Trends in the quality of human-intensive software engineering experiments – A quasi-experiment

Journal:	*Transactions on Software Engineering*
Manuscript ID:	TSE-2012-04-0082.R1
Manuscript Type:	Regular
Keywords:	D.2 Software Engineering < D Software/Software Engineering, K.7.m.a Codes of good practice < K.7.m Miscellaneous < K.7 The Computing Profession < K Computing Milieux

Trends in the quality of human-centric software engineering experiments – A quasi-experiment

Barbara Kitchenham[1], Dag I.K. Sjøberg[2], Tore Dybå[2,3], Pearl Brereton[1], David Budgen[4], Martin Höst[5], Per Runeson[5]

[1]School of Computing and Mathematics, Keele University, Keele, Staffordshire, ST5 5BG, UK

{B.A.Kitchenham, O.P.Brereton}@keele.ac.uk

[2]Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, NO-0316 Oslo, Norway

Dag.Sjoberg@ifi.uio.no

[3]SINTEF, P.O. Box 4760 Sluppen, NO-7465 Trondheim, Norway

Tore.Dyba@sintef.no

[4]School of Engineering and Computing Sciences, Durham University, Science Laboratories, Durham, DH1 3LE, UK

David.Budgen@durham.ac.uk

[5]Department of Computer Science, Lund University, SE-221 00 Lund, Sweden

{Martin.Host, Per.Runeson}@cs.lth.se

Abstract

Context: Several text books and articles published between 2000 and 2002 have attempted to introduce experimental design and statistical methods to software engineers undertaking empirical studies. Objective: This paper investigates whether there has been an increase in the quality of human-centric experimental and quasi-experimental journal papers over the time period 1993 to 2010. Method: 70 experimental and quasi-experimental papers published in four general software engineering journals in the years 1992-2002 and 2006-2010 were each assessed for quality by three empirical software engineering researchers using two quality assessment methods (a questionnaire based method and a subjective overall assessment). Regression analysis was used to assess the relationship between paper quality and the year of publication, publication date group (before 2003 and after 2005), source journal, average co-author experience, citation of statistical text books and articles, and paper length. The results were validated both by removing papers for which the quality score appeared unreliable and using an alternative quality measure. Results: Paper quality was significantly associated with year, citing general statistical texts and paper length ($p<0.05$). Paper length did not reach significance when quality was measured using an overall subjective assessment. Conclusions: The quality of experimental and quasi-experimental software engineering papers appears to have improved gradually since 1993.

Index terms: Quality evaluation, Empirical studies, Human-centric experiments, Experimentation, Software Engineering

1. Introduction

From the start of the 21st century, many researchers involved in human-centric software engineering experiments (ourselves included) became concerned about the methodological standard of software engineering (SE) experiments. In response to this concern a number of researchers published procedures and guidelines aimed at improving the rigour of conducting and reporting SE experiments (see for example [23], [27], [11], [17], [10]). In the context of this paper, we define human-centric SE experiments to be studies of SE methods, techniques and procedures that depend on human expertise. In such experiments the outcomes are determined by the efficacy of the combination of capability of the human participants and characteristics of the method, technique or procedure, e.g., studies of design methods or code reading methods. These can be compared with technology-centric studies, where the techniques, methods or procedures are implemented in tools, and it is assumed that outcomes depend on the task and the tools, with the impact of human capability on the outcome being considered to be negligible. For example experiments that compare the effectiveness of test cases generated by data flow analysis with test cases generated by mutation analysis.

In 2005, Sjøberg et al. reported the results of their major study of 103 papers describing human-centric experiments and quasi-experiments published in 13 leading journals and conferences [23]. Their study and three other related studies investigated the same set of papers which were published between 1983 and 2002 inclusive ([6], [14], [16]). These studies confirmed the view of empirical SE methodologists that there were problems with SE experiments. For example, independent replications of experiments often found contradictory results [23]; statistical power was poor [6]; few papers reported effect sizes [14]; and design and analysis of quasi-experiments needed to be improved [16].

As yet, however, there has been no assessment of later papers that might be expected to have benefitted from the recent spate of guidelines and text books. Thus, we believe that it is time to investigate whether there has been any noticeable improvement in SE experiments, and this is the goal of the study reported here.

In Section 2 we discuss related research. In Section 3 we present our methodology. We present our results in Section 4 and discuss them in Section 5. Section 6 concludes our paper.

2. Related Research

We have found no papers in the field of SE that investigated whether the quality of SE papers is changing over time. However, there are studies of quality evaluation procedures in many disciplines. In a recent paper, we summarised research related to quality criteria used to evaluate experiments [18], pointing out that quality criteria in medical studies were based on three issues:

1. Use of random allocation to experimental conditions.
2. Use of single-blind versus double blind procedures.
3. How dropouts were analysed.

Furthermore, we noted that there are some doubts about using checklists based on more general criteria to assess medical studies [12]. For SE studies, we argued that *double-blind procedures* and the *intention to treat* method [8] were inappropriate and, therefore, not being used in the context of SE experiments. (In *double-blind procedures,* the experimenter and the subjects do not know what experimental condition they are assigned. In the *intention to treat* method, the subjects are analysed within the experimental condition to which they were assigned even if they dropped out.) Consequently we argued that the use of another set of quality criteria was necessary for SE experiments, as it is for other disciplines such as education or psychology.

After we began work on this study, Dieste et al. published a study that investigated the relationship between internal validity and bias in SE experiments, where bias refers to “a tendency to produce results that depart systematically from the ‘true’ results” [5]. They identified a set of 10 quality evaluation questions and evaluated 25 studies that had been aggregated using meta-analysis (in two separate meta-analyses). They applied the 10 quality evaluation questions to each paper and correlated the results with bias (measured as the difference between the overall average effect size calculated in the meta-analysis and the mean effect size observed in the study). They found only three questions that were negatively and significantly correlated with bias (noting that a large negative correlation with bias is associated with high quality and vice versa) which were:

- Q3: “Are hypotheses being laid [*sic*] and are they synonymous with the goals discussed before in the introduction?” (Correlation of -0.744 with bias)
- Q6: “Does the researcher define the process by which he applies the treatment to objects and subjects (e.g. randomization)?” (Correlation of -0.694 with bias)
- Q9: “Are the statistical significances mentioned with the results?” (Correlation of -0.406 with bias)

They also noted that one question had a high, although not significant, positive correlation with bias, which (rather surprisingly) was:

- Q8: “Is mention made of threats to validity and also how these threats affect the results and findings?” (Correlation of 0.25 with bias)

3. Materials and Methods

The basic method used in this quasi-experiment was to select a set of paper reporting human-centric experimental and quasi-experimental published on or before 2002, and to compare them with a similar set of papers published between 2006 and 2010 inclusive. The comparison was based on a quality questionnaire described in detail in a previous paper [18]. Seventy papers were selected in such a way that they provided as even a spread of papers per year as possible. This means that our experimental design is similar to an interrupted time-series design [24] with the aim of investigating whether the publication of SE guidelines on performing experiments (i.e. [11], [17] and [27]) caused an interruption in the quality trends of papers reporting SE experiments. The material and methods used in this quasi-experiment are discussed in more detail in the following sections.

3.1 Research Goal

Formally, the goal of this paper is to investigate whether the quality of human-centric SE experiments and quasi-experiments is showing an improvement over time. In particular, we were interested to see whether the guidelines for SE experiments produced in the early 2000's had improved the quality of experiments.

We restricted ourselves to an investigation of papers published in international SE journals, so we would expect the experiments that we included in our study to be of higher quality than SE experiments in general.

