



Contents lists available at ScienceDirect

The Journal of Systems and Software

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

An update to experimental models for validating computer technology

Marvin V. Zelkowitz

Department of Computer Science, University of Maryland, College Park, Maryland 20742, USA

ARTICLE INFO

Article history:

Received 3 March 2008

Accepted 22 June 2008

Available online xxx

Keywords:

Data collection

Experimentation

Technology evaluation

ABSTRACT

In 1998 a survey was published on the extent to which software engineering papers validate the claims made in those papers. The survey looked at publications in 1985, 1990 and 1995. This current paper updates that survey with data from 2000 to 2005. The basic conclusion is that the situation is improving. One earlier complaint that access to data repositories was difficult is becoming less prevalent and the percentage of papers including validation is increasing.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

In 1998, a paper by Zelkowitz and Wallace (1998) surveyed the software engineering literature in order to classify the experimental methods used by authors to validate the technical claims made in those papers. The basic conclusion was that approximately half of the papers had an inadequate level of validation. A similar survey by Tichy et al. (1995) came up with a similar conclusion. However, one of the conclusions in the Zelkowitz and Wallace paper was that the situation seemed to be improving. Since the 1998 survey evaluated papers from 1985, 1990 and 1995, and since two more 5-year milestones have since passed, it is worthwhile to revisit that initial survey to see how the research world has changed in the approximately 10 years since the original survey was conducted.

Table 1 presents the basic data from both the original and current survey. The three data sources for this survey were:

- *icse* – Proceedings of the International Conference on Software Engineering.
- *tse* – IEEE Transactions on Software Engineering.
- *sw* – IEEE Software Magazine.

A total of 612 papers were evaluated earlier and 50 of them were deemed not applicable, leaving 562 research papers. In the current survey covering 2000 and 2005, an additional 346 papers were evaluated and 47 were considered not applicable, leaving 299 research papers.

Each of the research papers was classified according to a 14-category taxonomy. Eleven of categories represented various empiri-

cal validation methods. (See Appendix.) There was a twelfth theoretical method indicating the paper was a formal model of some property and there was a thirteenth quasi-validation method, called an assertion. Assertion papers were those where the author knew that an experimental validation would be appropriate, but only a weak form of validation was applied. (For example, a paper describing a new programming language might only show that it was possible to write programs in that language, not whether the programming language solved any underlying problem that needed to be solved.) All other papers were characterized as “No experimentation,” indicating that some form of validation was appropriate, but was lacking.

Note that in the published version of the 1998 paper, “Theoretical” and “No experimentation” were considered a single classification. Subsequent to its publication, the two categories were separated since we considered a purely formal paper as not necessarily being appropriate for empirical validation.

2. Observations

The percentages for each validation method are given in Fig. 1. Several trends are clear. Case study remains the most popular method, except for dynamic analysis in 2005 and lessons learned in 1995, and its popularity is slowly rising from 8% in 1985 to 21% in 2005. The “classical” experimentation method of a controlled replicated study (represented as both synthetic and replicated in Fig. 1) grew slightly to 7% of the papers in 2005.

More important than individual methods is the general “health” of the field. This is summarized by Fig. 2. Except for 2000, the percent of “No experimentation” papers dropped from 27% in 1985 to only 16% in 2005. Assertions dropped from 35% in 1985 to 19% in 2005. The percent of papers that used one of the 11 validation

E-mail addresses: mvz@cs.umd.edu, marv@zelkowitz.com

Table 1
Basic classification data from 958 papers: 1985–2005

	Project monitoring	Case study	Field study	Literature search	Legacy	Lessons learned	Static analysis	Replicated	Synthetic	Dynamic analysis	Simulation	Assertion	Theoretical	No experimentation	Not applicable	Total
icse	0	5	1	1	1	7	1	1	3	0	2	12	3	13	6	56
tse	0	12	1	3	2	4	1	0	1	0	10	54	18	38	3	147
sw	0	2	0	1	1	5	0	0	1	0	0	13	1	10	6	40
1985 Total	0	19	2	5	4	16	2	1	5	0	12	79	22	61	15	243
icse	0	7	0	1	2	1	0	0	0	0	0	12	1	7	4	35
tse	0	6	1	1	2	8	0	1	4	3	11	42	19	22	2	122
sw	1	6	0	5	0	4	0	0	1	0	0	19	0	8	16	60
1990 Total	1	19	1	7	4	13	0	1	5	3	11	73	20	37	22	217
icse	0	4	1	0	1	5	0	1	0	0	1	4	3	7	5	32
tse	0	10	2	2	1	8	2	3	4	4	6	22	7	7	1	77
sw	0	6	1	3	1	7	0	0	0	0	1	14	0	3	7	43
1995 Total	0	20	4	5	3	20	2	4	2	4	8	40	10	17	13	152
icse	0	10	0	0	1	4	0	2	2	4	1	11	3	20	10	68
tse	0	9	3	1	0	0	0	0	4	4	7	11	10	15	2	66
sw	0	6	4	0	0	0	0	0	1	3	1	3	0	19	14	51
2000 Total	0	25	7	1	1	4	0	2	7	11	9	25	13	54	26	185
icse	0	14	1	0	1	0	0	0	3	8	1	10	1	3	0	42
tse	0	9	4	1	5	0	2	1	2	13	5	13	1	8	2	66
sw	0	7	3	1	1	3	0	0	3	0	0	4	1	11	19	53
2005 Total	0	30	8	2	7	3	2	1	8	21	6	27	3	22	21	161

methods rose from 29% to 63% in 2005. (The percentage rose from 39% to 65% when theoretical papers were included.) Clearly the situation is improving.

