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.

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.