many judges are needed to obtain a reliable assessment of the quality of human-intensive software engineering experiments and quasi-experiments?
- RQ2: 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?
- RQ3: Is using a quality checklist better than performing a simple overall assessment?

Our first two studies were investigatory, rather than formal experiments; hence, we do not present formal hypotheses for them. The third study was designed more formally with the aim of determining whether discussion within teams of two or three persons leads to more reliable assessments of human-intensive experiments and quasi-experiments than do individual assessments. Based on the first two studies, we assumed that assessments based on discussion between either two or three persons would lead to better reliability (i.e. inter-rater agreement, see Section 3) than assessments by individuals, and we expected the reliability of assessments from three-person teams to outperform assessments based on two-person teams. Study 3 was intended to address RQ2 and to test our expectations more formally.

1.5. Paper structure and contents

Section 2 discusses related research. Section 3 describes the metrics that are used to measure inter-rater agreement and the materials we used in our studies (i.e. the quality evaluation questionnaire and the research papers). Section 4 describes the methods we adopted in each of the three studies. We present our results in Section 5 and discuss them in Section 6.

An earlier version of this article was presented at the ESEM 2010 conference [24]. The ESEM paper was based on Studies 1 and 2 alone. The data analysis in this paper has also been updated to use the ordinal scale Kappa metric to measure reliability [8] rather than the less appropriate basic Kappa reliability [7]. We have also used the Intra-Class Correlation [40] to investigate

[1] I.e. concealment of the treatment to which individual participants were allocated.

whether a simple overall assessment of quality is as good (or better) than a checklist-based assessment.

2. Related research

In the context of evaluating the quality of primary studies included in systematic literature reviews and meta-analyses, most of the available research regarding quality assessment is concerned with assessing the quality of randomised controlled trials in medicine. We found no empirical research that was related to the number of judges or the method of aggregating results from different judges.

Quality assessment of primary studies in systematic reviews and meta-analyses is based primarily on three specific issues:

- Whether or not the study used random allocation of treatment to participants.
- Whether the study was single-blind (i.e. the allocation of participant to treatment was concealed from participants) or double-blind (i.e. the allocation of participant to treatment was concealed from both the participants and the experimenters).
- Whether dropouts were analysed on an *intention-to-treat* basis or not. Intention-to-treat is an analysis based on the initial treatment intent, not on the treatment eventually administered. It is meant to reduce bias due to participants dropping out of an experiment, changing treatment group after assignment, or being ineligible for participation in the study. For example, in cross-over experiments where participants use one technique in a laboratory session followed by a second laboratory session using another technique, some participants might not attend the second session of the experiment, resulting in drop-outs. In software engineering experiments involving agile methods, participants might be assigned to use the test-first method but actually adopt a standard test-after approach, resulting in some participants using the wrong treatment. In studies where study materials are in English and participants are non-native English speakers, some participants may not have sufficiently good English skills to participate, but this is not recognised until the experiment has finished, meaning that some ineligible participants were included.

There is some doubt as to the validity of quality evaluation scales using a broad range of quality criteria. Jüni et al. performed a meta-analysis of 17 medical trials comparing low-molecular weight heptarin (LMWH) with standard heptarin for prevention of postoperative thrombosis using 25 different quality evaluation scales [17]. They found that for 6 of the scales, high quality studies found no difference whereas low quality studies favoured LMWH. Seven scales showed the opposite, with high quality studies favouring LMHW and low quality studies showing no effect; for the remaining 12 studies the effects were the same for high and low quality studies. Interestingly, when using two judges, Jüni et al. found excellent inter-rater reliability for each scale based on the Intra-Class Correlation. They conclude that the use of a summary score to identify high quality is problematic and that relevant methodological aspects should be assessed individually, with their influence on effect sizes being explored individually. This presents a problem for human-intensive software engineering experiments, since although software engineering experimenters do use random allocation, blinding of participants is usually impossible and blinding of experimenters is only rarely possible, and we are unaware of any software engineering studies using an intention-to-treat analysis for handling dropouts. So, on the one hand we cannot rely on the three validated quality criteria, but on the other hand evaluation scales using a broad range of quality criteria are problematic. In domains other than medicine, researchers have developed many evaluation scales using multiple quality criteria, but we found no example of them being evaluated.

In the context of software engineering, Kitchenham et al. [23] report the outcome of several different strategies they used to assess the quality of systematic literature reviews in software engineering using four criteria. They suggest that a process that they refer to as "consensus and minority report" is more reliable than either the median of three independent assessments or basing an assessment 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 reviewer's assessment. They also noted that the simple median of three independent assessments was almost as reliable as the consensus and minority report process.

In searches with Google Scholar, we found no other study that investigated the optimum number of judges for quality assessments or the level of reliability that might be expected for systematic reviews. Most research that is related to assessor reliability comes from peer review studies which, although not aimed at quality evaluation of systematic reviews, can give some indication of problems associated with assessing research papers. Weller [43] produced an extensive review of studies that investigated peer review, covering 1439 studies published between 1945 and 1997. The majority of studies have looked at either peer review of journal and conference papers (see, for example, [30,35–37]), or at the extent to which reviewers agree on whether to accept or reject research grant applications or research fellowships (see, for example, [28]).

Generally, researchers have found that reliability of peer review is poor. *Reliability* in this context is a measure of inter-rater agreement. It is usually measured in terms of the Kappa statistics or the Intra-Class Correlation coefficient (ICC). These metrics are described in Section 3. Bornmann [4] reports the results from 16 studies for which the Kappa or ICC "generally fall in the range from 0.2 to 0.4", which is regarded as *fair* (see Table 2). 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 addressing peer review that we found in the field of information science [44] also reported low levels of reliability in reviewing performed for two conferences: one conference had Kappa = −0.04, the other had Kappa = 0.30.

Neff and Olden [30] modelled the peer-review process, focussing on the editors' pre-screening 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 optimised at about eight referees". Looking at research proposals, Marsh et al. [28] 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)".

Several researchers have suggested that using a checklist increases reliability [35,36]. Reporting on experiences of evaluating abstracts over a 4-year period, Poolman et al. [35] 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. The assessments were made on an aggregate of several individual evaluation criteria. Rowe et al. [36] reported a study on the evaluation 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 when using 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.

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? (If random allocation was possible and not done, there is a possibility of biased assignment to treatment)* *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 Experiment/Quasi-Experiment 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?*

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

3. Materials and measures

This section discusses the checklist that we used, the way in which the agreement among judges was evaluated and the selection of research papers used in the studies. The materials and measures described in this section were used in all three studies to increase the comparability of the results.

3.1. Quality checklist construction

We decided to base our checklist on the checklist proposed by Dybå and Dingsøyr [9] for two reasons:

1. It was based on a checklist that is widely used by researchers in other disciplines.
2. It is currently being adopted by software engineering researchers performing systematic reviews.

Dybå and Dingsøyr developed the questionnaire for use in their systematic review of agile methods [10]. It was based on checklists used by the Critical Appraisal Skills Programme (CASP),[2] in particular that used for assessing the quality of qualitative research, and good practice for conducting empirical research in software engineering [22]. It covers the three main quality issues that need to be considered when appraising primary studies of any design:

[2] See http://www.phru.nhs.uk.

- Rigour: Whether the key research methods adopted a thorough and appropriate approach.
- Credibility: Whether the finding are well-presented and meaningful.
- Relevance: Whether the finding are useful to the software industry and the research community.

This checklist was reviewed and revised by five of us (BAK, DS, TD, PR, DP)[3] at a meeting in Oslo on 22 February 2009. The rationale behind the revisions were:

Removal of unnecessary questions: In particular since we were restricting ourselves to studies that had already been classified as formal experiments or quasi-experiments, we removed three superfluous questions. One question asking whether the paper was a research paper; the second asked whether there was a control group; the third asked whether the context of the study was adequately described (a question of far more importance in a qualitative study).

Refocusing of questions: Many of the questions asked whether some aspect of the study was appropriate in some way. After considerable discussion, it was decided to refocus the questions on whether the topic had been discussed rather than whether the methodological decisions addressed by the question were appropriate or not. The reason for this decision was to reduce the subjectivity involved in answering questions. However, this decision introduced the risk that an author could describe in detail a method that was completely inappropriate for the aims of the study and would score well on our questionnaire.

Introducing experiment-related questions: We introduced one question asking whether the authors described their participants and materials. This question took the place of the question that asked about context. We also introduced a question asking whether authors had described the limitations of their study.

We also reviewed the sub-questions associated with each question proposed by Dybå and Dingsøyr [9,10]. We added and revised sub-questions in order to address issues of importance in experiments e.g. random allocation to treatment and concealed allocation (i.e. not letting the experimenters know which participant was in which treatment group), or issues important in software engineering experiments, such as whether the paper authors were developers of the technique being investigated. However, we recognised that the set of questions was not guaranteed to be complete, that other issues might influence the answer to the question in a specific study, and that in some situations, some of the sub-questions might be irrelevant.