3.2 Experimental Units and Participants

There is one main experimental unit involved in this experiment: the set of papers to be assessed for quality. In addition, the human participants in this study are the seven co-authors of this paper.

The papers were obtained from two sources. Papers published on or before 2002 were selected from the 76 papers (of 103 papers) found in [23] that were published in four international journals *IEEE Transactions on Software Engineering* (TSE), *Empirical Software Engineering* (ESE), *Information and Software Technology* (IST) and the *Journal of Systems and Software* (JSS).

Relevant papers published between 2006 and 2010 inclusive were found by a search of the same four journals over the five-year period.

We excluded the years 2003 to 2005 from our analysis because we wanted to investigate whether guidelines for SE experiments (e.g. [27], [11], and [17]) had had an impact on experiment quality. If the guidelines had had an impact, it would have taken several years for that to become visible in journal citations, since given the time needed to get papers published, many SE experiments published in the years 2002-2005 would have been performed before the guidelines were published. The papers from the earlier time period (1993-2002) also fitted in well with the publication dates of the guidelines and provided a relatively long time period (i.e. 10 years) to establish any quality trends.

With respect to being active participants in the study, obviously, we are not a random selection of researchers. We are a group of SE researchers with an interest in, and experience of, undertaking SE experiments. Furthermore, we are often asked by journal editors and conference organisers to review empirical SE studies. Therefore, we are representative of reasonably expert empirical researchers with an understanding of issues related to the quality of SE papers, and we consider ourselves to be eligible to act as assessors of the quality of the papers assessed in this study.

3.2.1 Selection of papers available for inclusion in the study

We restricted the papers to those published in four journals because:

- These journals published the majority of papers on human-centric experiments and quasi-experiments that were found by Sjøberg et al. [23].
- Restricting ourselves to journal papers meant there was less likelihood of including duplicate reports of the same study from different sources (i.e. no likelihood of encountering both conference and journal versions of the same study).
- The restriction ensured that we had a homogeneous dataset with a reasonable number of papers from all the selected sources included in the two main time periods we analysed.

We also used the following exclusion/inclusion criteria:

- We excluded papers co-authored by any of the authors of this paper to avoid any possible bias in our quality evaluations.
- If a specific researcher was first author of many different papers (within each time period), we included no more than one paper with that researcher as first author to avoid biasing the results either for or against any individual researchers who published a large number of papers (and who are usually experienced researchers). To decide which paper from a particular author to include in the set of available studies, we either selected a paper published in the year that had fewest available papers or (if there was no clearly preferable year), we selected a paper at random.
- We excluded from the set of candidate papers those papers that we had used to test our quality questionnaire [18].

3.2.2 Available papers in the time-period 1993-2002

Of the 103 papers identified by Sjøberg et al. [23], we considered only the 76 journal papers from TSE, JSS, IST and ESE. Applying our exclusion criteria, we excluded six papers that were co-authored by co-authors of this study, and we excluded a further set of 10 papers to avoid including multiple papers by the same first author. The number of papers available from each of the four journals is shown in Table 1.

3.2.3 Search process used to find papers in the time period 2006-2010

In order to identify recent papers for inclusion in this study, it was necessary to search the four journals over the period 2006 to 2010 inclusive. It was not necessary to find every single paper with a human-centric experiment/quasi-experiment, only an unbiased set of papers.

Kitchenham performed the search for recent papers using a four stage process:

1. SCOPUS was searched using the following string recommended by Dieste and Padua [4]: TITLE-ABS-KEY("experiment" OR "empirical study" OR "empirical evaluation" OR experimentation) OR "experimental comparison" OR "experimental analysis" OR "experimental evidence "OR "experimental setting"). This search string identified 409 papers in the four specific journals for the years 2006-2010 inclusive. Kitchenham reviewed the title and abstract of each paper and identified 68 papers as possible human-centric experiments/quasi-experiments.
2. After completing the automated search, Kitchenham undertook a test-retest validity check [7], organised as follows: each issue of the four journals over the five-year period was inspected, and the abstract and title of each research paper were checked. This manual search identified 56 candidate papers from a total of 1568 papers.
3. The results of the two searches were compared. Overall 43 papers were included in both searches. The searches disagreed for 38 papers and agreed that 1487 papers (i.e. 1568-42-38) should be rejected. The manual search omitted 13 papers that were selected by the automated search, and the automated search omitted 25 papers that were included by the manual search. Comparing the two search method results gives a Kappa value of 0.68, which is categorised as "substantial".
4. The papers on which there was disagreement between the search processes were reviewed a second time, and 11 of the 13 papers selected by the manual search and rejected by the automated search and 13 papers of the 25 papers selected by the automated search and rejected by the manual search were included in the set of selected studies. Thus, 67 papers were initially available for selection into the study.

Although completeness was not absolutely essential for this study, we checked our set of studies with the eight studies that Kampenes [15] found after manually searching TSE, JSS, IST and ESE for the year 2007. Our search process found six of the eight studies, missing one paper where a human-centric experiment was only a small part of the validation exercise [13] and one paper we judged to be technology-centric [28].

After applying the exclusion criteria described previously:

- Two papers were rejected because, although they were found by the search process, they were actually published after 2010.
- Three papers were rejected because they were included in a previous study [18].
- Five papers were rejected because one of the co-authors of this paper was an author.
- Six papers were rejected in order to restrict authors that were *first* authors of multiple papers to at most one *first*-authored paper included in the study. Note that we did not place any restriction on co-authors other than first authors because that would have caused us to reject too many studies. In practice, this meant that some of the first authors also appeared as co-authors in other selected papers.

This left 51 papers available for inclusion in this study. The number of papers available from each of the four journals is shown in Table 1. Note that IST has far fewer older papers than more recent papers, whereas JSS has fewer recent papers than older papers.

Table 1 Number of papers available and selected from each journal

Journal	Older papers (1993-2002)		Recent papers (2006-2010)	
	Available	Selected	Available	Selected
TSE	12	10	7	5
IST	5	2	16	8
JSS	21	12	13	11

ESE	17	11	15	11
Total	55	35	51	35

3.2.4 Final selection of papers for inclusion in the study

With seven researchers available to conduct the study, and a decision that we wanted to obtain quality evaluations from three researchers for each paper (based on the recommendation of [18]), we decided to limit ourselves to 35 recent papers and 35 older papers. This meant that each researcher would have to assess 30 papers. We settled on these constraints because with 70 papers we were likely to achieve a reasonable power for any statistical tests [6], while keeping the workload required from each researcher to a manageable level.

In order to select the 35 older papers from the available 55 papers, with the goal of spreading the papers across 10 years we aimed to select three or four papers for each year and to include a maximum of one paper of all the authors that were first authors on multiple papers in the time period. We decided to select papers, in such a way that more recent years (1998-2002) were allocated four papers, while years earlier in the time period were allocated three papers (1993-1997). In practice, only two papers were available in 1994, so 1997 was allocated four papers.

If the available papers in a particular year corresponded to the allocated number of papers, all the available papers were allocated. If there were more papers than required for a specific year the required number of papers were selected at random. To do this, we allocated a random number to each of the available papers in that year, and selected the three or four papers that had been allocated the smallest random numbers.

In order to select the 35 recent papers from the available 51 papers with the goal of spreading the papers across the five years (2006 to 2010), we aimed to select seven papers for each year including a maximum of one paper from all the authors who were first authors of multiple papers in the time period. If seven papers were available in a year they were all selected. If more than seven papers were available they were selected at random using the same procedure as before. In 2010 only six papers were available so we allocated an additional paper to year 2009.

