SOA based and Quality of Human-Centric Experiments—A Quasi-Experiment over Software Engineering

Banu Chandra
M.Tech student
GIET
Rajahmundry, A.P, India

G.Rosline Nesa Kumari, Ph.D.
Associate Professor
Saveetha School of Engineering
Chennai, T.N, India

S.Maruthuperumal, Ph.D.
Prof. and HOD CSE and IT
GIET
Rajahmundry, A.P, India

ABSTRACT
Research into how humans interact with computers has a long and rich history. Only a small fraction of this research has considered how humans interact with computers when engineering software. A similarly small amount of research has considered how humans interact with humans when engineering software. For the last forty years, we have largely taken an artifact-centric approach to software engineering research. To meet the challenges of building future software systems, I argue that we need to balance the artifactcentric approach with a human-centric approach, in which the focus is on amplifying the human intelligence required to build great software systems. A human-centric approach involves performing empirical studies to understand how software engineers work with software and with each other, developing new methods for both ecomposing and composing models of software to to ease the cognitive load placed on engineers and on creating computationally intelligent tools aimed at focusing the humans on the tasks only the humans can solve. Context: Several text books and papers 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: Seventy experimental and quasi experimental papers published in four general software engineering journals in the years1992-2002 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, and average coauthor experience, citation of statistical text books and papers, 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 atistical texts, and paper length ($p < 0.05$). Paper length did not reach significance when quality was measured using an overall subjective assessment.

Keywords
Component; formatting; style; styling; insert (key words)

1. INTRODUCTION
Science action writers often speculate about situations in which software is intelligent, sufficiently so to perhaps even program itself. Perhaps luckily, we have not yet entered into situations where software can determine its own actions or evolve to meet new needs. Rather, software engineering is, and should remain, a human-intensive activity. Despite the central role of humans using the tools, methods and processes that support software engineering, the focus of much software engineering research is on improving artifacts that support, or are the end goal, of the engineering rather than on ensuring the abilities of the humans involved in the activities of engineering the software are amplied to the greatest degree possible. As one example, a substantial amount of research considers how to express the abstractions describing software in models rather than in source code. However, little attention has been paid to how the software engineers using the models reason about the eventual system through the models. I argue that a human-centric approach to software engineering can help accelerate our ability to build complex software systems with desired qualities. A human-centric approach would involve research focused on how humans work with computational structures and with each other. A human-centric approach would also consider extensions to existing research to consider how humans can work with artifact centric research results. Finally, such an approach would involve the development of limited intelligence models and tools to allow software engineers to focus those aspects of a development project requiring human creativity and judgment. To give a sense of human centered versus artifact-centered results, I first outline the differences between the two approaches in terms of vignettes in three areas of software engineering research results: tools, methods and processes. I then sketch how research agendas might change to accommodate human-centered software engineering.

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 summarized research related to quality criteria used to evaluate experiments [18], pointing out that quality criteria in medical studies were based on three issues: 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 summarized research related to quality criteria used to evaluate experiments [18], pointing out that quality criteria in medical studies were based on three issues:

- use of random allocation to experimental conditions,
- use of single-blind versus double blind procedures,
- How dropouts were analyzed.

Furthermore, we noted that there are some doubts about using checklists based on more general criteria to assess medical studies. For SE studies, we argued that double blind procedures and the intention to treat method 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 analyzed 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: The basic method used in this quasi-experiment was to select a set of papers 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. 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 with the aim of investigating whether the publication of SE guidelines on performing experiments 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 2000s 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 coauthors of this paper. The papers were obtained from two sources. Papers published in or before 2002 were selected from the 76 papers (of 103 papers) found in [23] that were published in four international journals: the 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 an impact on experiment quality. If the guidelines 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 organizers to review empirical SE studies. Therefore, we are representative of reasonably expert empirical researchers 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. 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 analyzed. We also used the following exclusion/inclusion criteria:. We excluded papers coauthored 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.

A quality questionnaire, which is 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 ¼ Moderate; 2 = Good; 3 = Excellent) 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 Intra Class 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 randomized 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.

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 Fig. 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 Fig. 3 that this effect might be happening in 2009 and 2010.

3. THE RELATIONSHIP BETWEEN AVERAGE COAUTHOR EXPERIENCE AND PAPER QUALITY

Using the average of the number of papers published by the coauthors in years prior to the paper included in our dataset 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 et al.). This analysis confirmed that after accounting for year and referencing statistical texts, average coauthor experience was not significantly associated with paper quality.

Table.1 Paper Quality

Year	Papers	Average quality	Variance of average quality	Average subjective assess-ment	Variance of subject-tive assess-ment
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

4. 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 1990s 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 in only four SE journals (TSE, JSS, ESE, IST). These are highquality 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 generalize to conference papers, which are usually constrained to be shorter than journal papers and so may score poorly on a quality instrument that favors 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 is best nor indeed whether it is possible to identify a best set of questions given the different suggestions made by different researchers. 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 . 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.

5. REFERENCE

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy Soc. London, vol. A247, pp. 529–551, April 1955, (references)

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

[3] T.D. Cook and D.T. Campbell, Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally Collage, 1979.

[4] I.K. Crombie, The Pocket Guide to Appraisal. BMJ Books,1996.

[5] O. Dieste and A.G. Padua, "Developing Search Strategies forDetecting Relevant Experiments for Systematic Reviews," Proc.First Int'l Symp, Empirical Software Eng. and Measurement,pp. 215-224, 2007.