Finally we decided to change from using a yes/no answer to each main question and adopted a four point ordinal scale where:

- "4 = Fully" means all questions listed in the "Things to consider" column can be answered with an unqualified "yes" and are relevant and all other relevant issues were fully described.
- "3 = Mostly" means the majority of all (but not all) questions listed in the "consider" column can be answered with an unqualified "yes" or some other relevant issues were either not addressed, or not fully addressed.
- "2 = Somewhat" means some (but the minority) of the questions listed in the "consider" column can be answered with unqualified "yes" and/or other relevant issues were either not addressed at all or not fully addressed.
- "1 = Not at all" means all the questions listed in the "consider" column can be answered with an unqualified "no" and other relevant issues were not addressed.

Thus, the sub-questions were advisory and the judge had to make-up his/her own mind how to answer the main question.

One of us (DP) produced the revised version of the checklist which was then reviewed by those of us that did not attend the Oslo meeting, i.e. MH, DB and PB, who were asked 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. The total score for a paper for a specific judge (or team of judges) was calculated by summing the numeric values for each of the nine main quality questions.

3.2. Reliability metrics and methods of data analysis

There is no well-accepted way of assessing the reliability of k judges (i.e. reviewers) evaluating n target objects (i.e. papers reporting experiments) in m dimensions (i.e. using a quality checklist with m questions) where each dimension is an ordinal-scale subjective variable taking values 1–4. In this paper, we report the results of using the weighted Kappa statistic [8] and the Intra-Class Correlation (ICC) [40]. Note that in our previous paper [24] we only used the simple unweighted Kappa, which is not really appropriate for ordinal scale metrics, and we did not use the ICC metric at all.

3.2.1. The weighted Kappa statistic

The basic Kappa statistic assumes a comparison based upon using 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 (κ), as applied to two judges is then as follows:

$$\kappa = \frac{PO - PC}{1 - PC} \quad (1)$$

where PO is the proportion of the target values that are the same for two judges; PC is 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, using an ordinal 4-point scale (i.e. 1, 2, 3, 4). Hence, we decided to use the weighted Kappa statistic rather than the basic Kappa, where the weighted Kappa κ_W is defined as:

$$\kappa_W = 1 - \frac{q_o}{q_e} \quad (2)$$

where q_o is the *observed weighted disagreement* among two judges and q_e is e*xpected weighted disagreement* among two judges. In this case we used the weights to penalise large disagreements. We used a linear weighted scale so that a disagreement of 0 points (i.e. agreement) was given a weight of zero, a disagreement of one point was given a weight of 1, a disagreement of two was given a weight of 2 and a disagreement of three was given a weight of 3.

The usual method of assessing Kappa (whether or not it is weighted) is to use the interpretation scale shown in Table 2. However, a statistical test can be based on the empirical distribution of

[3] We use the initials to designate a specific author.

Table 3
Intra-Class Correlation ANOVA.

Source of variation	df	Mean Square = Sum of squares/df
Between targets	$(n-1)$	BMS
Within targets	$n(k-1)$	WMS
Between judges	$(k-1)$	JMS
Residual	$(n-1)(k-1)$	EMS

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 3.2.3).

One problem with the use of the weighted 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.

3.2.2. Intra-Class Correlation

The Intra-Class Correlation (ICC) measure is used to assess ordinal scale ratings that have been made by multiple judges across multiple targets [40]. The ICC is based on an analysis of variance that considers the *between-target variation* and the *within-target variation*, which is split into a *between-judges* component and a *residual* term (see Table 3).

Shrout and Fleis [40] point out there are three different types of ICC measure. If the judges are different for each target (and assumed to be selected at random from a group of possible judges), the ANOVA collapses into a simple one-way ANOVA with a between-target element and a within-target element that cannot be further decomposed. In this case the ICC is calculated as:

$$\mathrm{ICC}(1,1) = (\mathrm{BMS} - \mathrm{WMS})/(\mathrm{BMS} + (k-1) \times \mathrm{WMS}) \quad (3)$$

This is meant to be the expected reliability of a single judge. If we want to assess the ICC of the set of k judges, we use the following formula:

$$\mathrm{ICC}(1,k) = (\mathrm{BMS} - \mathrm{WMS})/\mathrm{BMS} \quad (4)$$

The use of ICC(1,1) is appropriate for Study 2, where assessments were made by a pair of judges, and for each target (i.e. research paper) we have a different pair of judges (see Section 4.2).

If the same judges assess each target (as was the case for Study 1, see Section 4.1), the data must be analysed as two-way ANOVA (with *judge* as one factor and *target* as the other factor), but there are two separate cases. Firstly, there is the case where the judges are assumed to be drawn at random from a set of available judges. In this case the effect of the judges is assumed to be due to a random error term. This leads to a two-way random effects ANOVA model based on ICC(2,1) and ICC(2,k) reliability measures. Secondly, there is the case where the judges are a fixed set of individuals. In this case, the effect of the judges is assumed to be a fixed effect, with a constraint that the sum of the effects is zero. This is a two-way mixed effects model because the judge effect is fixed and the target effect is random. This leads to the ICC(3,1) and ICC(3,k) reliability measures.

For Study 1, the judges were fixed – since our goal was to see how well we, as a specific group of researchers, assessed research papers. Thus we used the following formulae to calculate ICC:

$$\mathrm{ICC}(3,1) = (\mathrm{BMS} - \mathrm{EMS})/(\mathrm{BMS} + (k-1) \times \mathrm{EMS}) \quad (5)$$

$$\mathrm{ICC}(3,k) = (\mathrm{BMS} - \mathrm{EMS})/\mathrm{BMS} \quad (6)$$

The advantages of the ICC measure is that we can use the appropriate F test from the relevant ANOVA to see whether the intra-class reliability is better than chance. In addition, the same "interpretation scale" is used for ICC as for Kappa (see Table 2).

The problem with the ICC is that it does not consider the reliability of judges across a set of variables, thus we can only judge the reliability separately for each question in our scale. However, it is particularly useful for assessing the reliability both of the overall subjective assessment of the paper, and also of the total score (i.e. the sum of the assessment for each checklist item).

Another issue is that in Study 2, we have only the equivalent of four judges for each target. Thus, we use different variants of the ICC measure for Study 1 and Study 2. In Study 2 we use the coarsest measure, i.e. ICC(1,1), with only 12 degrees of freedom for the WMS term. Generally the ICC measure of reliability needs more judges and/or more targets to give reliable results. This means that the intra-class reliability measures will not be fully comparable between Studies 1 and 2. In Study 3, we employed a randomised block design (see Section 4.3) and the ICC formula does not apply.

3.2.3. Obtaining null distribution of weighted Kappa

We planned to perform statistical analysis of the Kappa values by comparing the Kappa values obtained from our study with the null distribution of Kappa values obtained from random evaluation. In order to obtain the null distribution of the weighted Kappa, we generated 1000 vectors (v_i), where each v_i included 9 variables and each variable took on the values 1, 2, 3, 4 at random. The vectors were numbered 1–1000 and the weighted Kappa values were calculated from 1000 pairs of vectors such that v_i was compared with v_{i+1} – with the sole exception that v_{1000} was compared with v_1. Thus, each vector was involved in two comparisons. The upper 95 percentile of the null distribution weighted Kappa was 0.422 and the upper 99 percentile was 0.556.

3.3. Selection of research papers

In order to undertake our study of inter-rater reliability, we needed to select some suitable research papers to evaluate. Since our long term goal was to investigate trends in the quality of software experiments and quasi-experiments, we wanted to select papers that were representative of our intended sample. Sjøberg et al. identified 103 research papers reporting human-intensive experiments and quasi-experiments in the time period 1993–2002 [41]. These represent a sample from the set of papers that we were interested in studying, but as they had already been studied in depth by two of us (DS and TD), it was felt that using some of those papers might have biased our study.

We decided to look for other papers representative of our intended population but published more recently. Fortunately, Kampenes, who was a co-author of [41], undertook a short study, as part of her PhD research, to compare published experiments and quasi-experiments found in 2007 with those found in the time period 1993–2002 [18]. She found 8 papers from a manual search of 4 journals (2 from EMSE, 5 from JSS, 0 from TSE and 1 from IST). These papers represented a sample of the papers we were interested in that had been identified by an independent researcher with experience of identifying human intensive experiments and quasi-experiments. We selected 4 of these 8 papers at random to include in our first study. The selected papers were labelled as A [27]; B [20]; C [1]; D [19].

4. Study design

This section describes the design of the three empirical studies.

4.1. Study 1: Inter-rater reliability

4.1.1. Goals

After the final version of the checklist was agreed, we undertook Study 1 in which we assessed the checklist for inter-rate reliability. This study was intended to assess:

1. The reliability of the checklist items in terms of inter-rater agreement (RQ1 and RQ2).
2. Whether the checklist appears to give a reasonable evaluation of paper quality (RQ3).
3. How many independent reviewers are sufficient to obtain reliable results (RQ1).
4. How much time each researcher would be likely to need for the planned wider study.

4.1.2. Experimental units

The main experimental units in this study are the research papers, the judges and the quality evaluations produced by the judges. In analyses, based on ICC, paper and judge are treated as separate factors in the analysis of variance. In analyses based on the weighted Kappa, reliability is assessed among pairs of judges separately for each paper.