The number of papers finally selected for inclusion in the study is shown in Table 1.

3.2.5 Allocating Paper to Researchers

Each researcher assessed the quality of 30 papers. There are 35 different ways of selecting 3 individuals from a set of 7 individuals (${}^{3}C_{7}=7!\times(7-3)!/3!=35$). We itemised each of the 35 ways twice in a spreadsheet giving a list of 70 allocations of three numbers, i.e. the first row and the 36th row in the sheet comprised three columns with the number 1 in column 1, 2 in column 2 and 3 in column 3, the second and 37th row had column values 1, 2 and 4, etc. Each researcher was allocated a random number between 0 and 1. Then the researcher with the lowest number was allocated the number “1”, the next lowest was allocated the number “2”, etc. The 70 research papers were each allocated a random number between 0 and 1, ordered according to the value of the number and then allocated to one of the 70 available combinations. In this way 30 papers were allocated at random to each researcher. Note, we used the random number function supplied by Excel, which although limited because it will repeat values in very long sequences, is sufficient to ensure that the allocation of researchers to papers was not intentionally biased.

3.3 Experimental Material

The experimental materials used in this study were:

- A quality questionnaire which used nine individual questions about the quality of a human-centric experiment/quasi-experiment plus one question asking for an overall subjective assessment of the quality of the study. The questionnaire was the same as that used in our previous research [18], [19]. The only difference was in how it was scored, with the assessors being encouraged to interpolate between the 4-point ordinal assessment scale (0 to 3) for each question if they wanted, rather than select one of the discrete points. For convenience, a copy of the questions used in the questionnaire is shown in Table 2. Note that many of the questions relate to reporting practice. In addition to the nine basic questions, we also asked reviewers to make an overall subjective assessment of the paper on a 4-point ordinal scale (0=Poor; 1=Moderate; 2=Good; 3= Excellent) allowing interpolation. This was intended to act as a check on the validity of the questionnaire.
- A spreadsheet tailored for each researcher that included a list of all the papers allocated to that researcher. Columns in the spreadsheet corresponded to each of the nine quality questions and the overall subjective assessment. The spreadsheet also included a column to record whether the paper reported

quasi-experiments or formal experiments and how long (minutes) the researcher took to assess the paper.

Note that when individual papers reported more than one study, we did not undertake separate evaluations for each study reported in a specific paper: we just gave an overall assessment of the paper.

Table 2 Quality Questionnaire

#	Question	Related Question Number in Dieste et al. (2011)
Category: Questions on Aims		
1.	Do the authors clearly state the aims of the research?	Q3
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)?	Q4
3.	Do the authors describe the design of the experiment?	n/a
4.	Do the authors describe the data collection procedures and define the measures?	n/a
5.	Do the authors define the data analysis procedures?	Overlaps somewhat with Q9
6.	Do the authors discuss potential experimenter bias?	Overlaps somewhat with Q6 and Q10
7.	Do the authors discuss the limitations of their study?	Q8
Category: Questions on Study Outcome		
8.	Do the authors state the findings clearly?	n/a
9.	Is there evidence that the Experiment/Quasi-Experiment can be used by other researchers / practitioners?	n/a
Score: 0=Not at all; 1=Somewhat; 2=Mostly; 3= Fully; Interpolate if you want		

Table 2 also identifies the relationship between the quality questions that we used in our questionnaire and the questions used by Dieste et al. [5] in their study (see Section 2).

3.4 Tasks

Each researcher was responsible for assessing each of the 30 papers allocated to him/her and recording the assessment in their spreadsheet. We imposed no time limits for assessments of individual papers.

Researchers who decided that the paper they had been allocated was not in fact a human-centric experiment/quasi-experiment were instructed to consult with other researchers dealing with the same paper and if necessary approach Kitchenham to be allocated a replacement paper. However, in no case did an allocated paper need to be replaced.

3.5 Hypotheses, parameters and variables

Our null hypothesis is that there has been no difference in the quality trends observed in human-centric experiments and quasi-experiments in the years 1993 to 2002 and 2006 to 2010. There are two conditions that would support the null hypothesis:

- In the absence of any linear relationship between publication date and quality, the average quality of papers published in the period 1993 to 2002 (TP1, i.e. Time Period 1) is not significantly different from the average quality of papers published in the period 2006 to 2010 (TP2, i.e. Time Period 2).
- In the presence of a linear relationship between publication date and quality, the gradient of the regression line between publication date and quality of papers published in TP1 is not significantly different from the gradient of the regression line between publication date and quality of papers published in TP2.

The alternative hypothesis is supported if:

- In the absence of any linear relationship between publication date and quality, the average quality of papers published in TP1 is significantly lower than the average quality of papers published in TP2.

- In the presence of a linear relationship between publication date and quality: the gradient of the regression line between publication date and quality of papers published in TP1 is significantly less than the gradient of the relationship between publication date and quality of papers published in TP2.

3.5.1 Assessing paper quality

The measure of total quality for a paper obtained from an individual researcher is the sum of the nine quality questions (i.e. varies from 0 to 27). Our hypotheses are based on the average quality of the paper, that is, the average of the three total quality scores obtained from the researchers who assessed the paper. Each assessor also allocated an overall subjective assessment of quality to each paper. We assessed the subjective quality of a paper by taking the average of the three subjective assessments.

The level of agreement among individual researchers for the total score and the subjective overall score of each paper was assessed using the IntraClass Correlation (ICC) coefficient [25]. There are three variants of the ICC depending on whether the same judges are used for each paper or different judges are used for each paper; see [18] for a more detailed discussion of the ICC and its variants. Since we randomised the allocation of three judges to each paper (as opposed to having the same set of judges evaluate each paper), we used the simplest version of ICC based on the within and between paper variance. Since a two-way analysis of variance suggested that the effect of individual judges was statistically significant, our ICC values are conservative. The ICC value for the total score was 0.51, which is considered moderate agreement. The ICC value for the overall subjective assessment was 0.61, which is considered substantial. However, the overall subjective assessment is represented as an ordinal scale number, and the ICC value is based on analysis of variance which assumes a normally (or approximately normally) distributed variable, so the ICC value must be treated with some caution.

3.5.2 Reliability and validity of the quality variable

If the basic reliability of our assessments had been unacceptable (i.e., the ICC value was not statistically significant or interpreted as poor or slight), we would have needed to undertake additional assessments of papers exhibiting particularly poor reliability (i.e., papers where the variability of a total quality score was substantially larger than the average). Given that the ICC value for the total score was moderate, this proved unnecessary.

Nonetheless, in the event of the variability of a total quality score being substantially larger than the average for specific papers, or the overall subjective assessment being out of alignment with the quality score, we planned to repeat our analysis omitting the papers whose assessment appeared particularly unreliable.

With respect to the construct validity of our questionnaire, it is necessary to investigate whether our set of questions are truly related to paper quality. Dieste et al. [5] identified three questions significantly negatively related to bias in his quality questionnaire (where high bias represents poor quality) and one other question that appeared to have a positive relationship with bias. Three of our questions were related to the questions that Dieste found negatively related to bias (although only one was very clearly equivalent). We also have a question similar to the question that Dieste found positively related to bias. To assess the construct validity of our questionnaire, we analysed the score for each of these questions (i.e. our questions Q1, Q5, Q6 and Q7). If our results were broadly consistent with the results for the total quality score we could have some confidence that our total quality score, at least, relates to one aspect of paper quality. Note, although each question was based on an ordinal scale, the score for an individual question for a specific paper was based on the mean of three independent assessments. Since the central limit theorem states that the mean of a set of values will be approximately Normal irrespective of the distribution of the individual variables, we believe it is valid to apply the same regression style analysis that we used for the total score (as described in Section 3.7).