This result is consistent with an alternative study of the ICSE conferences (Zannier et al., 2006). Using a sampling technique over all 29 ICSE proceedings (since 1975), they found that 19 of 63 papers included no empirical study (30%). This present study indicates that 50 out of 208 ICSE papers (24%) since 1985 had no experimentation. As a simple comparison, if we assume that (Zannier et al., 2006) represents papers from all ICSEs and our study of the five ICSE proceedings is representative of all ICSEs from 1985 though 2005, then (Zannier et al., 2006) represents about a 36% no experimentation rate for the early ICSEs in order to get to a 30% overall rate. That is then a 1/3 drop from 36% to 24% over the life of the conference. They also found a statistically significant increase in evaluation papers between those proceedings prior to 1990 and those since.

Unlike in Zannier et al. (2006), no attempt was made to evaluate the quality of the validation. (It was beyond our knowledge to attempt to understand and evaluate all 861 research papers.) If the paper stated some hypothesis about the technology described in the paper (even if stated indirectly) and then proceeded to describe a validation method of that hypothesis, we considered it as validated. For example, an experiment could be extremely well done from a statistical viewpoint and still be pointless due to poor design (e.g., if industrial experience necessary to answer a question addressed by a survey of students).

Several other anecdotal observations are buried in the data. A common complaint 20 years ago was the lack of published data sources that others could have access to. That is changing. Many of the papers used the various open source repositories, looking at the development history of products such as the Apache Web server or Mozilla, as sources for data. This use of historical data using open source and other data repositories was one of the reasons for the rise in the dynamic simulation category in Fig. 1. Similarly, data mining through these sources led to the rise in the legacy data category.

3. Conclusions

There are several threats to the validity of this study.

1. The recent classification was performed about nine years after the earlier study. While every attempt was made to use the same classification process, undoubtedly the intervening years may have changed our views of some of the validation methods (e.g., in Sjøberg et al., 2005, the authors report 0 and 3 controlled studies in ICSE for 1995 and 2000, while Table 1 in this paper shows 1 and 4, respectively, for those years). While this may have affected individual percentages, it should not have had much of an impact on the overall results.
2. As with the earlier study, each publication for each year was managed by a different editor or conference chair. This has an effect on the overall acceptance rate of various papers submitted to that source. For example, the rise in “No experimentation” in 2000 was partially due to the largest number of ICSE papers (68) and the relatively large number of “No experimentation” papers (20) accepted to that proceeding. Although such variances affect individual sources in a given year, the overall trends seem consistent.
3. There was a change in the scope of *IEEE Software* between 1995 and 2000. In the earlier survey, this magazine often published longer articles that had a research component. However, more recently the papers are limited to be shorter with

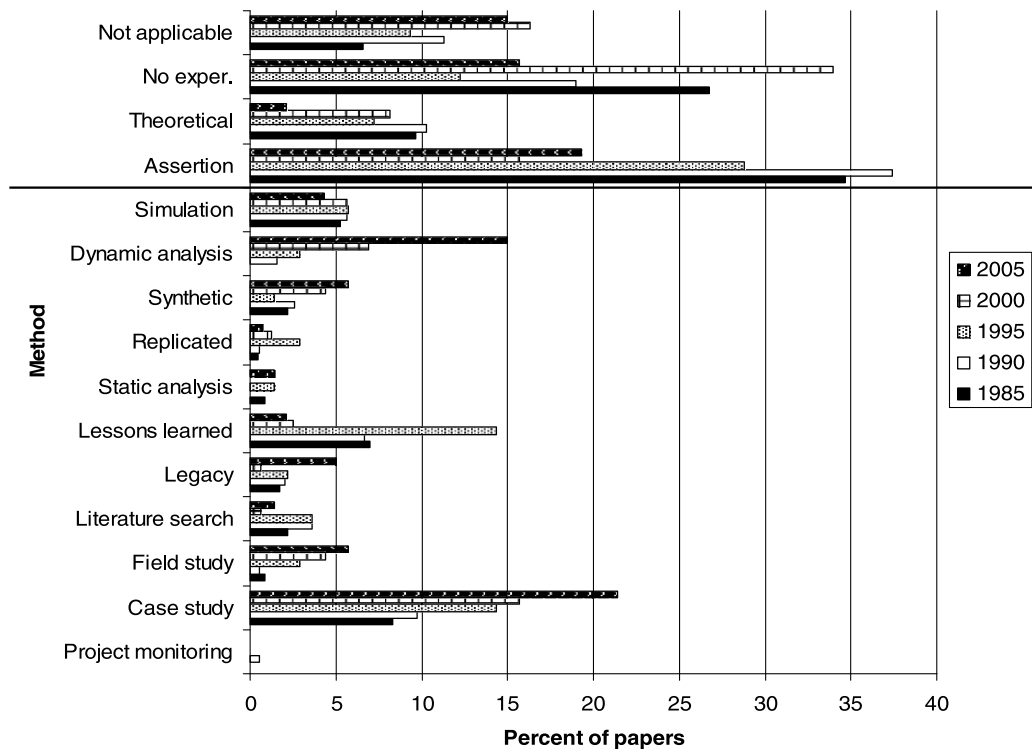


Fig. 1. Percentages of each validation method.