4.1.3. Tasks

In this study each of the 8 co-authors of this paper acted as a judge and evaluated each paper independently, using the criteria for determining quality that are presented in Table 1, noting:

1. The answers to each main question from the quality checklist for each paper.
2. The time taken to evaluate each paper (including reading time and answering the checklist questions). We agreed to try to restrict ourselves to about 30 min per paper. This schedule was suggested by TD as a result of his experience using his checklist. However, 30 min was intended as a guideline not a strict limit and each researcher took the time that they needed and kept a record of how long they actually took.
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.

4.1.4. Design

To ensure 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 assessment of the papers occurred in a different order, we needed to ensure that the researchers were assigned at random to different orders for performing the assessment. There are two Latin Square designs that deal with 3 factors each of 4 levels (in this case Paper, Judge, Treatment Order). We chose one of the two designs (i.e. 1: A,B,C,D; 2: D,A,B,C; 3: C,D,B,A; 4: B,C,D,A) and replicated it to cater for 8 judges.

4.1.5. Analysis

The weighted Kappa results for all possible pairs of individuals for a specific paper was compared with a null distribution to assess whether the inter-rater reliability was better than random chance. The ICC was used to assess the extent to which evaluations of the same paper were consistent compared with the difference among paper. This was done for each quality checklist question and the sum of the numerical values assigned to each question.

4.1.6. Procedure

Each judge evaluated each paper and completed the quality questionnaire. Once he/she was happy with his/her evaluation, the time was recorded and the evaluation form sent to BK for aggregation and analysis.

4.2. Study 2: Investigation of the value of discussion among judges

4.2.1. Goals

In Study 2, we investigated the impact of pairs of reviewers adding a period of discussion to their assessment process. The study addresses RQ1 and RQ2 by providing information about the reliability obtained by pairs of judges after a period discussion with those obtained from individuals and from simple arithmetic aggregations (obtained from Study 1). Like Study 1, it also addresses RQ3 by comparing the reliability obtained from using a score derived by summing the numerical code for the answer to each of the 9 quality evaluation questions, with that obtained from a subjective overall assessment.

4.2.2. Experimental units

The main experimental units in this study were the joint evaluation and the research papers.

4.2.3. Tasks

In this study, each paper was again reviewed by each judge, but this time each judge was paired with another judge to obtain a joint evaluation. Thus, we obtained 4 joint evaluations of each paper.

4.2.4. Design

Allocation to pairs was not done at random, but was done in such a manner that the pairing was different for each paper.

4.2.5. Analysis

The analysis was similar to that of Study 1, but based on the paired evaluations.

4.2.6. Procedure

We re-read 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 re-reading each paper. After we had reviewed the papers again and revised (if necessary) the preliminary assessment, each pair exchanged their revised assessments and worked together through e-mail to make a joint evaluation. We also answered each of the nine quality related questions and gave an overall assessment of the paper, as for Study 1. Again, no time limit was set on the process.

Interaction between the pairs was done entirely by e-mail (even when researchers worked at the same establishment), we did not rely on face-to-face meetings. Thus, the process had some aspects of a Delphi process [5], in that we worked together to define the checklist (i.e. Table 1); our initial contributions were collected as answers to the main questions in our checklist, with comments associated with each answer; we commented on our own and our partners' answers; and we did not (apart from the initial checklist design) rely on face-to-face meetings. There were of course significant differences from the Delphi process, i.e. there was no panel director, and our answers were not anonymous.

Once the judges agreed upon their joint evaluation, the results were sent to BK for aggregation and analysis.

4.2.7. Design limitations

A limitation of Study 2 is that the same set of judges reviewed the same set of papers for a second time. Thus any improvement in the results might not arise from the discussion among pairs of

judges but alternatively might be due to the increased familiarisation of the judges with the specific research papers. Thus, to gain a better evaluation of the impact of discussion and number of judges, we undertook a third study.

4.3. Study 3: The number of judges per paper when discussion is permitted

4.3.1. Goal

The goal of Study 3 was to investigate the impact of discussion among judges and the impact of using two or three person teams. It addressed RQ2 by providing information comparing the evaluation obtained from individuals with those obtained from teams of two people and teams of three people.

4.3.2. Experimental units

In Study 3, we used a completely different set of judges who were organised into three treatment groups. One treatment group comprised 8 individuals, one treatment group comprised 8 two-person teams and the final treatment group comprised 9 three-person teams (in total 51 judges). Thus the "unit" of analysis was a team (albeit 8 "teams" were made up of a single person). One three-person team failed to complete the checklist, so they were excluded from the analysis leaving us with 8 three-person teams, and a total of 48 judges.

The judges were volunteers who were attending the ISERN meeting in Bolzano, Italy, 2010. They were, therefore, interested in, and generally experienced with, human-intensive software engineering experiments. Furthermore, they were the type of researchers who might be expected to perform systematic reviews of empirical studies. However, they ranged from senior academics with many years of experience to postgraduate students with more limited experience. They volunteered to take part in the study after one of the co-authors of this paper (DS) gave a brief overview of the rationale and goals of the study.

4.3.3. Tasks

Within each treatment group, four of the individuals/teams were asked to assess paper C and four were asked to assess paper D. We chose to use papers C and D because:

1. We wanted to use the same papers as previously, to ensure the results of Study 3 were comparable with Studies 1 and 2.
2. We had limited time (only 1 h during the ISERN meeting could be allocated to the experiment) and limited participants, so could only use at most two papers. For that reason we excluded paper A which was rather atypical of human-intensive experiments, and also excluded paper B, which was the longest paper.

The teams/individuals were allowed 1 h to complete their task, which involved reading the paper and completing the assessment form (i.e. the checklist shown in Table 1 together with questions asking about their overall assessment and the time they spent on the task, as well as an additional space asking them to make any comments they wanted about the experiment).

4.3.4. Hypotheses

If R_{TX} is the average inter-rater reliability of quality assessments made by teams of size X (where $X = 1, 2$, or 3), our null hypothesis is that:

- H0: $R_{T1} \geqslant R_{T2}$ and $R_{T1} \geqslant R_{T3}$ i.e. reliability is not improved by working in teams, that is, assessments made by single judges are as reliable as, or more reliable than, assessments made by teams of two or three judges working together.

Our alternative hypothesis is:

- H1: $R_{T1} < R_{T2}$ and $R_{T2} < R_{T3}$ i.e. reliability is improved by team working and the more people in the team the better the reliability, that is, assessments made by teams of two judges are more reliable than assessments made by individual judges, and assessments made by teams of three judges are more reliable than assessments made by teams of two judges.

As in Studies 1 and 2, we measured inter-reliability by calculating the weighted Kappa of assessments made by all possible pairs of "judges" (where a "judge" was either an individual or a team of two or three persons) from the same treatment group.

4.3.5. Design

To test these hypotheses statistically, it is necessary to use a two-way analysis of variance (with one factor corresponding to team size, and a blocking variable corresponding to the specific papers being evaluated) to determine the within-treatment residual variance, then to use t-tests of the average team size effect to test the specific hypotheses. As it happened, detailed statistical analysis was unnecessary (see Section 5.2).

The basic design of the experiment was a replicated randomised block design. The blocking factor was the research paper (C or D); the treatment factor was the team size (1, 2, or 3 persons). Based on the team structure (and excluding the one team that dropped out of the study), the design was completely balanced. Individuals were assigned at random to team/individual treatment groups (assignments were prepared in sealed envelopes which were allocated at random to the participants), and within each treatment group, teams/individuals were assigned at random to each paper (whilst constraining the numbers assigned to each paper to be equal).

4.3.6. Analysis

The hypotheses we wanted to test require three separate t-tests. First we need to test whether the reliability of two person teams is significantly greater than the reliability of individuals, then whether the reliability of three person tests is significantly greater than the reliability of individuals. Together these two tests are used to see whether or not teams perform better than individuals. Finally, if teams are better than individuals, we can test whether three person teams perform better than two person teams.

The availability of timing information also allowed us to consider the relative costs of using simple aggregates of multiple individual assessments compared with including a discussion element.

4.3.7. Power analysis

The study included 48 participants, but the "unit" of analysis was a team of which we had 24 (albeit 8 "teams" were made up of a single person). Furthermore, the hypotheses we wanted to test required three separate t-tests. Both factors (i.e. the number of "units" and the number of tests) have an impact on the power of the experiment. It is usual, when performing multiple tests, to decrease the level of significance of individual tests. In our case, the level of significance for a specific test would be 0.017 to achieve an overall level of significance of 0.05, 0.037 to achieve an overall level of significance of 0.10. In addition, the sample size for each individual test is 16. Looking at the tables provided by Dybå et al. [11], it is clear that we have a low power experiment. Even using an overall level of significance of 0.10, we would only be able to detect a large effect with a power of 0.70.

4.3.8. Procedure

The conduct of the study was overseen by two of the co-authors (DP and DS) and observed by two more of the co-authors (TD and

Table 4
Median assessments for eight individual judges and four pairs of judges.