3.6 Design

This study is a quasi-experiment, specifically an interrupted time series. We used two groups of papers, those published before 2003 and those published after 2005, and we sought to determine whether the quality of the more recent papers is greater than the quality of the papers published in the earlier time period. Furthermore if any improvement was found, we wanted to know whether the SE guidelines published between 2000 and 2002 had contributed to that improvement

Shadish et al. [24] point out several problems with interrupted time series that are relevant to our study:

- Changes are often adopted slowly and diffuse through a population rather than taking place abruptly. Since we are interested in the impact of guidelines published in between 2000 and

2002, we have omitted the papers published in the years 2003-2005 from our study.

- Many data series are much shorter than the 100 observations recommended for statistical analysis. This is certainly the case in our study and means we may be unable to detect small changes.

Another potential problem with this design is that when the "interruption" is not due to an organisational or legal policy, it is always possible that some other factor has caused any observed change and not the one being suggested. For example, we are aware of other initiatives such as the *International Software Engineering Research Network* (ISERN) established in 1997, the *Empirical Software Engineering Journal* established in 1997 that both aimed to encourage researchers to undertake human-centric experiments and quasi-experiments. However, these initiatives started in a slightly earlier time period. If these were the cause of any changes, we would expect to see the beginning of quality changes to occur during the time period 1999-2002.

In addition, we noticed that the more recent papers appeared to be longer than the papers published before 2003, so we identified the length in pages of each paper to investigate whether there was a confounding effect between the length of the paper and its quality (bearing in mind that the quality questions were oriented to reporting quality).

Finally, even if the observed changes are restricted to the time period 2006-2010, so that they appear to be due to the guidelines and textbooks, we cannot tell whether any favourable change was due to the experimenters directly adopting the suggested practices or due to journal reviewers taking a more critical attitude to experiments and quasi-experiments (particularly as some frequent journal reviewers were responsible for producing those guidelines and text books). To investigate this issue, we have checked whether authors actually referenced the guidelines and text books. However, bearing in mind that most expert researchers are aware of good practice and are therefore unlikely to regard it as necessary to make reference to general guidelines or text books, we also considered the publication record of co-authors of papers to identify whether or not the authors were experienced researchers. To do so, for each paper included in our analysis, we found the specific paper in the DBLP database[1], and counted how many papers by each co-author were published prior to the year of publication of the specific paper. We used three variables based on this data: the average experience of all the co-authors, the maximum experience of any co-author and the experience of the first co-author.

We also checked the citation list of each paper to determine for all papers whether they cited a statistical text book (including manuals for statistical tools). We looked at the specific statistical references (if any) cited by each paper to see whether citation practice had been influenced by the statistical guidelines and text books authored by SE researchers. We considered only statistical texts and texts discussing quasi-experiments because the papers we were assessing were restricted to experiments and quasi-experiments.

3.7 Analysis Process

In order to test the main hypothesis we used segmented regression [26]. A segmented regression analysis allows the relationship between an outcome measure and the time after an interruption of some kind to be assessed in terms of differences with respect to both the gradient and intercept of the relationship in the time period before the interruption. Segmented regression is based on the following model:

$$Score_{ijk} = a + b_1 \times Year_i + b_2 \times Group_j + b_3 \times TP2Year_i + \varepsilon_{ijk} \qquad (1)$$

Where:

$Score_{ijk}$ is the total quality score for paper k in Year i and Group j and TP2Year i.

$Group_j$ identifies the time period in which the paper was published, where j=0 if the paper was published in the period 1993 to 2002, and j=1 if the paper was published in the period 2006 to 2010. If the estimate of b_2 is significantly different from zero, then the linear relationship in Time Period 2 has a significantly larger intercept than the relationship in Time Period 1.

$Year_i$ identifies the year in which the paper was published, where i=1993, …, 2002, 2006,…, 2010.

$TP2Year_i$ takes the value 0 if the paper was published in the first time period (i.e., before 2003) and is equal to $Year_i$ in the second time period. If the estimate of b_3 is significantly different from zero, then the linear relationship in Time Period 2 has a significantly different gradient compared with the relationship in Time Period 1.

ε_{ijk} is the error term associated with the score for paper k in Year i and Group j and TP2Year i.

The model parameters a, b_1, b_2, and b_3 are estimated using least squares.

After assessing the main hypothesis, the model was extended to include other factors that might have influenced the quality score:

[1] http://www.informatik.uni-trier.de/~ley/db/index.html

- A dummy variable indicating whether the paper cited any statistical reference.
- Dummy variables indicating what type of statistical references (if any) were cited.
- Three variables based on the number of papers published by co-authors in preceding years, which were intended to measure the author experience. These were the average of the number of papers published by each co-author in previous years, the maximum experience of any co-author and the experience of the first author.
- Three dummy variables indicating which journal the paper was published in.
- The length of the paper in pages.

The variables were introduced one at a time into the model, and only those that were statistically significant were retained. This process was not used to provide a predictive model but to test whether these variables had any significant relationship with paper quality.

4. Results

4.1 Descriptive Statistics

Figure 1 shows the total quality score per paper per year (averaged over the three assessments for each paper).[2] This figure seems to indicate an increase in quality across the time period with the recent papers including fewer poor quality papers (i.e. a quality score<15). Figure 2 shows the relationship between the total quality score and the subjective assessment of quality. There is a significant linear relationship between the two measures, and the relationship shows no major inconsistencies. The average quality score and the average subjective assessment per year are reported in Table 3. The relationship between the total quality score and year of publication is shown more clearly in Figure 3. This suggests that apart from the years 2001 and 2002, there has been a steady increase in quality over the time period.

The average score for each question in each time period is shown in Table 4. This confirms that the score for each question has improved, particularly Question 7 (Do the authors discuss the limitations of their studies?). However, Table 4 shows that the scores for Question 6 (Do authors discuss potential experimenter bias?) and Question 9 (Is there evidence that the experiment/quasi experiment can be used by other researchers/practitioners?) are still relatively low.

4.2 Regression Analysis

A statistical analysis of the relationship between the quality of individual papers (averaged over the three independent assessments) and year, group and TP2year is shown in Table 5. This analysis, which is based on the data shown in Figure 1, confirms that there is a significant positive linear relationship between year and paper quality, but there is no significant relationship between group and paper quality nor is there a significant change in the gradient of the linear model in TP2. This means that the general trend is one of increasing quality, but there was no major change in the overall trend before 2003 and after 2005.

However, since the quality score has an upper bound of 27, we would expect the gradient of the linear relationship between year and quality to decrease in years following 2010 and indeed there is a slight indication visible in Figure 3, that this effect might be happening in 2009 and 2010

4.3 The Relationship between Statistical Citations and Paper Quality

We checked the references cited in each paper and identified whether the paper cited:

- Statistical texts not written by SE academics.
- Statistical texts and articles produced by SE academics, i.e. [10], [11], [17], [22], [27]. Note that, although most such references occurred in the more recent papers, one paper published in 2001 did make reference to [27].
- Statistical texts written by Campbell and his collaborators that cover issues such as types of validity problems and how to handle quasi-experiments (e.g. [1], [2], [24]).

The number of papers that cited any statistical texts of the above three types in each time period is shown in Table 6. Clearly a majority of papers published after 2005 have cited statistical texts produced by SE academics, but the rate of citing statistical texts from other sources has not changed.

