Seminar "Open Source Software Engineering"

Winterterm 2004

Empiricism in Computer Science

Eike Send
send@inf.fu-berlin.de
Advisor: Christopher Oezbek

August 29, 2005

Abstract

As the name computer science already implies, the study of computers is a field of science. Computer Science (CS) as it exists today lacks to a certain extent, what other sciences rely most on: An empirical body of knowledge. This paper looks at several meta studies which have analyzed the presence of empirical data on CS subjects. It also provides an overview of empiricism in general, some empirical concepts and where computer science and empiricism intersect.

Contents

1 Introduction

What I will do in the following chapters is provide some insight in the scientific method, empiricism and empirical work. Having done so, I will continue with the intersection of scientific method and computer science, namely by presenting studies and experiments.
I set off with an introduction to what the scientific method is, how it came to be and why empiricism is so directly connected with it. I proceed with a definition of a formal experiment and how it is to be conducted. Which kinds of experiments and studies there are and how they can be grouped is found before a brief summary of which scale types exist and how to not confuse them. The section on the scientific method is ended by a summary of Prechelt's introduction to data analysis.
The section on empiricism in computer science begins with a positioning of CS in science in general, followed by a brief overview of how empirical work is viewed and evaluated on a meta-level. A subsection which summarizes a meta study on empirical work in CS is included with a summary of a tangible study which has been reviewed in the meta study.

2 The Scientific Method

Let me begin this section by quoting Lutz Prechelt in one of his lectures:
"There are different ways how people obtain understanding

- by intuition (direct insight)
- from some authority (tradition, teacher, book etc.)
- by rational thought (reasoning, deduction)
- by direct observation
- via the scientific method" [Pre05]

He continues to state, that even though all of these methods can lead to a correct result, the scientific method will most likely do and it will also be the one most convincing to others. But what is the scientific method?
"Empiricism [...] is generally regarded as being at the heart of the modern scientific method, that our theories should be based on our observations of the world rather than on intuition or faith [...]" [Wik05c]

2.1 Falsification and Karl Popper

Karl Popper is the man who took empiricism to the level where it is today by introducing the philosophy of critical rationalism, which states that "no number of positive outcomes at the level of experimental testing can confirm a scientific theory, but a single genuine counter-instance is logically decisive: it shows the theory, from which the implication is derived, to be false. Popper's account of the logical asymmetry between verification and falsification lies at the heart of his philosophy of science."[Wik05e]
"Falsification is the contradiction of hypotheses or theories through empirical statements (for example by observation or experiment)" [Tic05b]
In the words of Prechelt this lead to the current state in which "a theory [that] cannot be refuted for a long time, [...] will gradually be accepted as confirmed."[Pre05]

2.2 The Experiment

Deligiannis et al. "consider an experiment to be a controlled empirical investigation into some phenomenon with a clearly stated hypothesis and random allocation of subjects to different treatments." [DSWR02]
They also provide "a formal terminology for describing the components of an experiment. Object of study is the entity that is studied in the experiment. They can be products, processes, resources, models, metrics or theories. Treatments are the different activities, methods or tools we wish to compare or evaluate. When we are comparing using a treatment with not using it, a control must be established, which provides a benchmark. A trial is an individual test run, where only one treatment is used. Experimental subjects are the people applying the treatment, for example using an OO programming language to solve a particular problem. The response or dependent variables are those factors that are expected to change or differ as a result of applying the treatment, for example, time taken or accuracy. By contrast, state or independent variables are those variables that may influence the application of a treatment and thus indirectly the result of the experiment. The number of, and relationships among, subjects, objects and variables must be carefully described in the experimental plan. Criteria for measuring and judging effects need to be defined, as well as methods for obtaining the measures. Finally, two important concepts are involved in the experimental design: experimental units which are the experimental objects to which a single treatment is applied, and experimental error which is the failure of two identically treated experimental units to yield identical results."[DSWR02]

Summarizing Fenton and Pfleger [FP97] Deligiannis et al. provide six steps for "carrying out a formal experiment:

1. **Conception:** deciding what we wish to learn more about, and define the goals of the experiment. From this, we must state clearly and precisely the objective of the study.
2. **Design:** to translate the objective into a formal hypothesis. The goal for the research needs to be re-expressed as a hypothesis that we want to test. The hypothesis is a tentative theory or supposition that we think explains the behavior we want to explore.[...]
3. **Preparation:** to make ready the subjects and the environment. If possible, a pilot study of the experiment should be conducted.
4. **Execution**
5. **Analysis:** this phase consists of two parts. First, all the measurements taken must be reviewed in order to ensure that they are valid and useful. Second, there follows the analysis of the sets of data according to usual statistical principles.
6. **Dissemination and Decision-making:** to document the experimental materials and conclusions in a way that will allow others to replicate and confirm the conclusions in a similar setting. [...]"

But there are different kinds of research methods, Tichy [Tic05a] identifies seven types:

1. **Case study:** "Rather than using large samples and following a rigid protocol to examine a limited number of variables, case study methods involve an in-depth, longitudinal examination of a single instance or event: a case." [Wik05b]
2. **Field experiment:** "A field experiment applies the scientific method to experimentally examine an intervention in the real world rather than in the laboratory." [Wik05d]
3. **Controlled experiment:** see section 2.2
4. **Survey:** "Statistical surveys are used to collect quantitative information [...] social science research. A survey may focus on opinions or factual information depending on its purpose, but all surveys involve administering questions to individuals." [Wik05j]
5. **Meta study:** A meta study "combines the results of several studies that address a set of related research hypotheses." [Wik05f]
6. **Simulation:** "A simulation is an imitation of some real device or state of affairs. Simulation attempts to represent certain features of the behavior of a physical or abstract system by the behavior of another system." [Wik05i]
7. **Benchmark:** "In computing, a benchmark is the result of running a computer program, or a set of programs, in order to assess the relative performance of an object, by running a number of standard tests and trials against it." [Wik05a]

2.3 Quantitative and Qualitative Research

Research can also be divided into two other groups: quantitative and qualitative research. "Quantitative research is the numerical representation and manipulation of observations for the purpose of describing and explaining the phenomena that those observations reflect." [Wik05h] "Qualitative research is a broad term that describes research that focuses on how individuals and groups view and understand the world and construct meaning out of their experiences; it is essentially narrative-oriented." [Wik05g]
Even though the types Tichy found tend to be more quantitative than qualitative or vice versa they "can combine both approaches, which is almost always a good idea." [Pre05] E.g. how much time is spent on a type of work (quantitative) and on what activities is it spent (qualitative). [Pre05]

2.3.1 Scale Types

Quantitative more than qualitative research uses scales to measure and note results. "In the early 1940s, the Harvard psychologist S.S. Stevens coined the terms nominal, ordinal, interval, and ratio to describe a hierarchy of measurement scales."[VW93]
Warren S. Sarle defines them like this [Sar96]:

Nominal Two things are assigned the same symbol if they have the same value of the attribute. Examples: numbers assigned to religions in alphabetical order, e.g. Atheist=1, Buddhist=2, Christian=3, etc.

Ordinal Things are assigned numbers such that the order of the numbers reflects an order relation defined on the attribute. Two things x and y with attribute values a(x) and a(y) are assigned numbers m(x) and m(y) such that if m(x) $>$ m(y), then a(x) $>$ a(y).
Example: school grades (1, 2,...,6); very good, good, OK, not so good, bad

Interval Things are assigned numbers such that differences between the numbers reflect differences of the attribute. If m(x) - m(y) $>$ m(u) - m(v), then a(x) - a(y) $>$ a(u) - a(v).
Examples: temperature in degrees Fahrenheit or Celsius; calendar date.

Ratio Things are assigned numbers such that differences and ratios between the numbers reflect differences and ratios of the attribute.
Examples: Length in centimeters; duration in seconds; temperature in degrees Kelvin.