Paper and Study	Judges	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Total score	Subjective overall assessment
S1_A	8 individuals	2.5	2	2	2	2	1	1	2	2	16.5	2
S2_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
S1_B	8 individuals	4	3	4	3.5	3.5	3	3	3	2.5	29.5	4
S2_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
S1_C	8 individuals	4	3	4	3.5	4	3	3.5	4	2.5	31.5	4.5
S2_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
S1_D	8 individuals	4	3	3	3	3	2	3	3	3	27	4
S2_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 5
Intra-class correlations.

Criterion	Study 1			Study 2		
	ICC(3,1) Reliability of Individual judge	ICC(3,8) Reliability of 8 Judges	*p*-value	ICC(1,1) Reliability of one pair of judges	ICC(1,4) Reliability of 4 pairs of judges	*p*-value
Overall assessment	0.599 (Moderate)	0.923	0.0001	0.714 (Substantial)	0.909	0.0009
Q1	0.639 (Substantial)	0.934	<0.00001	0.868 (Almost Perfect)	0.963	<0.00001
Q2	0.466 (Moderate)	0.875	0.0010	0.429 (Moderate)	0.750	0.0346
Q3	0.745 (Substantial)	0.959	<0.00001	0.867 (Almost Perfect)	0.963	<0.00001
Q4	0.456 (Moderate)	0.870	0.0012	0.513 (Moderate)	0.808	0.0156
Q5	0.412 (Moderate)	0.849	0.0026	0.650 (Substantial)	0.882	0.0028
Q6	0.556 (Moderate)	0.909	0.0001	0.596 (Moderate)	0.855	0.0059
Q7	0.844 (Almost perfect)	0.977	<0.00001	0.748 (Substantial)	0.922	0.0005
Q8	0.579 (Moderate)	0.917	0.0001	0.833 (Almost perfect)	0.952	<0.00001
Q9	0.181 (Slight)	0.639	0.067 (n.s.)	0.143 (Slight)	0.400	0.227 (n.s)
Total score (Q1–Q9)	0.893 (Almost perfect)	0.985	<0.00001	0.923 (Almost perfect)	0.980	<0.00001

Table 6
Number of judges to achieve required level of ICC (Study 1).

Criterion	ICC $\geqslant$ 0.75	ICC $\geqslant$ 0.8
Total score	2	2
Overall assessment	10	14
Q1	9	11
Q2	22	30
Q3	5	6
Q4	24	32
Q5	33	44
Q6	13	17
Q7	3	3
Q8	12	15
Q9	n/a	n/a

PR). Individuals were asked to read their assigned paper and complete an assessment, while the teams of two or three persons were asked to read the paper individually and then discuss the paper together to produce a joint assessment. However, in practice some teams almost immediately engaged in team discussion, while others spent longer on individual assessments before turning to team discussions. Teams and individuals were free to leave as soon as they had completed their task and submitted an assessment form to the study supervisors.

5. Data analysis and results

This section presents the analysis of the data. The raw data for the three studies is shown in Tables 7–12 at the end of the paper. First, we present the combined results for Studies 1 and 2. Then we present the results for Study 3.

5.1. Results for Studies 1 and 2

Initial results for Studies 1 and 2 have already been presented in [24]. However, for this paper the analyses have been changed to use the more appropriate weighted Kappa and ICC reliability metrics. The analysis of the aggregation strategies (see Section 5.1.2) has not been changed but is included for completeness.

5.1.1. Descriptive statistics

The median assessment of the eight individual assessments and median of the four joint assessments are shown in Table 4. Table 4 shows the results for each question, the total score (i.e. the sum of the median assessments), and the median of the subjective overall assessment for the paper (the additional question scored on a five point ordinal scale). The level of agreement is remarkable for all four papers and for all questions. However, the overall assessment suggests that papers B and D are of equivalent quality, whereas the sum of the nine quality questions suggests that paper B is marginally better than paper D.

Clearly, good reliability can be obtained by using a large number of judges but the goals of Study 1 and Study 2 were to investigate whether we could assess the quality of papers with substantially fewer than eight reviewers per paper, and whether or not allowing judges to have a round of discussion is useful. These issues are discussed in the following sections.

5.1.2. Reliability

For Study 1, there were eight assessments for each paper. Since the weighted Kappa statistic is based on pairs of assessments, and there are 28 possible ways of comparing the eight assessments (i.e. ${}^{2}C_{8} = 8!/[6! \times 2!]$), we obtained 28 values for weighted Kappa. For

Table 7
Study 1 and 2 paper A raw data.

Judge	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Overall	Total score
Study 1											
BAK	3	2	3	4	3	1	1	2	2	2	21
PB	2	1	2	3	3	1	1	4	2	4	19
MH	3	3	3	3	3	1	1	3	1	4	21
DP	2	2	2	1	1	1	1	2	1	2	13
PR	2	2	1	1	1	1	1	1	2	1	12
TD	3	2	2	2	2	1	1	2	2	2	17
DS	3	2	2	2	2	1	1	2	2	2	17
DB	2	2	3	1	1	1	1	2	3	2	16
Study 2											
DS&DB	2	2	2	2	1	1	1	2	2	2	15
BAK&PB	2	2	3	3	3	1	1	2	2	2	19
MH&DP	2	3	3	2	1	1	1	2	1	2	16
PR&TD	2	2	2	1	2	1	1	2	2	2	15

Table 8
Study 1 and 2 paper B raw data.

Judge	Q1N	Q2N	Q3N	Q4N	Q5N	Q6N	Q7N	Q8N	Q9N	Overall	Total score
Study 1											
BAK	4	4	4	4	4	4	4	4	3	5	34
PB	4	2	4	4	4	4	4	4	2	4	30
MH	3	3	3	4	4	3	3	3	2	4	28
DP	3	2	4	3	2	1	4	4	3	4	26
PR	4	2	3	4	4	2	3	3	4	4	29
TD	4	3	4	3	3	3	3	3	2	4	28
DS	4	3	4	4	3	3	3	3	3	4	29
DB	4	3	4	3	4	2	3	3	2	5	28
Study 2											
BAK&PR	4	3	4	4	3	2	3	3	4	3	30
DP&DS	3	2	4	3	3	1	3	4	3	4	26
MH&PB	4	3	4	4	4	3	4	4	2	4	32
TD&DB	4	3	4	3	4	3	3	3	2	4	29

Table 9
Study 1 and 2 paper C raw data.

Judge	Q1N	Q2N	Q3N	Q4N	Q5N	Q6N	Q7N	Q8N	Q9N	Overall	Total score
Study 1											
BAK	4	4	4	4	4	4	4	4	4	5	36
PB	4	3	4	3	3	3	4	4	2	4	30
MH	4	3	4	3	4	2	3	4	2	4	29
DP	4	4	4	4	4	2	4	4	2	5	32
PR	4	3	4	4	4	1	3	4	4	4	31
TD	4	3	4	4	4	3	3	4	3	5	32
DS	4	3	4	3	2	3	3	4	3	3	29
DB	3	3	4	3	4	3	4	4	2	5	30
Study 2											
BAK&DS	4	3	4	3	3	4	2	4	3	3	30
PB&TD	4	3	4	4	4	3	4	4	2	5	32
PR&DP	4	3	4	4	4	2	4	4	3	5	32
MH&DB	4	3	4	3	4	3	3	4	2	4	30

Study 2, there were four joint assessments. Since there are six possible ways of comparing four assessments, we obtained six values for weighted Kappa.

Fig. 1 shows the distribution of the weighted Kappa values for each paper and for each study compared with the null distribution of weighted Kappa (derived from 1000 random vectors of 9 four-point ordinal scale independent variables). The different box plots are characterised by study as Study 1 (S1) or Study 2 (S2) and paper (A–D). The dotted horizontal line shows the upper 95 percentile for the null distribution which has the value 0.42 (see Section 3.2.3). The results for the different papers are separated by a vertical dashed line.

Fig. 1 shows that the weighted Kappa values obtained in Study 1 were better than chance and were usually in the range *slight* to *fair*, although some were non-existent and some were *substantial* or *almost perfect*. That is on average the reliability of the judges is significantly better than random, but for some pairs of judges, for every paper, agreement was no better than random. With respect to RQ1, the results suggest that two judges are insufficient to obtain reliable evaluations. This was why we undertook Study 2 to investigate whether allowing judges to discuss their assessments would improve the reliability.

For Study 2, the reliability improved substantially for the papers A and D, slightly for paper B (the lower tail increased but the upper

Table 10
Study 1 and 2 paper D raw data.

Judge	Q1N	Q2N	Q3N	Q4N	Q5N	Q6N	Q7N	Q8N	Q9N	Overall	Total score
Study 1											
BAK	4	4	3	3	2	2	2	4	4	3	28
PB	4	4	4	4	3	2	3	4	4	4	32
MH	3	3	3	3	3	4	4	3	2	4	28
DP	4	2	3	3	3	2	3	2	2	4	24
PR	4	3	3	4	4	2	3	3	3	4	29
TD	4	3	3	3	4	2	3	3	3	4	28
DS	3	3	3	3	3	3	3	2	3	3	26
DB	4	3	4	3	4	2	4	4	1	4	29
Study 2											
BAK&DP	4	3	3	3	3	2	2	2	2	3	24
PB&DS	4	3	3	4	3	2	3	3	3	4	28
MH&TD	3	3	3	3	3	2	3	3	2	4	25
PR&DB	4	3	3	3	3	2	3	3	2	4	26

Table 11
Study 3 paper C raw data.