The number of papers that cited *any* statistical source in each time period is shown in Table 7. Note, many papers cited statistical texts of more than one type, so the values in Table 7 cannot be directly derived from Table 6. Table 7 suggests that more recent papers were more likely to cite a statistical text than older papers. A chi-squared test confirmed that there is a statistically significant difference in the citation rate ($p=0.036$).

A linear regression model relating average paper quality to year and citation group (where citation group is a dummy variable taking the value 1 if a paper cites a statistical text and 0 otherwise), shown in Table 8 indicated that both year and citation group are jointly significantly related to average paper quality.

Thus, it appears that better quality papers are likely to cite statistical texts. However, an investigation of

[2] The full data set is available on request from the first author.

the different types of statistical texts being cited (see Table 9), suggests that it is referencing general statistical texts that is most strongly associated with high quality, since the effect of the other types of text is not significant.

4.4 The Relationship between Average Co-Author Experience and Paper Quality

Using the average of the number of papers published by the co-authors in years prior to the paper included in our data set as a measure of experience, we investigated the effect of experience on paper quality in a model including year and statistical texts cited (i.e., only general statistical texts not ones by SE researchers or by Campbell and his colleagues). This analysis confirmed that after accounting for year and referencing statistical texts, average co-author experience was not significantly associated with paper quality (see Table 10). We also tested the maximum experience of the co-authors and the first author experience. Neither of these variables was associated with paper quality.

4.5 The Relationship between Source Journal and Paper Quality

Adding to the basic model three dummy variables identifying which journal the paper was published in indicated that there was no observable difference in the quality of papers from the different journals (see Table 11). Note, it is only possible to include three dummy variables since the effect of the fourth journal is found when the other three journal variables take the value zero.

4.6 The Relationship between Paper Length and Paper Quality

Including paper length as a variable in the baseline model indicated that longer papers were likely to be of higher quality than shorter papers, to a statistically significant degree (see Table 12).

We re-ran the regression analysis for average co-author experience and source journal including paper length in our baseline model (with year and citation of statistical texts) and found co-author experience and source journal were still non-significant.

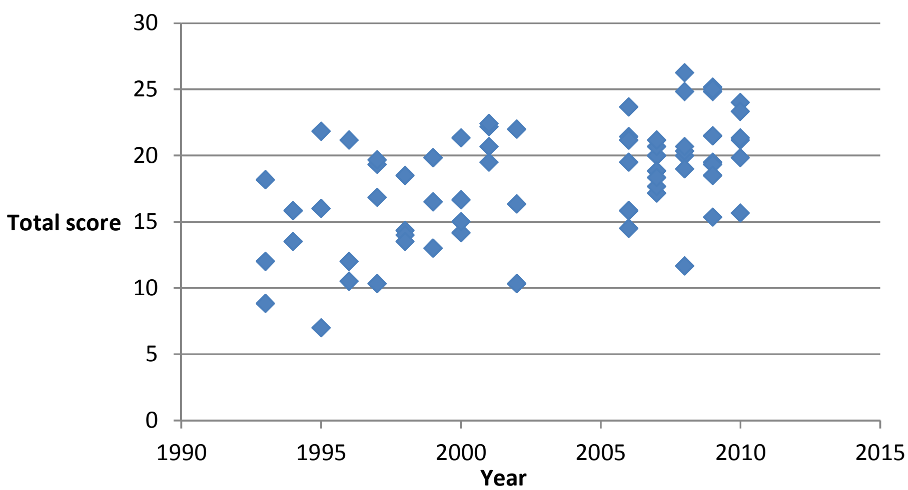

Figure 1 Total quality score per paper per year averaged over the three assessments of each paper

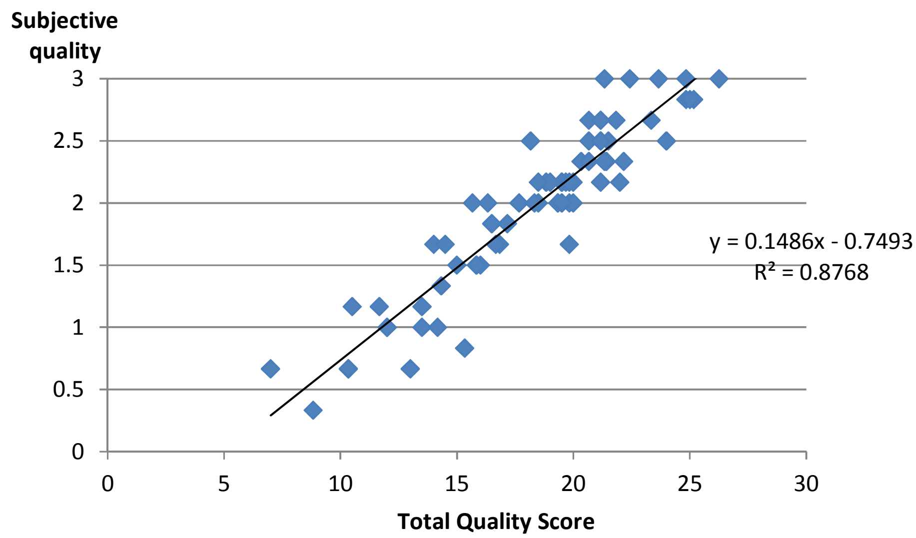

Figure 2 Subjective assessment versus total score per paper

Table 3 Average Quality score and subjective assessment per year

Year	Papers	Average Quality	Variance of average quality	Average Subjective Assessment	Variance of subjective assessment
1993	3	13.000	4.7463	1.278	1.1097
1994	2	14.667	1.6499	1.250	0.3536
1995	3	14.944	7.4728	1.611	1.0046
1996	3	14.556	5.7743	1.611	0.9179
1997	4	16.542	4.3277	1.625	0.6719
1998	4	15.083	2.3034	1.542	0.3696
1999	4	17.292	3.2642	1.583	0.6455
2000	4	16.792	3.2012	1.792	0.8539
2001	4	21.188	1.3649	2.500	0.3600
2002	4	14.750	5.6001	1.375	0.8207
2006	7	19.607	3.2902	2.238	0.5431
2007	7	19.119	1.5267	2.119	0.2673
2008	7	20.393	4.6965	2.310	0.6194
2009	8	21.146	3.6159	2.271	0.6722
2010	6	20.889	2.9771	2.333	0.2789

Table 4 Average score for each quality question

Question Number	Average score for TP1	Average Score for TP2
Q1	2.34	2.65
Q2	2.04	2.39
2.39	2.30	2.63
Q4	2.06	2.40
Q5	2.18	2.52
Q6	0.82	1.26
Q7	1.22	2.57
Q8	1.99	2.44
Q9	1.13	1.69

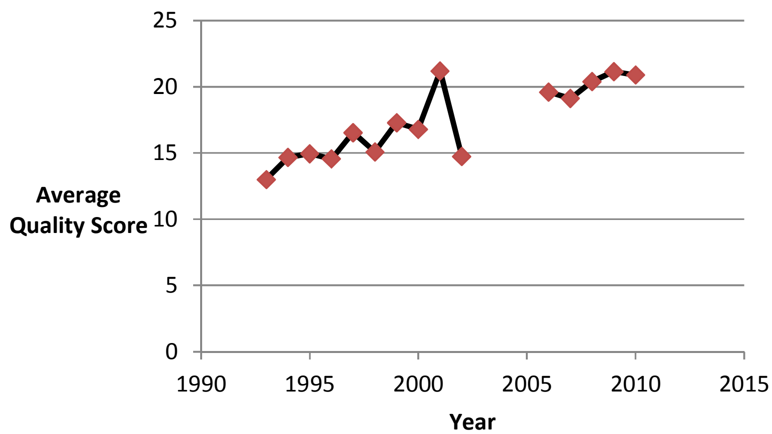

Figure 3 Average quality score per year of papers reporting human-centric experiments