2.4 Data Analysis

Prechelt says that "there are several different kinds of general goals when analyzing data" [Pre05] which can be summarized as follows:

Exploring something "You do not know in advance what to expect in the data. You try to get an overview of the data you have and to find interesting structure in the data." [Pre05]

Measuring something "You know exactly what aspect of an object you are interested in." [Pre05]

Modeling something for explanation "You want to describe the mechanism that has produced the data" [Pre05]

Modeling something for prediction "You consider your data to be examples. You want to find out how to predict output values given the input values." [Pre05]

Comparing something "You have two or more "things" and want to compare them with respect to one or more attributes." [Pre05]

"Data analysis has to support the primary quality attributes of the empirical study overall: credibility and relevance." [Pre05] Relevance can not be provided by the analysis itself, but in order to render the study's credibility higher it must be correct, illustrative and informative: "

Correct Data has not been mis-collected nor mis-processed and we trust the analysis (and hence its results).

Illustrative It is easy to understand what the results say and how they came to be, given the data. The analysis makes us understand the data itself.

Informative The analysis reports results that are relevant and helpful for answering the study question. " [Pre05]

Prechelt also provides 4 steps for carrying out the analysis:

1. **Make data available:** Make it machine readable, collect it, anonymize it, encode it, etc.
2. **Validate data:** Compare actual and expected data, remove impossible, corrupt or mislabeled data, etc.
3. **Explore data:** Get an overview, check single variables and pairs, quick check specific expectations.
4. **Perform analysis:** Measure, model, or compare

3 Empiricism in Computer Science

When I say in the abstract of this paper, that computer science has a shallow empirical body of knowledge, does it imply CS is not much of a science at all?
Prechelt mentions in a lecture that the modern sciences consist, to different degrees, of "theory, construction, empiricism" [1] [Pre05] "all three are essential for successful work, empiricism is slowly gaining ground." [Pre05]
He says computer science is a mixture of the purely theoretical mathematics and the mostly constructive but also theoretical and empirical electrical engineering.

3.1 The State of Empirical Research in Computer Science

What the quality of research in CS is concerned a meta study concludes that "the results suggest that large parts of CS may not meet standards long established in the natural and engineering sciences." [LTHP94] For example has it been noticed that the supremacy of object oriented technology (OOT) over procedural technology has been proven "more by intuitive feelings and anecdotal evidence, than by empirical and quantitative evidence." [Bur95]
This, put together with a situation described by Paul Schneck as "[...] many (most?) students of computer science are not educated as scientists. They are trained as programmers." [Mud96]
This leads Braught, Miller and Reed to demand to introduce "important empirical concepts and skills"[BMR04] to students from early on.
In conclusion a panel of computer scientists experienced in experimentation and system building developed a consensus "that favored fostering a culture closer to the traditional sciences as far as experimentation is concerned." [Mud96]
"This weakness should be rectifed for the long-term health of the field." [LTHP94]
But also the amount of research which reaches established standards is low compared to other sciences: "A survey of over 400 recent research articles suggests that computer scientists publish relatively few papers with experimentally validated results." [LTHP94] Another more recent study suggests the same result:

[1] He defines these parts of the scientific method as:
Theory: produces formalisms, derives results about them, revolves around logical issues
Construction: produces systems designs, constructs systems, revolves around practical issues
Empiricism: produces observations of systems and interprets them, revolves around behavior in and of the real world

Table 2. Classification of 612 evaluated papers.