Group	Overall	Q1N	Q2N	Q3N	Q4N	Q5N	Q6N	Q7N	Q8N	Q9N	Total score	Time (min)
1	4	4	3	4	2	3	2.5	3	4	3	28.5	15
1	4	3	3	3	3	3	3	3	3	2	26	20
1	5	4	4	4	4	4	4	4	4	3	35	35
1	5	4	4	4	3	3	3	4	3	3	31	30
2	4	3	4	4	2	3	2	3	3	3	27	n/a
2	2	2	3	3	3	3	2	2	3	1	22	45
2	n/a	4	4	3	2	3	3	3	3	2	27	n/a
2	4	4	3	4	3	4	3	4	3	2	30	60
3	4	4	4	4	4	4	4	3	3	3	33	52
3	4	3	3	2	3	3	3	3	3	2	25	50
3	5	4	3	4	4	4	4	3	4	4	34	65
3	3	3	2	3	3	2	3	4	2	2	24	60

Table 12
Paper D study 3 raw data.

Group	Overall	Q1N	Q2N	Q3N	Q4N	Q5N	Q6N	Q7N	Q8N	Q9N	Total score	Time (min)
1	2	3	3	2	2	2	2	4	2	1	21	45
1	3	4	3	2	3	3	2	2	4	1	24	50
1	4	3	4	4	3	2	3	3	3	2	27	40
1	4	2	3	2	3	1	2	3	3	2	21	30
2	5	4	4	4	4	3	4	4	4	3	34	60
2	1	3	4	4	2	3	4	2	2	1	25	20
2	2	2	3	2	3	1	2	3	2	1	19	55
2	5	4	4	4	3	4	3	4	4	4	34	48
3	3	3	3	3	3	2	2	3	3	2	24	n/a
3	1	3	2	1	1.5	1	2	1.5	2	1	15	55
3	3	3	4	4	4	3	3	2	3	1	27	n/a
3	4	3	3	3	3	3	2	3	3	3	26	25–30

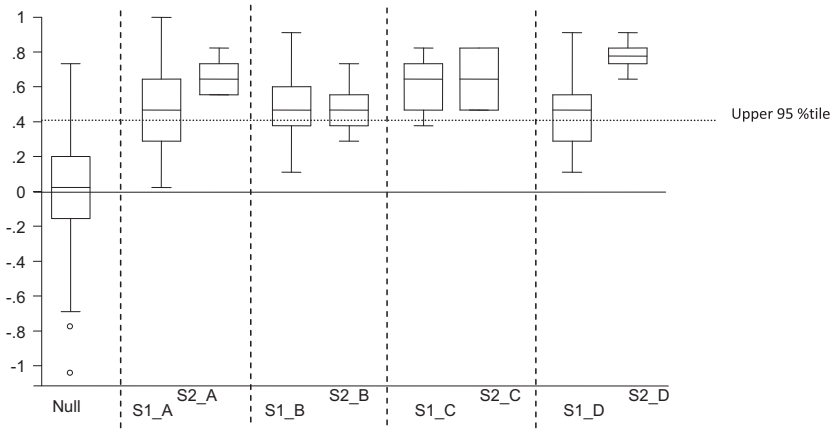

Fig. 1. Comparison of the weighted Kappa null distribution and the observed distribution of Kappa values for each paper for Studies 1 and 2.

tail decreased), and hardly changed for paper C. Again on average, pairs of joint assessments were significantly better than random and in this case only paper B exhibited any pairs of joint assessments that were not better than random. With respect to RQ1, these results suggest that using two reviewers is sufficient providing that there is a period of discussion among the reviewers. These results were similar to those found using the unweighted Kappa [24].

The Intra-Class Correlation (ICC) is based on multiple judges assessing multiple objects using a single assessment criterion measured on an ordinal scale. The calculated ICC values for each question, the total score, which is the sum of the numerical value of the nine individual questions, and the overall subjective assessment, are shown in Table 5. The equations for the different variants of the ICC are given in Section 3.2.2.

These results suggest that averaged across the four papers, the reliability achieved by an individual judge is better than the weighted Kappa values suggest. Excluding question 9, which shows no significant reliability for Studies 1 or 2, the reliability for an individual judge is *moderate* or better for all individual questions. However, this may be more a reflection of difference between the papers rather than good reliability among the judges.

For the overall subjective assessment the ICC value is *moderate* for Study 1 and *substantial* for Study 2 while the reliability of the total score is *almost perfect* in both studies. For 6 of the 8 significant questions and for the total score and overall subjective assessment, the reliability achieved in Study 2 is greater than the reliability achieved in Study 1. This suggests that using a pair of judges to achieve a joint evaluation has some benefit. However, the ICC is calculated differently for the two studies, so the values are not directly comparable.

The ICC analysis also makes it possible to estimate the number of judges needed to achieve a required reliability level. The values for Study 1 are shown in Table 6. This suggests that *almost perfect* reliability can be achieved for the total score with only two judges, whereas many more are needed for the overall subjective assessment or the individual checklist items. However, these results must be treated with some caution since they are based on applying an analysis of variance to non-normally distributed ordinal-scale variables. In fact, only the results for the total score are likely to be robust because the Central Limit Theorem confirms that the distribution of a variable that is the sum of n other identically distributed variables is approximately Normal, irrespective of the individual variable's original distribution. We do not present a similar analysis for Study 2 because there are only four assessments available for each paper, which is too few for an accurate analysis.

Thus, these results for the ICC analysis suggest that the answer to RQ1 is that two judges are sufficient with or without discussion as long as the total quality score obtained by summing the answers to the individual questions is used. With respect to RQ3, these results also suggest that using a quality checklist to provide a total score is preferable to using a simple subjective assessment.

5.1.3. Timing for Study 1

The time taken to perform the Study 1 evaluations is not directly related to the research questions raised in this paper but we found that the average time for reading and assessing a paper to be 29.7 min, with a standard deviation of 11.4, and with little differences in the timing for the different papers (average time in minutes for paper A = 28.1, paper B = 29.6, paper C = 29.1 and paper D = 31.7). The relatively consistent times for each paper provided the justification for believing that we could run Study 3 in a 1 h session.

5.1.4. Aggregation strategies

The ICC analysis has addressed the issue of whether to use simple arithmetic aggregations of results or joint aggregations after discussion, but that analysis has some limitations, so we present various strategies for aggregating assessments for papers A–D in Figs. 2–5, respectively. This analysis is restricted to the total score derived from the sum of the nine questions.

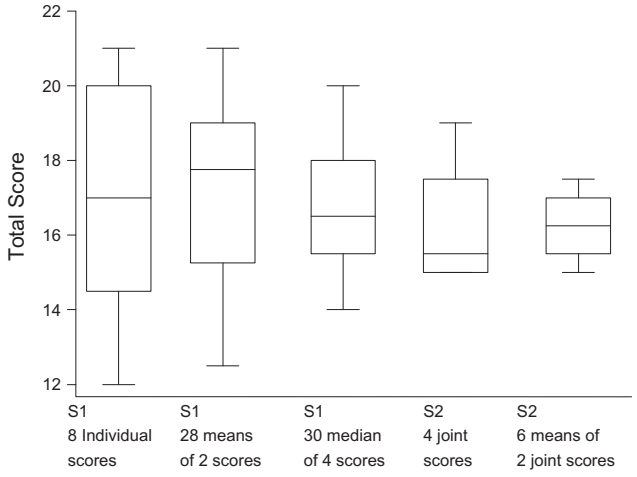

Fig. 2. Paper A distribution of total score for different aggregation strategies.

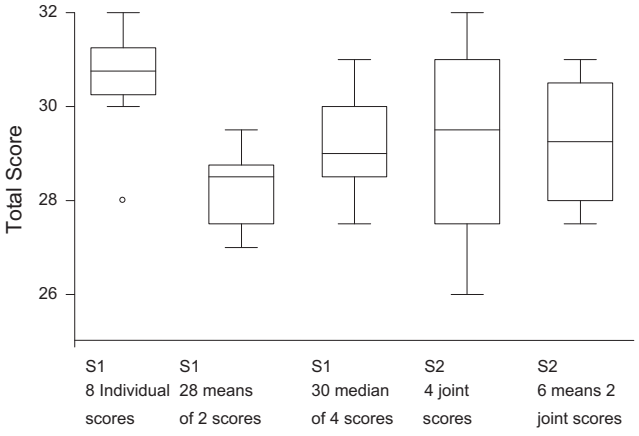

Fig. 3. Paper B distribution of total score for different aggregation strategies.