[6] O. Dieste, A. Grima´n, N. Juristo, and H. Saxena, "Quantitative Determination of the Relationship between Internal Validity and Bias in Software Engineering: Consequences for Systematic Literature Reviews," Proc. Int'l Symp. Empirical Software Eng. And Metrics, pp. 285-288, 2011

[7] T. Dyba°, V.B. Kampenes, and D.I.K. Sjøberg, "A Systematic Review of Statistical Power in Software Engineering Experiments," Information and Software Technology, vol. 48, no. 8, pp. 745- 755, 2006

[8] L.D. Fisher, D.O. Dixon, J. Herson, R.K. Frankowski, M.S.Hearon, and K.E. Pearce, "Intention to Treat in Clinical Trials," Statistical Issues in Drug Research and Development, K.E. Pearce, ed., pp. 331- 350, Marcel Dekker, 1990.

[9] A. Fink, Conducting Research Literature Reviews: From the Internet to Paper. Sage Publication, Inc., 2005.

[10] T. Greenhalgh, How to Read a Paper: The Basics of Evidence-Based Medicine. BMJ Books, 2000. A. Jedlitschka, M. Ciolkowski, and D. Pfahl, "Reporting Experiments in Software Engineering," Guide to Advanced Empirical Software Eng., F. Shull, J. Singer, and D.I.K. Sjøberg, eds., Springer- Verlag, 2008.

[11] J. Juristo and A. Moreno, Basics of Software Engineering Experimentation. Kluwer Academic Publishers, 2001

[12] P. Ju¨ ni, A. Witschi, R. Bloch, and M. Egger, "The Hazards of Scoring the Quality of Clinical Trials for Meta-Analysis," J. Am. Medical Assoc., vol. 282, no. 11, pp. 1054-1060, 1999.

[13] H. Liu and H.B.K. Tan, "Testing Input Validation in Web Applications through Automated Model Recovery," J. Systems and Software, vol. 81, pp. 222-233, 2007

[14] V.B. Kampenes, T. Dyba°, J.E. Hannay, and D.I.K. Sjøberg, "A Systematic Review of Effect Size in Software Engineering Experiments," Information and Software Technology, vol. 49, no. 11/12, pp. 1073-1086, 2007

[15] V.B. Kampenes, "Quality of Design Analysis and Reporting of Software Engineering Experiments: A Systematic Review," PhD thesis, Dept. of Informatics, Univ. of Oslo, 2007.

[16] V.B. Kampenes, T. Dyba°, J.E. Hannay, and D.I.K. Sjøberg, "A Systematic Review of Quasi-Experiments in Software Engineering," Information and Software Technology, vol. 51, no. 1, pp. 71- 82, 2009.

[17] B. Kitchenham, S.L. Pfleeger, L.M. Pickard, P. Jones, D.Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary Guidelines for Empirical Research in Software Engineering," IEEE Trans. Software Eng., vol. 28, no. 8, pp. 721-734, Aug. 2002.

[18] B.A. Kitchenham, D.I.K. Sjøberg, T. Dyba°, D. Pfhal, P. Brereton, D. Budgen, M. Ho¨ st, and P. Runeson, "Three Empirical Studies on the Agreement of Reviewers about the Quality of Software Engineering Experiments," Information and Software Technology, vol. 54, pp. 804-819, 2012.

[19] B.A. Kitchenham, D.I.K. Sjøberg, O.P. Brereton, D. Budgen, T. Dyba° , M. Høst, D. Pfahl, and P. Runeson, "Can We Evaluate the Quality of Software Engineering Experiments?" Proc. Conf. Empirical Software Eng. and Metrics, 2010.

[20] W.F.Rosenberger, "Dealing with Multiplicities in Pharmacoepidermioloical Studies," Pharmacoepidemiology and Drug Safety, vol. 5, pp. 95-100, 1996.

[21] R.L. Rosnow and R. Rosenthal, People Studying People. Artifacts and Ethics in Behavioural Research. W.H. Freeman and Company, 1997.

[22] J. Singer, "Using the APA Style Guidelines to Report Experimental Results," Proc. Workshop Empirical Studies in Software Maintenance, pp. 71-75, 1999

[23] D.I.K. Sjøberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N.K. Liborg, and A.C. Rekdal, "A Survey of Controlled Experiments in Software Engineering," IEEE Trans. Software Eng., vol. 31, no. 9, pp.733-753, Sept. 2005.

[24] W.R. Shadish, T.D Cook, and D.T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, 2002.

[25] P.E. Shrout and J.L. Fleiss, "Intraclass Correlations: Uses in Assessing Rater Reliability," Psychological Bull., vol. 86, no. 2, pp. 420-428, 1979.

[26] A.K. Wagner, S.B. Soumerai, F. Zhang, and D. Ross-Degnan, "Segmented Regression Analysis of Interrupted Time Series Studies in Medication Use Research," J. Clinical Pharmacy Therapeutics, vol. 27, pp. 299-309, 2002

[27] C. Wohlin, P. Runeson, M. Ho¨ st, M.C. Ohlsson, B. Regnell, and A. essle´n, Experimentation in Software Engineering—An Introduction. Kluwer, Academic Press, 2000.

[28] M.A. Wojcicki and P. Strooper, "Maximising the Information Gained by a Study of Static Analysis Technologies for Current Software," Empirical Software Eng., vol. 12, no. 6, pp. 617-645, 2007