	1985			1990			1995			
Method	ICSE	*IEEE Software*	TSE	ICSE	*IEEE Software*	TSE	ICSE	*IEEE Software*	TSE	Total
Not applicable	6	6	3	4	16	2	5	7	1	50
No experimentation	16	11	56	8	8	41	10	3	14	167
Replicated	1	0	0	0	0	1	1	0	3	6
Synthetic	3	1	1	0	1	4	0	0	2	12
Dynamic analysis	0	0	0	0	0	3	0	0	4	7
Simulation	2	0	10	0	0	11	1	1	6	31
Project monitoring	0	0	0	0	1	0	0	0	0	1
Case study	5	2	12	7	6	6	4	6	10	58
Assertion	12	13	54	12	19	42	4	14	22	192
Field study	1	0	1	0	0	1	1	1	2	7
Literature search	1	1	3	1	5	1	0	3	2	17
Legacy data	1	1	2	2	0	2	1	1	1	11
Lessons learned	7	5	4	1	4	8	5	7	8	49
Static analysis	1	0	1	0	0	0	0	0	2	4
Yearly totals	56	40	147	35	60	122	32	43	77	612

As you can see a total of 359 studies of 612 total have either no experimentation at all or are mere assertions. The authors of this study conclude that "the state of experimentation in software engineering is poor. [...] On the whole, we consider this situation as unacceptable, even alarming." [ZW98]
In order to show some representative empirical work in computer science I chose an example on the question of object oriented technology (OOT) vs procedural technology (PT).

3.2 Programming: Object Oriented vs Procedural

To illustrate the point made in the section above I will give a brief summary of the meta study by Deligiannis et al. [DSWR02] which has been mentioned earlier on. The main subject of the study are formal experiments on object oriented technology vs procedural technology because "the results of an experiment can be more easily generalised than those of a case study." [DSWR02]
Of the 27 Experiments they had identified only 18 met their requirements of a formal experiment. (see 2.2 "The Experiment") 10 of these compared OOT with PT. All of the experiments were evaluated using a framework by Wohlin et al. [WRH+00] which divides the analysis into the following parts:

- Motivation
- Hypotheses
- Variables
- Participants
- Experiment design
- Results and interpretation
- Critique

The main critique was that most papers merely used students as participants, which supposedly distorts the picture, because in OOT expertise is important for decision making.[2] Other points

[2]Tichy notes on this subject "Studies have found that mere length of professional experience has little to do with competence. In other words, you can't use the argument that professionals with years of experience will necessarily solve a given problem better than appropriately prepared (graduate) students." [Tic00]

that have been criticized were that in some experiments the design was flawed or too simple to be realistic or that the experiment setup lacked a documentation of the domain.
For what the results are concerned the authors found that in only two of the ten OOT vs PT experiments a significant benefit of OOT was found, which could be due to the choice of subjects. For the sake of illustration I will summarize one of the two studies.

3.2.1 A Sample Controlled Experiment

"This paper describes an experiment which compares the maintainability of two functionally equivalent systems, in order to explore the claim that systems developed with object-oriented languages are more easily maintained than those programmed with procedural languages." [HH90] The summary of the evaluation of this experiment by Deligiannis et al. follows.

Motivation Lack of scientific evidence for OOT superiority

Hypotheses Alternative H_0: It is easier to maintain object oriented (OO) programs than structured programs; H_1: OO programmers take less time to perform a maintenance task, H_2: OO maintenance requires fewer changes to the code, H_3: OO programmers perceive the changes as conceptually easier, and H_4: OO programmers make fewer errors during the maintenance task.

Variables "There were four independent variables: subject (students), group (A or B), programming language, and the modification task. The dependent variables were maintenance times, error counts, change counts, and programmers subjective impression." [DSWR02]

Participants "Twenty students participated." [DSWR02]

Experiment design Using a counterbalancing procedure the subjects were to perform enhancement maintenance tasks. The subjects performed each task twice, once in C and once in Objective C and completed a post-experimental questionnaire.

Results and interpretation "This experiment supports the hypothesis that subjects produce more maintainable code with an OO language than with a PT language." [DSWR02]

Critique The counting of maintenance time relied solely "upon the accuracy of subjects reporting minutes of thinking time" and it is also noted "that inexperienced subjects were used and design documentation was not provided." [DSWR02]

4 Conclusion

What the philosophical, theoretical and conceptional, but also the very tangible sides of empirical work in general is concerned it is evident, that a massive body of knowledge has been accumulated literally over centuries.
The empirical body of knowledge in computer science on the other hand is shallow and the little there is needs confirmation from other experimenters. A lot of work still needs to be done, if CS wants to hold up to scientific standards. But it should not remain unmentioned that there also are scientists who are well aware of this situation and who are doing their share to keep the process of empirical work going.