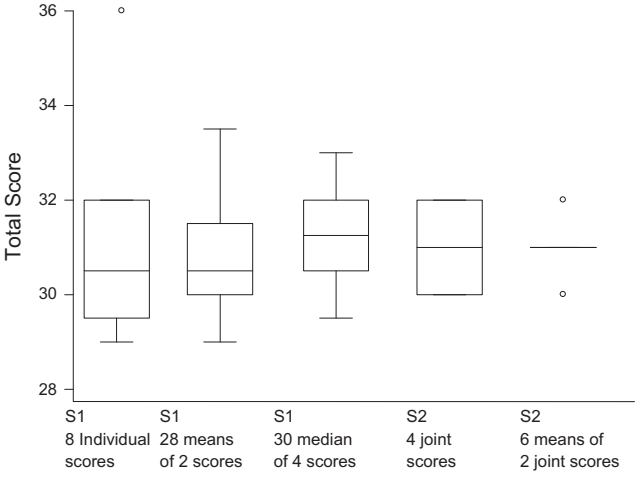

Fig. 4. Paper C distribution of total score for different aggregation strategies.

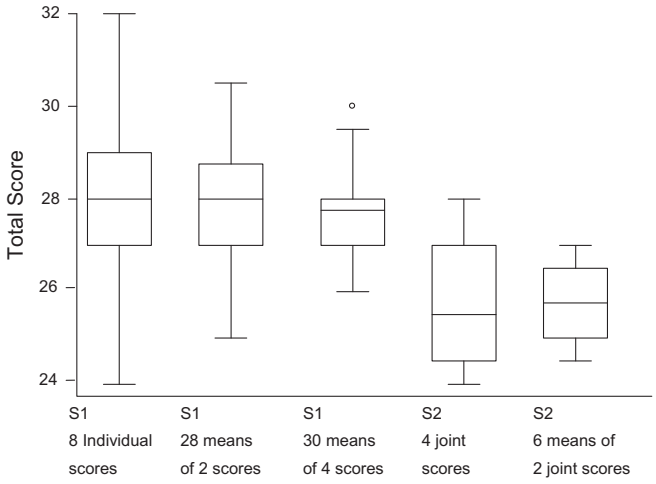

Fig. 5. Paper D distribution of total score for different aggregation strategies.

The five strategies used in each of the Figs. 2–5 are (in order of increasing effort per assessment):

- *Strategy 1 (S1)*: A single independent evaluation (i.e. no aggregation). This shows the variation in scores obtained by a single judge obtained from Study 1 data.
- *Strategy 2 (S2)*: The average of any two individual evaluations obtained from Study 1 data. We used the average in this case because although the median is preferred to assess the central point of ordinal scale variables, when two variables are involved, the median and the average are the same. There are 28 possible ways of aggregating two evaluations from eight. We provide the average for each aggregation. This shows the variation in scores obtained by aggregating the results from two judges without discussion.
- *Strategy 3 (S3)*: The median of four individual evaluations obtained from Study 1 data. There are a total of 70 different ways in which four assessments can be aggregated (i.e. ${}^{4}C_{8} = 8!/4! \times 4! = 70$). We selected 30 such combinations at random and derived the median for each question. We restricted the analysis to 30 combinations to reduce the manual effort involved in setting up the formulae for calculating the medians. This shows the variation in scores obtained by aggregating the results from four judges without discussion.
- *Strategy 4 (S4)*: A joint evaluation produced by a pair of judges obtained from Study 2 data. This shows the variation in scores obtained by aggregating the results of two judges and incorporating discussion.
- *Strategy 5 (S5)*: The average of two joint evaluations obtained from Study 2 data. There are six possible ways of aggregating four evaluations (i.e. ${}^{2}C_{4} = 4!/2! \times 2! = 6$). We obtained the average of each pair. This shows the variation in scores obtained by aggregating the results of four judges and incorporating discussion among pairs of judges.

Paper A (Fig. 2) and to a lesser extent paper C (Fig. 4), showed the results we had hoped for, i.e. the more judges involved, and the more discussion employed in the evaluation, the more consistent (i.e. less variable) the evaluation. However, in general results were mixed:

- Individual scores were less reliable (more variable) than any aggregated score. This result confirms that, as expected, more than one evaluation is necessary.
- In three of the four cases, the median of 4 individual scores (middle box plot) was more reliable than the average of 2 individual scores (second from left box plot).
- In three of four cases, the joint evaluation was more reliable than the mean of two individuals.
- In two cases, the median of 4 individual scores was more reliable than the joint evaluation.
- The average of two joint evaluations (the most effort intensive method since it involves four judges and discussion between two pairs of judges) was the most reliable aggregation in three of the four cases.

Generally, the results support RQ2 by suggesting that incorporating a period of discussion increases reliability but they also show that using more reviewers without discussion will also improve reliability. This implies that the answer to question RQ1 and RQ2 may be a matter of available effort. If discussion among two reviewers takes a long time, then it may be more cost effective to allocate extra reviewers and take a simple arithmetic aggregate of the individual reviewers.

5.2. Results for Study 3

The results from Studies 1 and 2 indicated that a joint evaluation based on a team of two was more reliable than the average of two independent evaluations, and also that increasing the number of judges further increased reliability. They also suggested that evaluations based on checklists were more reliable than simple overall subjective assessments. To further investigate the issue of the number of judges and the importance of discussion, Study 3 investigated the reliability achieved by one-person, two-person and three-person teams.

5.2.1. Descriptive statistics

The total scores for each group and each paper are shown in Fig. 6. Comparing these results with the results for Studies 1 and 2 (see Figs. 4 and 5 respectively), it appears that most judges in this study agreed that paper C was better than paper D, but that they scored both of these papers down when compared with the scores from Studies 1 and 2.

5.2.2. The reliability achieved by teams of different sizes

In order to use the weighted Kappa, we rounded three scores that were recorded as fractions up to the nearest whole number. The weighted Kappa values for the three groups and two papers are shown in Fig. 7. The horizontal dotted line shows the upper 95% of the null Kappa distribution. Each box plot in Fig. 7 identifies the Kappa distribution for a specific paper and a specific team size: the letter C or D identifies the paper while the term T1, T2 or T3 indicates the number of persons in each treatment group (1, 2, or 3). It does not require statistical analysis to confirm that our null hypothesis cannot be rejected. The reliability among individuals

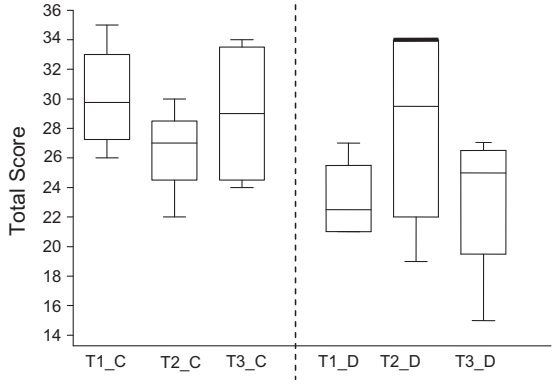

Fig. 6. Study 3 total scores for paper C and D for teams of 1, 2 and 3 persons.

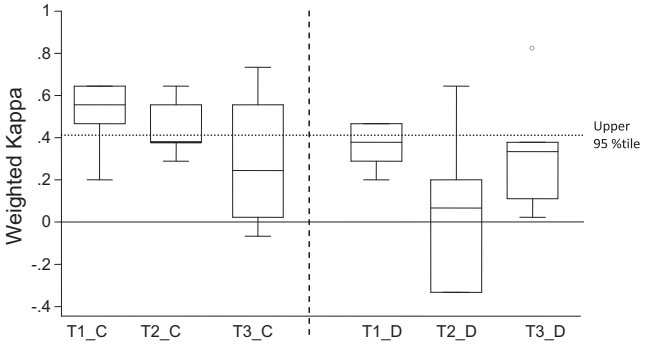

Fig. 7. Study 3 weighted Kappa values for papers C and D for teams of 1, 2 and 3 persons.

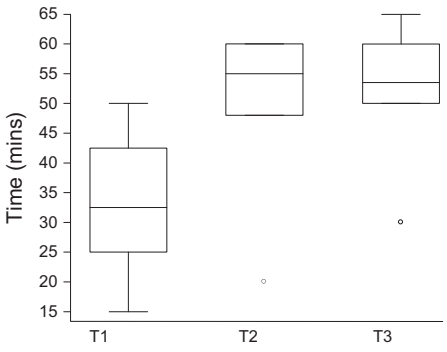

Fig. 8. Elapsed time for evaluation for the three treatment groups.

appears better than the reliability among the teams of two or three persons; although even for the single individuals the reliability is not very high, particularly for paper D. The dropout team was one that involved three persons, supporting the view that obtaining agreement among three persons is difficult. These results suggest that discussion among judges provides no benefit at all to reliability. One reason could be that in contrast to Study 2, the time for discussion was limited. Indeed two of the eight three-person teams commented that insufficient time was allowed for the task, and one of the three-person teams dropped out of the experiment.