Table 5 Regression analysis of average quality score

Model parameter	Coefficient	Std. Err.	t	P>|t|	[95% Conf. Interval]
year	0.465	0.2276	2.045	0.045	(0.0111,0.9198)
group	-17.552	1023.845	-0.017	0.986	(-2061.725, 2026.62)
TP2year	.0084789	0.5104	0.017	0.987	(-1.0106, 1.0276
constant	-9163.8	454.7	-2.010	0.049	(-1821.6, -5.986)

Table 6 Citation rate of statistical texts

Publication date group	Statistical texts		Statistical texts by SE authors		Texts written by Campbell and colleagues	
	Cited	Not cited	Cited	Not cited	Cited	Not cited
Papers published before 2003	18	17	1	34	7	28
Papers published after 2005	17	18	20	15	13	22

Table 7 Extent to which papers cite statistical texts

Publication date group	Cited no statistical texts	Cited statistical texts
Papers published before 2003	14	21
Papers published after 2005	6	29

Table 8 Regression analysis relating average quality to year and citation group

Model Parameter	Coefficient	Std. Err.	t	P>|t|	[95% Conf. Interval]
year	0.3588	0.0805	4.459	0.000	(0.1982, 0.5193)
citation group	2.627	0.9747	2.696	0.009	(0.6819, 4.5730)
constant	-702.28	160.9	-4.363	0.000	(-1023.5, -381.0)

Table 9 The relationship between citation and paper quality for different types of statistical texts

Parameter	Coefficient	Std. Err.	t	P>|t|	[95% Confidence Interval]
year	0.3268	0.0954	3.411	0.001	(0.135, 0.517)
statistical texts	1.922	0.87667	2.192	0.032	(0.171, 3.673)
statistical texts by SE researchers	1.440	1.1225	1.283	0.204	(-0.802, 3.682)
statistical texts by Campbell and colleagues	1.308	0.9968	1.313	0.194	(-0.682, 3.299)
constant	636.7	191.24	-3.329	0.001	(-1018.7, -254.8)

Table 10 The relationship between the average experience of the co-authors and paper quality

Parameter	Coefficient	Std. Err.	t	P>|t|	[95% Confidence Interval]
year	0.375	0.0905	4.142	0.000	(0.194, 0.556)
statistical texts	1.957	0.8610	2.273	0.026	(0.238, 3.656)
average co-authors experience	0.0222	0.0219	1.014	0.314	(-0.022, 0.066)
constant	-734.0	181.1	-4.054	0.000	(-1095.6, -372.5)

Table 11 The relationship between source journal and paper quality

Parameter	Coefficient	Std. Err.	t	P>|t|	[95% Confidence Interval]
year	0.422	0.0797	5.292	0.000	(0.262, 0.581)
statistical texts	0.735	0.34633	2.124	0.038	(0.044, 1.427)
JSS	1.360	1.4004	-0.971	0.335	(-4.157, 1.438)
ESE	0.622	1.3697	0.454	0.652	(-2.115, 3.358)
TSE	0.426	1.5031	0.284	0.778	(-2.576, 3.429)
constant	-826.8	159.86	5.172	0.000	(-1146.2, -507.5)

Table 12 The relationship between paper length and paper quality

Parameter	Coefficient	Std. Err.	t	P>|t|	[95% Confidence Interval]
year	0.367	0.0776	4.726	0.000	(0.2118, 0.5218)
statistical texts	0.865	0.2974	2.910	0.005	(0.272, 1.459)
paper length	0.0955	0.04119	2.319	0.024	(0.0133, 0.1777)
constant	-719.1	155.22	-4.633	0.000	(-1029.0, -409.2)

4.7 Validation of Results

To assess the stability of our results, we ran our final model (i.e., the regression analysis including year, citation of statistical texts and paper length as independent variables) with the variable *subjective quality* instead of the quality score. Subjective quality was based on an overall assessment of the paper quality made on a four-point ordinal scale (0=Poor, 1=Moderate, 2=Good, 3=Excellent) with interpolation permitted. We used the average of the three values for each paper as our dependent variable. The only difference in our results was that the variable *paper length* just failed to achieve statistical significance at the $p<0.05$ level (the p-value for paper length was 0.055).

The distribution of the variance of the three total quality scores obtained for each paper identifies five papers with unusually large variance (see Figure 4). Four of the papers were published before 2003 and one of the papers was published after 2005. We re-ran our final regression model with the total score as our dependent variable, and omitting these papers. The results were the same in terms of which factors were significant.

With respect to construct validity we analysed four questions separately (Q1, Q5, Q6, Q7). These questions were related to the questions assessed by Dieste et al. (2011), see Table 2. In each case we performed a forward stepwise regression including all the factors investigated previously as independent variables (i.e., year, publication time period, years in the second time period, paper length, average co-author experience, first author experience, maximum co-author experience, citation of statistical texts, citation of statistical texts authored by SE researchers, citation of statistical texts authored by Campbell, whether the paper was published in JSS, TSE or ESE). The results were as follows:

Q1 (corresponding to Dieste et al., Q3): Year and Cites Statistical texts were included in the final model.

Q5 (related to Dieste et al., Q9): Cited texts authored by Campbell and Cited texts authored by SE researchers.

Q6 (related to Dieste et al., Q6): Year and First Author experience were included in the model.

Q7 (related to Dieste et al., Q8): Year and Paper length were included in the model.

In three cases the individual questions were consistent with the total score results with respect to selecting the Year variable, and never selecting the Publication Time Period variable, but were not consistent with respect to other factors. This suggests some confounding between the remaining variables, but generally supports our main hypothesis.

The regression analysis with Q5 (Do authors define data collection procedures and define the measures?) was the only one that did not select Year as a dependent variable. It may be that given the problems associated with defining and collecting metrics in the software and sociology fields, the texts by SE authors and those by Campbell and his co-authors have emphasised defining measurements and data collection issues more than standard statistical texts which tend to assume that data is easy to measure and collect.

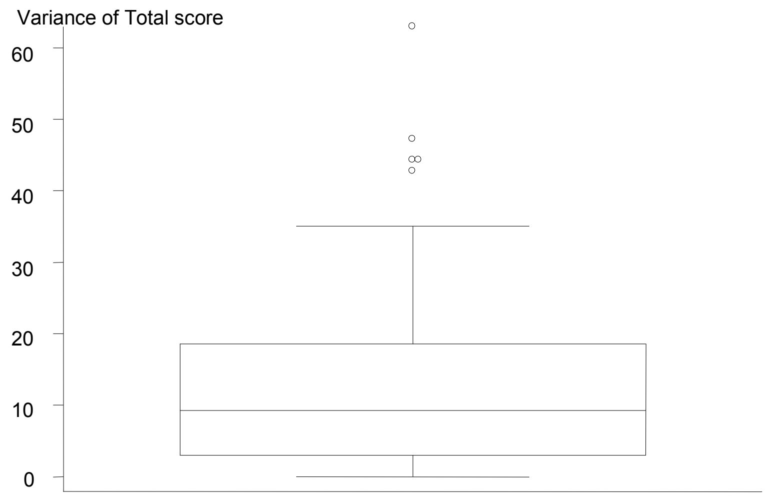

Figure 4 Box plot of the variance of the total score per paper

Table 13 Correlations among quality variables

Variable	Subjective quality	Q1	Q5	Q6
Q1	0.67			
Q5	0.80	0.50		
Q6	0.62	0.51	0.47	
Q7	0.80	0.43	0.68	0.46

In addition, all four variables were significantly correlated to one another and to average subjective quality (see Table 13). This suggests that in spite of the counter-intuitive finding reported in Dieste's study that papers discussing limitations exhibited increased bias, in our study, discussing limitations was positively related to quality. Furthermore, applying principal component analysis to the scores for all the 9 individual questions, the first factor, which was a weighted average of the 9 questions (with all weights being fairly similar), accounted for 61% of the variation, providing added confirmation that the total score (as a simple sum of the 9 questions) is a reasonable measure of quality.