References

[BMR04] Grant Braught, Craig S. Miller, and David Reed. Core empirical concepts and skills for computer science. In *SIGCSE '04: Proceedings of the 35th SIGCSE technical symposium on Computer science education*, pages 245–249, New York, NY, USA, 2004. ACM Press.

[Bur95] Angela Burgess. Finding an experimental basis for software engineering: Interview with victor basili, software engineering lab. *j-IEEE-SOFTWARE*, 12(3):92–93, may 1995.

[DSWR02] Ignatios S. Deligiannis, Martin Shepperd, Steve Webster, and Manos Roumeliotis. A review of experimental investigations into object-oriented technology. *Empirical Softw. Engg.*, 7(3):193–231, 2002.

[FP97] Norman Fenton and Shari Lawrence Pfleeger. *Software metrics (2nd ed.): a rigorous and practical approach*. PWS Publishing Co., Boston, MA, USA, 1997.

[HH90] Sallie M. Henry and Matthew C. Humphrey. A controlled experiment to evaluate maintainability ofobject-oriented software. Technical report, Blacksburg, VA, USA, 1990.

[LTHP94] Paul Lukowicz, Walter F. Tichy, Ernst A. Heinz, and Lutz Prechelt. Experimental evaluation in computer science: a quantitative study. Technical Report iratr-1994-17, 1994.

[Mud96] Trevor Mudge. Report on the panel: how can computer architecture researchers avoid becoming the society for irreproducible results? *SIGARCH Comput. Archit. News*, 24(1):1–5, 1996.

[Pre05] Prof. Dr. Lutz Prechelt. Empirische bewertung in der informatik: Einführung. *http://projects.mi.fu-berlin.de/w/bin/view/SE/VorlesungEmpirie2005*, 2005.

[Sar96] Warren S. Sarle. Measurement theory: Frequently asked questions. *http://originresearch.com/documents/measurement1.cfm*, 1996.

[Tic00] Walter F. Tichy. Hints for reviewing empirical work in software engineering. *Empirical Softw. Engg.*, 5(4):309–312, 2000.

[Tic05a] Walter F. Tichy. Ausbildung in empirischer softwaretechnik. *http://www.ipd.uka.de/exp/otherwork/AusbildungEmpirischeSoftwaretechnik.pdf*, 2005.

[Tic05b] Walter F. Tichy. Die rolle der empirie in der softwaretechnik. *http://www.ipd.uka.de/exp/otherwork/RolleDerEmpirie.pdf*, 2005.

[VW93] Paul Velleman and Leland Wilkinson. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 1993.

[Wik05a] Wikipedia. Benchmark, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/Benchmark*, 2005.

[Wik05b] Wikipedia. Case study, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/Casestudy*, 2005.

[Wik05c] Wikipedia. Empiricism, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/Empiricism*, 2005.

[Wik05d] Wikipedia. Field experiment, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/Fieldexperiment*, 2005.

[Wik05e] Wikipedia. Karl popper, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/KarlPopper*, 2005.

[Wik05f] Wikipedia. Meta-analysis, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/Metaanalysis*, 2005.

[Wik05g] Wikipedia. Qualitative research, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/QualitativeResearch*, 2005.

[Wik05h] Wikipedia. Quantitative research, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/QuantitativeResearch*, 2005.

[Wik05i] Wikipedia. Simulation, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/Simulation*, 2005.

[Wik05j] Wikipedia. Statistical survey, from wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/Statisticalsurvey*, 2005.

[WRH+00] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in software engineering: an introduction*. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[ZW98] Marvin V. Zelkowitz and Dolores R. Wallace. Experimental models for validating technology. *IEEE Computer*, 31(5):23–31, 1998.