5.2.3. The cost of discussion

The distribution of elapsed task time (for those teams that recorded it) shows that teams of two or three took on average 20 min longer than individuals to complete their task (see Fig. 8). Furthermore, if we assume that an individual takes 35 min, and a team of two or three takes 55 min, then this corresponds to 35 min effort for an individual compared with 110 min effort for a team of two and 165 min for a team of three. The time for three individuals would be 105 min, so if three individuals are at least as reliable as a team of two people, it is more cost effective to use three individuals and employ the median evaluation. However, we must treat these results with caution because some of the recorded times may have excluded reading time.

5.2.4. Comments from participants

The teams and individuals taking part in Study 3 were fairly critical of the evaluation criteria. 50% of the teams criticised the checklist in terms of the items themselves (e.g. overlapping items, irrelevant items, missing questions and assessing the quality of the reporting rather than the quality of the experiment) and the scales (one team wanted a three or five point scale not a four point scale – while another team wanted to have four point scales everywhere including the overall assessment).

6. Discussion and conclusions

In this section we discuss the answers to our research questions, discuss the contrasting assessments of the value of discussion among judges found in Studies 2 and 3, report study limitations, and provide recommendations for quality evaluation in the context of systematic literature reviews.

6.1. Answers to research questions

Unfortunately there are few unequivocal answers to any of our research questions. RQ1 was concerned with the number of judges necessary to achieve a reliable assessment. For Studies 1, 2 and 3, the weighted Kappa results all suggest that more than one person is necessary. Study 1 showed that simple aggregation using the median of the assessments of two or four judges is more reliable than individual assessments. Study 2 suggested that two judges with a period of discussion was comparable with the median of results from four independent judges.

RQ2 was concerned with the value of discussion compared with arithmetic aggregation of judges' assessments. Comparison of aggregation strategies used by Studies 1 and 2 suggested that the median of 8 individual assessments was as good as the median of 4 joint assessments. However, a more detailed comparison of the results from Studies 1 and 2 indicated that discussion improved reliability but so did increasing the number of judges. The choice between introducing discussions or increasing the number of judges may need to be based on an assessment of cost effectiveness. However, in stark contrast to Study 2, the results of Study 3 suggested strongly that discussion did not improve reliability. We discuss possible reasons for this contradiction in the next section but currently we do not have an answer to this question.

RQ3 was concerned with assessing the value of using a checklist compared with an overall subjective assessment. The ICC results for Studies 1 and 2 (Table 5) suggest that using an aggregate score delivers better reliability than a simple overall assessment. For Study 1 a more detailed analysis shown in Table 6 using the sum of the individual questions (i.e. the total score), indicates that two judges are sufficient to obtain substantial reliability, although 10–14 judges would be necessary to achieve comparable reliability if a single overall subjective assessment were used.

Our results show some consistency with previous studies. For example, as suggested by Neff and Olden [30], we found that the aggregated results from eight reviewers gave very reliable results. Our results concerning the use of a quality checklist to improve reliability are also consistent with those from other researchers, but it is clear from comments made by participants in Study 3 that constructing a checklist that is acceptable to a large number of judges and applicable to a large number of different types of scientific research paper is far from being an easy task.

6.2. Team working contradictions

The results of Studies 2 and 3 appear to contradict one another. In Study 2, the results of team working appeared beneficial but in Study 3 the results were not beneficial. One major difference between the two studies is that in Study 3, the evaluation was time constrained, and there are also other differences that might have influenced the results. The one-person teams seemed to have more than sufficient time to read and assess the papers, so it is likely that the consultation period required by the teams of two and three persons was the major time constraint. Also, we must not forget that the power of our experiment is low, although there are other possible explanations.

Looking at the literature on team working it appears that in most cases, teams are better at choosing, judging, estimating and

problem solving than individuals [13]. However, this superiority of teams to individuals depends on a variety of situational and task factors: teams do not perform well when tasks are difficult, complex, unfamiliar, uninteresting, or when time is limited [21,39]. In Study 3 the participants were given an unfamiliar assessment form and had only limited time. In Study 2 the members of the teams had designed the form used for the assessment and already applied it to the research papers. Thus, in contrast to Study 3, Study 2 participants were more familiar with the task and furthermore were not limited with respect to time.

In addition, team effectiveness depends on the extent to which the task has a demonstrable correct solution or whether, as in the case for our studies, it is a judgmental task, for which no correct answer can be authoritatively determined [25,26]. Teams are more clearly superior in the former situation than the latter, as the latter often suffers from polarisation [29] and groupthink [15]. Our results from Study 3 might therefore be due to polarisation, in which the teams are more extreme than the individuals, although in the same direction as the average individual responses; or because of groupthink arising from overestimation on the part of the team, closed-mindedness or pressures toward uniformity [15]. The contrasting results of Study 2 might be because we did not rely on face-to-face meetings. The Delphi process [5] and groupware software systems (e.g. [31,32]) are believed to deliver value because they are designed to avoid some of the negative effects of face-to-face team meetings such as:

- domination of the conversation by one or more members,
- individuals fearing criticism or negative evaluation,
- members failing to participate because they perceive that their input is not required, and
- pressure to conform with senior members of the team.

6.3. Limitations

Our studies have several limitations. One major limitation is the number of targets. Although we used a random sample from a set of research papers selected by an independent experienced researcher with similar aims to ours, with only four papers we cannot be sure how well our results will generalise to our target of all human-intensive software engineering experiments. For example, the papers do not constitute a homogeneous sample. In particular, paper A is rather different from the other papers because the human-intensive experiment presented in the paper was only a small part of a wider evaluation exercise. Overall, paper A was good, but the human-intensive experiment was weak. Furthermore, all the papers were journal papers and had therefore undergone a stringent review process; we would expect assessments of conference papers to be less reliable than those of journal papers.

Another limitation is that we, as a group of researchers, have extensive experience of empirical software engineering, so our Study 1 and Study 2 results may be more reliable than those that would have been obtained by a random selection of researchers. This problem does not affect Study 3 where a group of 48 individuals with varying levels of experience took part.

The checklist that we used was a modified version of a checklist that had already been developed and used by others. As reported by Jüni et al. [17], using another checklist with another set of criteria for determining quality, might have given other results. Also, from the comments from Study 3 participants, the checklist itself might be problematic for people who had not had experience using it. The agreement among reviewers may depend on the actual checklist being used. It is likely that the better the checklist, the higher the agreement. Our checklist was based on a checklist used in other disciplines and has been used several time in software engineering before, and was also improved during our work. Other, less well-founded checklists might have given even larger disagreement than that which we detected.

In addition, we calculated the total score for a paper by summing the scores for the individual questions; strictly such a procedure is only valid if the individual questions are independent and each question has an equal impact on the overall quality of the paper.

It should also be noted that our studies were based on *published* papers. It may be that assessments of draft papers would have been even more diverse. This means that although our conclusions apply to quality evaluation during the systematic review process, they may not generalise well to the quality assessment of manuscripts performed during the journal and conference peer review process.

We have identified substantial limitations among our three studies. However, being able to judge the quality of experiments in software engineering is crucial to study the effect of methodological or technical support to improve the way we carry out such experiments. Given that nobody else has systematically investigated how we agree on the quality of reported experiments in the context of systematic reviews, we believe this set of three studies offers some insight as to the complexity of the problem. However, given the novelty of this line of meta-research and the challenges involved, our studies should all be considered as explorative or pilot studies.

6.4. Recommendations for quality assessments for systematic reviews

Current recommendations for quality reviews of primary studies in systematic reviews recommend two judges and discussion in the event of disagreement. Given the results of Study 3, it appears that the benefits of discussion may be more limited than we had expected. Study 2 and the study by Kitchenham et al. [23] both favoured discussion among participants but they did not consider whether the additional costs of discussion would lead to significantly better reliability than simply adding another judge and aggregating scores arithmetically. Kitchenham et al. [23] found relatively good results by simply taking the median value of three independent scores without any discussion. Furthermore, in Study 1, the median of four independent scores led to quite good reliability.

Thus, our current advice to researchers performing quality evaluations that are part of a systematic review is to:

- Use a quality checklist but ensure all judges understand how to use it.
- In contrast to current systematic literature review standards, begin by using three independent judges. Study 1 suggested that some papers were easier to assess than others (i.e. better reliability was found for papers A and D). In particular, for papers A and D, two judges with discussion produced very good reliability. However, with two reviewers it is less likely that disagreements will be uncovered, so we believe a third independent review would be more beneficial than a period of discussion. Take the median of the scores from the independent judges for individual criteria – however if the values of individual criteria are summed to produce an overall score, the mean of the total scores should be used.
- If there is strong disagreement among judges for a specific paper, apply an exception process by either finding a fourth judge or by instituting a round of discussion. Assuming judges are equally experienced, using such a procedure is intended to ensure that extra assessment effort will be applied to the papers that are most difficult to assess.

Given the limitations of our three studies, these recommendations should be treated with some caution. However, in the related

papers we found that, with the exception of Kitchenham et al. [23], and apart from studies of randomised controlled trials, there was little empirical evaluation of the use of criteria to determine the quality of studies that are to be used in meta-analysis or systematic reviews, so any evidence-based recommendations are better than nothing.