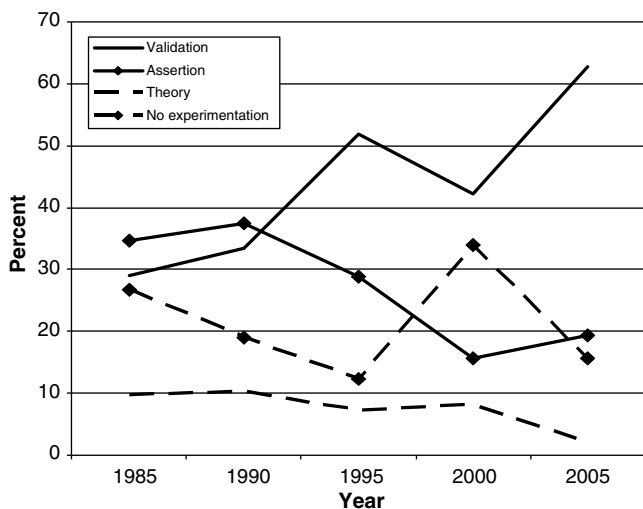


Fig. 2. Changes over time in validated papers.

more regular columns appearing in each issue. Regular columns were not included in this survey and a value judgment was made on the remainder of the papers. If a paper discussed many solutions to a given problem, the paper was considered a tutorial or survey and listed as “Not applicable,” but if the paper focused on a particular technique (often the author’s), then it was considered a research paper. This is clear from the “Not applicable” column of Table 1, where the percent of “Not applicable” for *IEEE Software* for 1985–1995 was 20% and rose to 32% for 2000–2005. In addition, the shorter length of *IEEE Software* papers since 1995 probably also contributed to the rise in “No experimentation” for this magazine from 18% in the earlier period to 42% in the new study since valuable space to describe a

technology had to compete with the validation of that technology.

In spite of these limitations, the results should prove of interest to the community. It provides a general overview of the forms of validation generally used by the computer science community to validate the various research results that are published and it does show that the field is maturing. Computer science seems to be developing an empirical culture so necessary to allow it to mature as a scientific discipline.

Appendix

The following is the taxonomy used to classify the 11 empirical validation methods.

1. *Project monitoring*. Collect the usual accounting data from a project and then study it.
2. *Case study*. Collect detailed project data to determine if the developed product is easier to produce than similar projects in the past.
3. *Field study*. Monitor several projects to collect data on impact of the technology (e.g., survey).
4. *Literature search*. Evaluate published studies that analyze the behavior of similar tools.
5. *Legacy data*. Evaluate data from a previously-completed project to see if technology was effective.
6. *Lessons learned*. Perform a qualitative analysis on a completed project to see if technology had an impact on the project.
7. *Static analysis*. Use a control flow analysis tool on the completed project or tool.
8. *Replicated experiment*. Develop multiple instances of a project in order to measure differences.

9. *Synthetic*. Replicate a simpler version of the technology in a laboratory to see its effect.
10. *Dynamic analysis*. Execute a program using actual data to compare performance with other solutions to the problem.
11. *Simulation*. Generate data randomly according to a theoretical distribution to determine effectiveness of the technology.

References

- Sjøberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanovic, A., Liborg, N.-K., Rekdal, A.C., 2005. A survey of controlled experiments in software engineering. *IEEE Trans. Soft. Eng.* 31 (9), 733–753.
- Tichy, W.F., Lukowicz, P., Prechelt, L., Heinz, E.A., 1995. Experimental evaluation in computer science: a quantitative study. *J. Syst. Software* 28 (1), 9–18.
- Zannier, C., Melnik, G., Maurer, F., 2006. On the success of empirical studies in the international conference on software engineering. In: *International Conference on Software Engineering*, Shanghai, China, pp. 341–350.
- Zelkowitz, M.V., Wallace, D., 1998. Experimental models for validating computer technology. *IEEE Comput.* 31 (5), 23–31.

Marvin V. Zelkowitz is a research professor of Computer Science at the University of Maryland in College Park Maryland. Prof. Zelkowitz received the MS and PhD degrees from Cornell University in Computer Science in 1969 and 1971, respectively and the BS in mathematics from Rensselaer Polytechnic Institute in 1967. He is a fellow of the IEEE, a Golden Core member of the IEEE Computer Society, and a member of ACM. He is the series editor of Elsevier's "Advances in Computers" book series, and on the editorial board of *Empirical Software Engineering*.