1st International Workshop on Conducting Empirical Studies in Industry (CESI 2013) – Post-workshop Report

Xavier Franch
Univ. Politènica de Catalunya
Barcelona, Spain
franch@essi.upc.edu

Nazim H. Madhavji
Univ. of Western Ontario
London, Canada
madhavji@gmail.com

Bill Curtis
CAST Software
Fort Worth, Texas, USA
b.curtis@castsoftware.com

Larry Votta
Brincos
Sammamish, WA, USA
larry@brincos.com

ABSTRACT
The quality of empirical studies is critical for the success of the Software Engineering (SE) discipline. More and more SE researchers are conducting empirical studies involving the software industry. While there are established empirical procedures, relatively little is known about the dynamics of conducting empirical studies in the complex industrial environments. For example, what are the impediments and how to best handle them. This was the primary driver for organising CESI 2013, held on 20th May, 2013. Thus, the theme of the workshop was "*conducting empirical studies in industry*". This report summarises the workshop details and the proceedings of the day.

General Terms
Measurement, Experimentation.

Keywords
Empirical studies, software industry.

1. INTRODUCTION
An "empirical study" is an investigation, using established procedures (also called "empirical methods"), for the purpose of gaining knowledge through observation. Empirical methods fall under the broad categories of case studies, scientific experiments and surveys. Investigative questions of interest are posed and related data is gathered and analysed to answer these questions. Briefly, with experiments, we are in search of quantitative, cause-and-effect relationships, and involve control of treatment. With case studies, we are in search of qualitative or quantitative relationships among the identified variables in the case under study in the real-world setting (and hence do not involve any control). With surveys, we are in search of qualitative or quantitative responses from a sample representative of the population under study. There are various "research designs" to cater for different investigative situations. Examples include: independent measures, repeated measures, matched pairs, etc.; exploratory case studies, longitudinal case studies, ethnographic studies, action research, etc.; and online surveys, focus groups, interviews, etc. With empirical studies being widely entrenched in fields such as social sciences, psychology, management sciences, and medicine, there is obviously much more in the general literature on empirical studies than that what is hinted above; still, this brief introduction will suffice for our purpose here. The following sample publications can serve as a starter [1][2][3].

In so far as software engineering (SE) is concerned, empirical studies lie at the heart of this burgeoning field. The quality of these studies is a determinant of the validity of the research findings and proposed solutions (i.e., methods, techniques, tools, etc.) and of the success of the evolution of the SE discipline. With increased awareness, more and more researchers are conducting empirical research in SE and increasingly so involving the software industry.

While there are established empirical methods in the general literature, relatively little is known about conducting empirical studies involving the software industry. What pitfalls to avoid when investigating phenomena in an organisation; what challenges to anticipate when evaluating the efficacy of methods and tools in actual projects; what are the dos and don'ts when conducting practitioner surveys; etc. Such questions abound and they formed the primary trigger for organising this workshop. The theme of the workshop was thus "*conducting empirical studies in industry*".

Experience suggests that empirical studies conducted in industrial settings are particularly challenging because the actual environments are complex and what is first observable by researchers (typically from academia) may only be a tip of the iceberg. Yet, relevant investigative questions must be formulated, valid constructs need to be defined, trust needs to be in place, relevant and quality data must be gathered within small time-frames available, industry-relevant results need to be delivered in real-time, etc. In essence, researchers often need to be able to run while they are still learning how to walk.

2. WORKSHOP GOALS AND PROCEEDINGS
The goals of the workshop were:

- to deliberate on challenges and experiences in conducting empirical studies in industrial settings;
- to discuss strategies for overcoming impediments;
- to debate on the limitations of contemporary research methods; and
- to project towards their resolutions.

Several mechanisms were used to realise these goals: a keynote speaker and its respondent, paper presentations, and the not-so-common "wall of ideas" session. These are described below.

3. KEYNOTE SPEAKER AND RESPONDENT
The keynote speaker for this workshop was Barry Boehm (University of Southern California), and the Respondent to the keynote was Dewayne Perry (University of Texas at Austin). Boehm summarised decades of conducting empirical studies in industry in the areas of: (i) *methods*: inspection, testing and pair programming; (ii) *emerging technologies*: agile, model-driven and value-based; and (iii) *parametric modelling*: cost, schedule and estimation quality.

For *methods*, the benefits centred around cost-effectiveness and insights; whereas, challenges centred around obtaining representative projects, and gained access to personnel and the environment. For *emerging technologies*, the benefits centred around maturity of the projects, cost-effectiveness and insights; whereas, challenges centred around baselining projects, learning curve due to changes in the processes, and appropriate subject skills. For *parametric modelling*, the benefits centred around budget realism, progress monitoring, productivity and quality improvement areas; whereas, challenges centred around community representativeness for generalisability of the results, proprietary data, and data consistency.

Proportionally, Boehm spent more time on COCOMO family of cost models. Of particular interest here are the overall success criteria:

- Evidence of model demand.

- User willingness to support model definition, provide calibration data.
- Model focused on supporting major decision situations.
- Good match of estimation relationships to underlying phenomenology.
- Clear definition of inputs, outputs, and assumptions.
- Careful conditioning of calibration data.
- Flexibility in adapting the model to explain mismatches and outliers.
- Good balance of model success criteria.