5. Discussion

Our results show clearly that there has been an overall increase in quality as measured by our quality assessment instrument for the four journals across the time period of our study (Table 2). During the first four years of the time period, the average quality of papers each year was less than 15 (out of a total possible score of 27). During the final three years, the average quality of papers each year exceeded 20.

Our statistical analysis confirmed there was a significant relationship between year and quality but, because there was no significant effect due to time period (i.e., before 2002 and after 2005), there no evidence that this increase in quality was directly caused by referencing the articles and texts on methodological issues written by SE researchers. In this study, we observed a gradual increase in quality across the years rather than a dramatic change in the more recent papers.

In addition to a year effect, we also found that the total quality score was significantly associated with referencing general statistical text books but not with referencing text books written by SE researchers. Thus, although we observed a change in citation practice in papers published after 2005, it is likely that the improvement was due to a general increase in the understanding of empirical SE rather than the availability of statistical material written by SE researchers. However, although the relationship between quality and referencing statistical texts was stable when we analysed the average subjective assessment, it was not completely stable for the detailed analyses of individual questions, so there may be confounding effects among the variables we investigated. For example, it may be that researchers preferred to reference the primary source material for statistical methods rather than secondary sources provided by software engineers. It may also be the case that experienced authors, who write high-quality papers, do use the statistical literature published by SE researchers, but because they are aware that it is very well-known and can be considered part of the common body of knowledge within the field, they do not cite it explicitly. Conversely, newcomers may not be very familiar with this literature and therefore may be more inclined to cite it explicitly.

We also found that the total quality score was associated with paper length although this relationship was not observed when we analysed the average subjective assessment, nor was it stable for the individual questions we analysed. Thus, the relationship between paper length and quality assessed using the total quality score may be due to the emphasis on reporting in the quality questionnaire, since reporting more information is easier in longer papers.

6. Limitations

A major limitation of our study is that our quality questionnaire may not properly represent the quality of an experiment, but rather the quality of the reporting in a paper. To address this issue, we used two different methods of measuring quality, one based on a questionnaire and one based on an overall subjective assessment, which confirmed that our results with the exception of the relationship with paper length were essentially the same. Furthermore, we checked that individual questions related to those that Dieste et al. [5] found to be related to be negatively related to bias (i.e., positively related to quality) behaved in a similar way to our overall quality score. All the questions confirmed that *year* was a significant variable and *time period* was not, although the inclusion of other factors varied.

Another important limitation is the possibility of *experimenter expectation*, i.e., our basic hypothesis was that paper quality would be improving over time and we might, therefore, have unintentionally marked papers published before 2003 lower than they deserved. The use of the quality questionnaire was intended to lower the probability of experimenter bias and our results for year 2001 (which scored higher than any other year) tends to support the view that our assessments were not biased against papers published before 2003. However, we all knew the date of each paper, so we cannot be certain that we were not subconsciously influenced. We note that attempts to blind ourselves to the date of papers would not have been effective since Kitchenham needed to assign the papers and, therefore, had to know the papers' pubication dates, and Sjøberg and Dybå had already studied the papers published in TP1 in considerable detail. Furthermore all of us act as reviewers for the four journals and had seen some of the papers in that capacity and/or had studied some of the papers as part of our own research.

Another limitation is that our assessments of quality may not be reliable in the sense that two judges using the same quality instrument might come to very different conclusions. We reported a series of studies that used the questionnaire [18] and found that 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 the second study and suggested that inter-rater reliability was poor for all groups but worse for teams of two or three than for individuals. These results confirm that it is difficult to achieve high levels of reliability. We have addressed this difficulty by using three judges per paper and checking that our results were robust when removing papers that did not exhibit good reliability (i.e., exhibited large variance in the total quality scores).

We have restricted ourselves to assessing the quality of journal papers, although there were 14 papers appearing in the International Conference of Software Engineering (ICSE) in TP1. This has increased the homogeneity of our set of papers at the expense of reducing the generality of our conclusions.

A problem with respect to statistical conclusion validity is our stepwise introduction of additional variables into our initial model (see Equation 1). It is dangerous to undertake a large number of tests on the same data set since some effects may be found by chance. However, an important part of our study was to investigate the impact of the SE guidelines, and we felt it was also important to also investigate possible confounding variables such as paper length, other sources of guidelines, and specific journals. In our final model (see Table 12), the sum of the p-values for the three variables included in our model was less than 0.05, so we have not substantially inflated the type-1 errors [20]. Nonetheless, the test of our initial model must be treated as the test of our main hypothesis, and the tests of other variables should be treated as exploratory.

We found it difficult to devise a good measure of the capability of the set of authors working in a specific paper. We used several surrogate measures but we are not convinced that they adequately captured the concept of team capability.

Finally, we assumed that limiting the number of papers with the same author as *first* author would avoid potentially biasing our results. Table 14 suggests that *failing* to restrict papers from the same first author would have increased the average first author experience and average quality of papers published before 2003. However, among papers published after 2006 the average first author experience would have increased without greatly affecting the average quality score. Thus, if we had included more papers by the same first authors, we would have increased the average quality score for papers published before 2003 due to the excellence of individual authors rather than the general standard of authors. This implies that our selection process has successfully removed a possible cause of bias.

Table 14 Mean and median (in parenthesis) first author experience and average quality score for first authors with multiple papers compared with other authors

Data set	Variable	TP1	TP2
Excluding multiple papers with the same first author	Observations	32	30
	Experience	5.9 (3)	15.6 (9)
	Average Quality Score	15.6 (15.9)	20.5 (20.9)
Including multiple papers with the same first author	Observations	3	5
	Experience	13 (10)	54 (35)
	Average Quality Score	21.6 (21.8)	18.5 (19.5)

7. Conclusions

As SE researchers, we are pleased to find that the quality of experimental and quasi-experimental SE papers appears to be improving. However, although the recent texts authored by SE researchers have had a significant impact on citation practices, there is no evidence that the change in citation practice is *directly* associated with the improvement in quality over the monitored time period.

The results of our study suggest that the quality improvement is due to a gradual increase across the entire time period 1993-2010. Our analysis of citations attributes this to a general increase in the level of understanding of experimental and statistical methods rather than specific initiatives by SE researchers. Indeed, the initiatives that led to new SE conferences and journals addressing empirical SE in the late 1990's and the later statistical text books and guidelines could actually have been a result of the initial increase in understanding of statistical methods and experimental design.

Our study was based on papers that were published inonly four SE journals (TSE, JSS, ESE, IST). These are high quality venues for SE experiments. Thus, we would expect the quality of software experiments and quasi-experiments published in these sources to be higher than that obtained in other sources. In particular, we do not know whether the results generalise to conference papers, which are usually constrained to be shorter than journal papers and so may score poorly on a quality instrument that favours reporting quality. Nonetheless, performing a similar study based on papers from ICSE and Empirical Software Engineering and Measurement (ESEM) might be an interesting topic for future research. However, ESEM papers would have to be compared with papers from the Metrics and ISESE conferences if the same time periods were used.