However, it is clear that this topic would benefit from additional studies. In particular, some of the major limitations of Study 3 need to be addressed. For example, it would be useful to provide the participants with a more extensive introduction to the checklist and give them more time to read the research papers. It would also be useful to use a wider sample of research papers. Any team-based discussions and assessments should take place after the individuals comprising the teams have completed an individual questionnaire and each part of the experiment (individual assessments and team assessments) needs to be accurately timed. This would allow the impact of discussion and the impact of simple aggregation of results to be compared in terms of both reliability and cost effectiveness. The problem here would be finding enough suitably qualified participants willing to spend up to 2 h on such an experiment.

Acknowledgements

We thank all the ISERN members who took part in our third study. We also thank the anonymous reviewers for their comments which have substantially improved the quality of our paper.

References

[1] S. Abrahão, G. Poels, Experimental evaluation of an object-oriented function, Information and Software Technology 49 (4) (2007) 366–380.
[2] W. Afzal, R. Torkar, R. Feldt, A systematic review of search-based testing for non-functional system properties, Information and Software Technology 51 (6) (2009) 959–976.
[3] V. Alves, N. Niu, C. Alves, G. Valença, Requirements engineering for software product lines: a systematic review, Information and Software Technology 52 (8) (2010) 806–820.
[4] L. Bornmann, R. Mutz, H-D. Daniel, A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants, PLoS ONE 5 (12) (2010) e14331, doi:10.1371/journal.pone.0014331.
[5] B.B. Brown, Delphi Process. A Methodology used for Elicitation of Opinions of Experts. Rand Paper P-3925, 1968.
[6] L. Chen, M.A. Babar, C. Cawley, A status report on the evaluation of variability management, in: Proceedings of the Conference on Evaluation and Assessment in Software Engineering EASE 2009, BCS eWic, 2009.
[7] J. Cohen, A coefficient of agreement for nominal scales, Education and Psychological Measurement 20 (1960) 37–46.
[8] J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychological Bulletin 70 (1968) 213–220.
[9] T. Dybå, T. Dingsøyr, Strength of evidence in systematic reviews in software engineering, in: Proceedings of the Conference on Empirical Software Engineering and Metrics, ESEM'2008, 2008, pp 178–187.
[10] T. Dybå, T. Dingsøyr, Empirical studies of agile software development: a systematic review, Information and Software Technology 50 (9–10) (2008) 833–859.
[11] T. Dyba, V.B. Kampenes, D.I.K. Sjøberg, A systematic review of statistical power in software engineering experiments, Information and Software Technology 48 (8) (2006) 745–755.
[12] M. Egger, P. Jüni, C. Bartlett, F. Holenstein, J. Sterne, How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study, Health Technology Assessment 7 (1) (2003).
[13] D.R. Forsyth, Group Dynamics, fifth ed., Wadsworth Publishing, 2009.
[14] J.P.A. Ioannidis, A-B. Haidich, J. Lau, Any casualties in the clash of randomized and observational evidence? Editorial, BMJ 322 (2001) 879–880.
[15] I.L. Janis, Groupthink: Psychological Studies of Policy Decisions and Fiascoes, second ed., Houghton, Mifflin, 1982.
[16] M. Jørgensen, K. Moløkken-Østvold, Impact of effort estimates on software project work? A review of the 1994 Chaos report, Information and Software Technology 48 (4) (2000) 297–301.
[17] P. Jüni, A. Witschi, R. Bloch, M. Egger, The hazards of scoring the quality of clinical trials for meta-analysis, JAMA 282 (11) (1999) 1054–1060.
[18] V.B. Kampenes, Quality of Design Analysis and Reporting of Software Engineering Experiments, A Systematic Review, PhD Thesis, Dept. Informatics, University of Oslo, 2007.
[19] A. Karahasanović, A.K. Levine, R. Thomas, Comprehension strategies and difficulties in maintaining object-oriented systems: an explorative study, Journal of Systems and Software 80 (2007) 1541–1559.
[20] L. Karlsson, T. Thelin, B. Regnell, P. Berander, C. Wohlin, Pair-wise comparisons versus planning game partitioning – experts on requirements prioritisation techniques, Empirical Software Engineering 12 (2007) 3–33.
[21] S.J. Karau, J.R. Kelly, The effects of time scarcity and time abundance on group performance quality and interaction process, Journal of Experimental Social Psychology 28 (1992) 542–571.
[22] Barbara Kitchenham, Shari Lawrence Pfleeger, Lesley Pickard, Peter Jones, David Hoaglin, Khaled El Emam, Jarrett Rosenberg, Preliminary guidelines for empirical research in software engineering, IEEE Transactions on Software Engineering 28 (8) (2002) 721–734.
[23] B. Kitchenham, P. Brereton, M. Turner, M. Niazi, S. Linkman, R. Pretorius, D. Budgen, Refining the systematic literature review process – two observer-participant case studies, Empirical Software Engineering 15 (6) (2010) 619–653.
[24] B.A. Kitchenham, D.I.K. Sjøberg, O.P. Brereton, D. Budgen, T. Dybå, M. Høst, D. Pfahl, P. Runeson, Can we evaluate the quality of software engineering experiments? in: Proceedings of the Conference on Empirical Software Engineering and Metrics ESEM 2010, 2010.
[25] P.R. Laughlin, Collective induction: twelve postulates, Organizational Behavior and Human Decision Processes 80 (1) (1999) 50–69.
[26] P.R. Laughlin, E.C. Hatch, J.S. Silver, L. Boh, Groups perform better than the best individuals on letters-to-numbers problems: effects of group size, Journal of Personality and Social Psychology 90 (4) (2006) 644–651.
[27] H. Liu, H.B.K. Tan, Testing input validation in Web applications through automated model recovery, Journal of Systems and Software 81 (2007) 222–233.
[28] H.W. Marsh, U.W. Jayasinghe, N.W. Bond, Improving the peer-review process for grant application. Reliability, validity, bias and generalizability, American Psychologist 63 (3) (2008) 160–168.
[29] D.G. Myers, H. Lamm, The group polarization phenomenon, Psychological Bulletin 83 (4) (1976) 602–627.
[30] B.D. Neff, J.D. Olden, Is peer review a game of chance, BioScience 56 (4) (2006) 333–340.
[31] J.F. Nunamaker, R.O. Briggs, D.D. Mittleman, D.R. Vogel, P.A. Balthazard, Lessons from a dozen years of group support systems research: a discussion of lab and field findings, Journal of Management Information Systems 13 (3) (1996) 163–207.
[32] J.F. Nunamaker, A.R. Dennis, J.S. Valacich, D. Vogel, J.F. George, Electronic meeting systems to support group work, Communications of the ACM 34 (7) (1991) 40–61.
[33] G.S. Omenn, G.E. Goodman, M.D. Thornquist, J. Balmes, M.R. Cullen, S. Hammar, Alpha-tocppherol, Beta carotene cancer prevention study group, 1996, Effects of a combination of beta carotene and vitamin a on lung cancer and cardiovascular disease, New England Journal of Medicine 334 (1996) 1150–1155.
[34] M. Petticrew, H. Roberts, Systematic Reviews in the Social Sciences: A Practical Guide, Blackwell Publishing, 2005.
[35] R.W. Poolman, L.C. Keijser, M.C. de Waal Malefijt, L. Blankevoort, F. Farrokhyar, M. Bhandari, Reviewer agreement in scoring 419 abstracts for scientific orthopedics meetings, Acta Orthopaedica 78 (2) (2007) 278–284.
[36] B.H. Rowe, T.L. Strome, C. Spooner, S. Blitz, E. Grafstein, A. Worster, Reviewer agreement trends from four years of electronic submissions of conference abstract, BMC Medical Research Methodology 6 (14) (2006).
[37] D.M. Schultz, Are three heads better than two? How the number of reviewers and editor behaviour affect the rejection rate, Scientometrics (2009), doi:10.1007/s11192-009-0084-0.
[38] A. Shang, K. Huwwiler-Müntener, L. Nartney, P. Jüni, S. Dörig, D. Pwesner, M. Egger, Are the clinical effects of homeopathy placebo effects? Comparative study of placebo-controlled trials of homeopathy and allopathy, Lancet 366 (9487) (2005) 726–732.
[39] M.E. Shaw, Group Dynamics: The Psychology of Small Group Behavior, third ed., McGraw-Hill, 1981.
[40] P.E. Shrout, J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability, Psychological Bulletin 86 (2) (1979) 420–428.
[41] D.I.K. Sjøberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N.K. Liborg, A.C. Rekdal, A survey of controlled experiments in software engineering, IEEE Transactions on SE 31 (9) (2005) 733–753.
[42] S. Yusuf, G. Dagenais, J. Pogue, J. Bosch, P. Sleight, Vitamin E supplementation and cardiovascular events in high-risk patients. The heart outcomes prevention evaluation study investigators, New England Medical Journal 342 (2000) 150–160.
[43] A.C. Weller, Editorial Peer Review: Its Strengths and Weaknesses, Information Today, Inc., Medford, NJ, USA, 2002.
[44] M. Wood, M. Roberts, B. Howell, The reliability of peer reviews of papers on information systems, Journal of Information Science 30 (1) (2004) 2–11.