Complementing the overall success criteria are technical success criteria for a model. These are:

- *Scope*: Covers desired range of situations?
- *Granularity*: Level of detail sufficient for needs?
- *Accuracy*: Estimates close to actuals?
- *Objectivity*: Inputs repeatable across estimators?
- *Calibratability*: Sufficient calibration data available?
- *Contructiveness*: Helps to understand job to be done?
- *Ease of use*: Parameters easy to understand, specify?
- *Prospectiveness*: Parameters values knowable early?
- *Parsimony*: Avoids unnecessary parameters, features?
- *Stability*: Small input changes mean small output changes?
- *Interoperability*: Easy to compare with related models?

When we stack up the challenges, described earlier (i.e., community representativeness for generalisability of the results, proprietary data, and data consistency), against the overall and technical success criteria of models, it becomes readily clear how difficult empirical studies become in industrial settings. In particular, these difficulties can translate into: (i) study failures, (ii) incomplete studies, (iii) study costs escalating beyond the budgeted amount, (iv) study's cycle-time stretched into delays in the findings, (v) numerous threats to the validity and quality of the results, and more.

Boehm also identified similarities and differences in the critical success factors for developing parametric estimation models and other forms of empirical studies in industry, such as surveys, case studies, and controlled experiments. Similarities identified include:

- Need for careful definitions and data assessment.
- Half-life of results due to rapid changes in the software field.
- Need for access to scarce experts.
- Getting a critical mass of contributors.
- Need for accuracy of contributed data.
- Need to overcome industry reluctance to contribute competition-sensitive data.

Likewise, differences identified with some of the other forms of empirical studies in industry include:

- Stronger ability of surveys and case studies to cover multiple issues.
- More detailed coverage of individual practices.
- Uncertainties in comparability across different organizations.
- Uncertainties in generality across application domains.
- Scalability and comparability challenges for controlled experiments.

Needless to say, Perry was faced with a difficult task to respond to Boehm's presentation! Confrontation was diplomatically avoided (rather easily as he was in agreement with what Boehm presented) in favour of digging up his own experience from empirical studies conducted in the 5ESS switching system environment at AT&T and Lucent. Given that Boehm's "parametric model" was a software system interpretation (or analogy) of basic empirical study structure, Perry drilled down further on the issues of validity in empirical software engineering. Basically, this is providing answers to the following questions:

- Are we measuring/evaluating what we mean to measure/evaluate?
- Are the results due solely to our manipulations?
- Are our conclusions justified?
- What are the results applicable to?

The answers to these questions are sequentially dependent: the first question is about construct validity which, if you do not get them right, the rest does not matter (like building the wrong system); the second is about internal validity and the problem of confounding variables and alternative explanations which, again, if you do not get right, the rest doesn't matter; the third is about the analysis logic and statistical validity to justify ones conclusions; and the last is about external validity, or, generally, what are the results applicable to. Each of these issues was expanded on, providing more details for each validity concern.

While discussing these validity issues, Perry mixed in examples from his empirical studies at Bell Labs and UT Austin. One such example was the time studies that Perry did with Votta and Staudenmeyer in which they explored how developers spent their time in the evolution of an extremely large real-time system. Their first study was a longitudinal study based on the personal diary and project notebooks for a 32 month development where the time granularity was one day. There were two interesting results: first, that the developers were only 40% effective with 60% of their time blocked; second, there were significantly different phenomena depending on different phases of development. To determine whether this study was the result of the specific developer or a more general phenomenon, a cross-sectional study was done with a variety of developers who together covered most of the complete development process. There were two primary differences: most of the developers had two developments they were working on; second, a daily self- reporting diary structure with a half-hour granularity was used to indicate how much time was spent where in their defined process. The results were congruent with the first study: the developers were still 40% efficient, but the issue of being block was ambiguated by context switching. The primary question then was how accurate were the self-reported diaries. This led to the third study with the cross-sectional study participants in which Staudenmeyer spent several days with each developer from a subset of the participants to find out what was done at a minute by minute basis. The developers were self-consistent but not consistent with each other and on average where 80% accurate in their reporting. The fascinating results were what could be seen at this level of granularity that were not at all obvious from the previous two studies: an average of 75 minutes a day were spent in short (on the order of 3 minutes) unplanned, informal interactions, and face to face interaction was by far preferred over email and telephone interactions.

Perry finished his response with a cartoon from the Wizard of Id, where the court magician was working on the perfect placebo: "one that has all the side effects of the real pill". The analog of this would make a world of difference in the validity of our empirical work in software engineering.

4. THE REVIEW PROCESS AND PAPERS PRESENTED

Each submitted paper was reviewed by at least three reviewers. The outcome of this process yielded seven regular papers (6 pages), six short papers (4 pages), and two practitioner messages (2 pages). We may arrange these papers along different dimensions.