We used a quality instrument that we developed ourselves; see [18]. Although there were overlaps, Dieste et al. [4] used a rather different set of quality questions for their study of the relationship between bias and quality questions. This raises the question of whether there is a "best" set of criteria for human-centric SE experiments and quasi-experiments. Dieste et al.'s results suggested that only three of their 10 questions were negatively related to bias. In contrast, our results suggest that all our questions were positively associated with quality. Thus, we cannot be sure which set of questions are best, nor indeed whether it is possible to identify a best set of questions given the different suggestions made by different researchers (e.g. [3], [9]). An alternative approach to assessing study quality is to assess specific well-defined criteria such as power, effect size, and quasi-experiment practices as has been done for studies published prior to 2003 ([6], [14], [17]). These criteria could be used both to investigate improvements in study quality over the time period 1993-2010, and to assess the validity of alternative quality instruments.

Our study showed that all of the quality questions had a larger average score in TP2 than in TP1 (see Table 4). However, there are two low scores for questions that suggest how researchers might improve the quality of their studies, i.e. Question 6 and Question 9.

Question 6 (Do the authors discuss potential experimenter bias?) scored an average of only 1.26 in TP2. This is probably because experimenter bias is not emphasized in sources such as [27], [24] and [2], which provide lists of validity issues related to Conclusion Validity, Internal Validity, External Validity and Construct Validity. This can be contrasted with the average score for Question 7 (Do authors discuss the limitations of their study?), which increased from 1.22 in TP1 to 2.57 in TP2. However, the issue of human experimenters studying techniques, methods or procedures that depend on the skills of human participants has many opportunities for bias. We recommend researchers refer to [21] for a detailed discussion of relevant issues.

Question 9 (Is there evidence that the Experiment/Quasi-Experiment can be used by other researcher/practitioners?) scored an average of 1.13 in TP1 and 1.69 in TP2. This may be because it is easy to recommend that researchers report how their results can be used but it is quite hard to suggest more operational guidelines. We suggest authors report avenues for further which researchers might find valuable. Furthermore, including a discussion of how robust any benefits found in the experiment were and the implications of any limitations might benefit practitioners. For example authors might indicate whether a better method/procedure requires extensive training and/or tool support to be viable in practice.

Acknowledgements

We thank the referees for their constructive comments,

References

[1] Campbell, D.T. and Stanley, J.C. (1966) Experimental and Quasi-experimental Designs for Research. Houghton Mifflin Company.

[2] Cook, T.D. and Campbell, D.T. (1979) Quasi-experimentation: Design and Analysis Issues for Field Settings. Rand McNally Collage Chicago.

[3] Crombie, I.K. (1996) The Pocket Guide to Appraisal, BMJ Books.

[4] Dieste, O. and Padua, A.G. (2007) Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews. First International Symposium on Empirical Software Engineering and Measurement, pp. 215-224.

[5] Dieste, O. Grimán, A. Juristo, N. and Saxena, H. (2011) Quantitative determination of the relationship between internal validity and bias in software engineering: consequences for systematic literature reviews, International Symposium on Empirical Software Engineering and Metrics (ESEM), pp. 285-28.

[6] Dybå, T., Kampenes, V.B. and Sjøberg, D.I.K. (2006) A systematic review of statistical power in software engineering experiments, Information and Software Technology, 48(8), pp. 745-755.

[7] Fisher LD, Dixon DO, Herson J, Frankowski RK, Hearon MS, Pearce KE. (1990) Intention to treat in clinical trials. In: Pearce KE, editor. Statistical issues in drug research and development. New York: Marcel Dekker, pp. 331–350.

[8] Fink, A. (2005) Conducting Research Literature Reviews. From the Internet to Paper, Sage Publication, Inc.

[9] Greenhalgh, T. (2000) How to read a paper: The Basics of Evidence-Based Medicine. BMJ Books.

[10] Jedlitschka, A., Ciolkowski, M. and Pfahl, D. (2008) Reporting Experiments in Software Engineering. In Guide to Advanced Empirical Software Engineering, Shull, F., Singer, J. and Sjøberg, D.I.K. (eds), Springer-Verlag London Ltd.

[11] Juristo, J. and Moreno, A. (2001) Basics of software engineering experimentation, Kluwer Academic Publishers, Boston, M.A.

[12] Jüni, P. Witschi, A., Bloch, R. and Egger, M. (1999) The Hazards of Scoring the

Quality of Clinical Trials for Meta-Analysis. JAMA, 282 (11), pp. 1054-1060.

[13] Liu, H. and Tan, H.B.K. (2007) Testing input validation in Web applications through automated model recovery. Journal of Systems and Software, 81, pp 222–233.

[14] Kampenes, V.B., Dybå, T., Hannay. J.E. and D. I. K. Sjøberg. (2007) A Systematic Review of Effect Size in Software Engineering Experiments, Information and Software Technology, 49 (11-12), pp. 1073-1086.

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

[16] Kampenes, V.B., Dybå, T., Hannay. J.E. and D. I. K. Sjøberg. (2009) A Systematic Review of Quasi-Experiments in Software Engineering, Information and Software Technology, 51 (1), pp. 71-82.

[17] Kitchenham, B., Pfleeger, S.L., Pickard, L.M., Jones, P., Hoaglin, D., El Emam, K. and Rosenberg, J. (2002) Preliminary Guidelines for Empirical Research in Software Engineering. IEEE Transactions on Software Engineering, 28(8), pp. 721-734.

[18] Kitchenham, B.A., Sjøberg, D.I.K., Dybå, T., Pfhal, D., Brereton, P., Budgen, D., Höst, M. and Runeson, P. (2012) Three empirical studies on the agreement of reviewers about the quality of software engineering experiments. Information and Software Technology, 54, pp 804-819.

[19] Kitchenham, B.A., Sjøberg, D.I.K., Brereton, O.P., Budgen, D., Dybå, T., Høst, M., Pfahl, D. and Runeson, P. (2010) Can we evaluate the quality of Software Engineering Experiments? Proceedings of the Conference on Empirical Software Engineering and Metrics, ESEM 2010.

[20] Rosenberger, W.F. (1996) Dealing with multiplicities in pharmacoepidermioloical studies. Pharmacoepidemiology and Drug Safety, 5, pp 95-100.

[21] Rosnow, R.L. and Rosenthal, R. (1997) People studying people. Artifacts and Ethics in Behavioural Research. W.H. Freeman and Company, New York.

[22] Singer, J. (1999) Using the APA style guidelines to report experimental results. In Proceedings of the Workshop on Empirical Studies in Software Maintenance, pp. 71-75.

[23] Sjøberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanovic, A., Liborg, N.K. and Rekdal, A.C. (2005) A survey of controlled experiments in software engineering. IEEE Transactions on SE, 31 (9), pp.733-753.

[24] Shadish, W.R., Cook, T.D and Campbell, D.T. (2002) Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, Boston & New York.

[25] Shrout, P.E. and Fleiss, J.L. (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, 86(2), pp. 420-428.

[26] Wagner, A.K., Soumerai, S.B., Zhang, F., and Ross-Degnan, D. (2002) Segmented regression analysis of interrupted time series studies in medication use research. J.Clin.Pharm.Ther. 27, pp 299-309.

[27] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B. and Wesslén, A. (2000) Experimentation in Software Engineering – An Introduction. Kluwer, Academic press, Boston, M.A.

[28] Wojcicki, M.A. and Strooper, P. (2007) Maximising the information gained by a study of static analysis technologies for current software. ESE, 12 (6), pp. 617-645.