1. *Demographic data*.
   - *Region*. Nine of the works were authored or co-authored by academics and practitioners from the Northern and Middle Europe (9 papers), with predominance of Sweden (5). Three papers had Asian contribution with identical profile: collaboration academy-industry where the Asian were the practitioners, from India (Tata, Infosys). Israel was the other country with authors of more than one paper, but there were from the same group (even if not complete author overlapping). It was a bit surprising having just one paper with authors from USA, especially considering that the workshop was held in San Francisco.
   - *Industry or Academia*. There is a clear dominance of academy papers: 10 papers were written by authors that were all of them academics, and no single paper had only practitioner authors, although there is one especial case with author with double affiliation industry-academy. Three of the other papers had a practitioner as first author, whilst the fourth had an academic as first author.
2. *Type of study presented.* Almost half of the papers (7) were experience reports aimed at extracting some lessons learned, challenges, open issues, etc., from a series of primary studies. The rest were very disperse, including two surveys, two technical papers presenting some theory out of empirical studies, two position or vision papers (corresponding to the two short practitioner messages) and one evaluation paper.
3. *Domain of study*. Although in most cases the domain may have not influenced the observation, most of the works referred to some software domain. Agile projects was the most referred one (3 papers) followed by testing (2). We had then 5 papers in the context of 5 different domains, and one paper that presented 6 primary studies from different domains. The last 3 papers didn't specify domain or the domain was clearly a secondary issue (e.g., a systematic literature review performed).
4. *Type of studies analysed*. Case studies were the favourite type of studies subject of analysis (5 papers). Experiments and web based questionnaire surveys were also addressed by more of one paper (3 and 2, respectively), and finally we had 1 paper on interviews. The remaining 3 papers didn't mention the type of studies or the contribution seemed to be applicable to all of them.
5. *Own studies or studies from community*. The great majority of papers (13) reported on the work of the authors themselves, whilst just 2 papers involved primary studies from other authors: the systematic literature review and a paper on replication of experiments.
6. *Number of primary studies*. As could be expected, the systematic literature review **¡Error! No se encuentra el origen de la referencia.** was the paper involving the greatest number of primary studies (16). Then we had 6 papers with more than one primary study (among 3 and 6) and 5 papers involving only one primary study, although it must be mentioned that one of these 5 papers is a case study conducted in 7 different firms. Three of the papers didn't mention the number.

5. THE WALL OF IDEAS

In this session, all participants were invited to post their thoughts and ideas, asynchronously, on "*The Wall of Ideas*" – a structured wall (essentially a matrix) to capture individual contributions. The wall had the following columns and rows:

- Rows:

1. Feedback to and from the stakeholders
2. Stakeholder commitment
3. Challenges / Barriers
4. Alignment with business goals
5. Industry setting
6. Tips, lessons learnt & solutions
7. Principles & Fundamentals
8. Tool support
9. Organization dynamics

- Columns:

1. Recognising the need for a study
2. Definition of a research problem
3. Design of study
4. Data access & gathering
5. Data cleanup
6. Threat identification
7. Validation of results
8. Interpretation of results
9. Transfer of results
10. Generalisation
11. Theoretical background

As can be easily visualised, a 9 by 11 matrix is quite large and can potentially trigger many diverse ideas and thoughts to be captured and, indeed, within the limited time we had in this session, a substantial number of points were posted on the wall. Not all points posted can be included here due to space limitation, so we give below one illustrative example row of ideas only. Note that these points are in the raw form; there wasn't adequate amount of time to reach any conclusions on specific issues. Table 1 shows this by presenting the row "Definition of a research problem" (applied to various column headers) as responded by participants.

Table 1. Excerpt of the wall of ideas

Column headers	Posted notes
Feedback to/from the stakeholders	Do workplace analysis
Figures	Experts believe in myths, have vested interests, are subject to political pressure, practice "denial", etc. Be careful to find the truth
Challenges / Barriers	• Obfuscation. Blind leading blind • RQ must be "relevant" to company or industry in general • "As you ask in the forest you shall get answered" - Swedish proverb
Alignment with business goals	• Will question be relevant in 2 years? 5? • Keep asking company what their business needs are until you hear something that triggers an idea for a study about a solution

	• Do domain modeling before concluding on a research question
Industry setting	• You may have to slightly adapt your original RQs to the real industrial setting • Company is interested in solving practical problems • Practitioners want stories. Case studies in realistic settings are valuable even if not statistically valid

6. ACKNOWLEDGMENTS

We greatly acknowledge: the contributions from the authors of all the papers submitted to CESI; the valuable time and effort spent by the Program Committee members to review the papers; the presentations made by the keynote speaker (Barry Boehm), the respondent (Dewayne Perry), and the presenters of the accepted papers; and the workshop participants at large without whom the session on "Wall of Ideas" would not have materialised. Sincerely thank you all. The participation of Dr. Franch was supported by the Spanish project TIN2010-19130-C02-01.

7. REFERENCES

[1] Yin, R.K. 2009. *Case Study Research: Design and Methods (4th ed.)*. SAGE Publications.

[2] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A. 2012. *Experimentation in Software Engineering*. Springer Verlag.

[3] Fowler, F.J. 2009. *Survey Research Methods (4th ed.)*. SAGE Publications, 2009.