Session 4: Experimentation in the Process Framework
The first step to building a credible empirical side to process research

Carolyn Seaman
Department of Computer Science
University of Maryland
College Park, Maryland 20742

Lawrence G. Votta
Software and Systems Research Center
AT&T Bell Laboratories
Naperville, Illinois 60566

Abstract
What does experimentation in the process framework mean? After all, processes enacted as parts or all of a software development are one data point—right?

As you can observe from these two introductory sentences, our models and our language for discussing empirical work in software engineering, and in particular, a process framework are not well developed. This session was a start at trying to discuss and provide some of the foundations for models and language of experimental work when discussing processes involving more than one or two people.

The workshop session consisted of two parts. The first part was a general background section on areas relevant to experiments involving many people. We described some of the properties of good experiments and a research bibliography was made available to start the work of building a hybrid empirical research paradigm somewhere between the physical and social sciences.

The second part was a discussion of some of the key issues of concern to the participants at the workshop. These issues included the role of academia in work involving large non-repeatable developments and the role of tailorability desired/required by the project manager in large software developments.

Larry began by presenting his goals for the session, his assumptions, some terminology, some of his own personal "pearls", some good examples of experimental work, and his criteria for a good experiment.

The goals for the session were to:

- chart the path by identifying the big questions for empirical study;
- establish common terminology;
- discuss the credibility of empirical work; and
- share our pearls, lessons, and experience with experimentation.

Larry presented several points he considered assumptions on which to base the discussion. First, single person studies are well established. Second, and ironically, team and organizational studies have been poorly framed, even though this is where software process research should have the most impact. Finally, it will be necessary to incorporate multiple disciplines in order to make any headway.

In an effort to define a framework of terminology, Larry presented some groups of related terms. Empirical work can be divided into case studies, engineering experiments, and scientific experiments. People can be studied as individuals, in small groups (2-20 people, one level of management), or in organizations (more than 20 people, more than one level of management). Empirical study runs the spectrum from anecdotal to case studies to experiments.

Correlations can be simple relationships or causal relationships; but correlation and causality are not the same! Correlation is one of four properties that must be present to say that event A causes event B. The other three are: event A must occur before event B (temporal precedence), event A and B must not be spuriously correlated, and there must be a mechanism that explains how event A caused event B.

Experimentalists must also be concerned with both internal and external validity. Deductive testing of theory requires falsifiability, logical consistency, prediction. Case studies have difficulties with making controlled observations and deductions and allowing for repeatability and generalizability. An experimental design describes how the experimenter will manipulate independent variables, state hypotheses, and then observe the behavior of dependent variables.

Larry then presented his own personal pearls:

1. People must always be allowed to withdraw from a study without penalty.
2. People can only discriminate into 7 categories reliably.
3. People's behavior changes when they think they are being measured.
4. Experiments can always be improved.
5. Measure what you think you are measuring.
6. Clearly state hypotheses.

Larry then cited what he considers a few great experimental studies: Weinberg and Schulman experiments of the effects of goals on programmer performance [18], Schneiderman on control flow and data
Basili and Weiss on requirements changes experiments. First of all, sequences of experiments are especially useful because they show a successive refinement of the problem and the analysis. Credibility is, of course, a key issue. And it is crucial that the experiment answer an important question to the field.

At this point in the session, the participants were invited to propose issues for discussion. The proposed issues are listed below.

1. How do we empirically show that process can reduce the development interval?
2. How can academia help answer some questions that will help?
3. What kind of process tailorability do project managers need?
4. What do we do after we have the empirical stuff?
5. What are inappropriate questions for experimentation?
6. How do we validate data that people supply on the process?
7. How do we account for scale in empirical process work?
8. How do we get the software engineering community to realize the value of repeating experiments?
9. Is measuring for control different than measuring as part of an experiment?
10. Is the SEI data of 8:1 ROI (Return On Investment) a valid number?
11. How important are placebos for our experiments? Is it ethical to use them?
12. What are inappropriate questions for experimentation?

The group voted on which of the above issues to discuss for the second part of the session. The three chosen are presented below.

**Issue #1: How do we empirically show that process can reduce the development interval?**

One problem with efforts to study this issue is that often improvements in quality have been shown, but not in productivity or interval because developers tend to make the work fit the time available. Nazim Madhavji related similar experience with their experiments with PSP (Personal Software Process). Another important problem is identifying process as the causal agent of the improvement. Also raised, by Walcelio Melo and Prasanta Bose, was the problem of time; studying interval and process improvement must be done over long periods of time, during which no improvement may be seen until the end. Lolo Penedo's suggestion for helping to shorten development intervals was to look at COTS (Custom Off the Shelf Software) products, despite their many problems. Another suggestion, from Bob Balzer, was that we look to the BPR community to see how they justify process changes. However, according to Lolo, most BPR (Business Process Re-engineering) efforts use simple function point analysis to show the value of proposed changes.

Watts related his experience with the F16 study, where it was found that the bottleneck in cycle time was documentation, and so substantial improvement was made. However, the problems were uncovered by modeling, not measurement. Alex Wolf pointed out that the F16 process was a good process to study because it was significantly flawed, and so dramatic improvement could be shown through process improvement. The question is how to justify the study and incremental improvement of processes that are not obviously broken. This kind of tinkering, he observed, sometimes leads to processes that are too tightly wound, that is, all paths through the process are close in length to the critical path, so the process is highly vulnerable to slight variations, e.g., JIT (Just In Time manufacturing techniques). The problem is to define and recognize when a process is wound too tightly.

**Issue #2: How can academia help answer some questions that will help? What are some "test tube" experiment ideas that would be relevant and feasible in an academic environment?**

Some suggestions for test tube experiments were: studying the usefulness of graph models vs. other formalisms; cost vs. usability (a current experiment at University of Maryland); Process Weaver usability (at Grenoble); process tailorability.

Much of the discussion on this issue centered around the importance of industry/academic cooperation in experimentation. Carolyn Seaman raised the point that industry collaboration is becoming more important, in terms of funding, and that this will make research experimentation results more useful in the future. Kathleen Culver-Lozo and Lolo Penedo also stressed the need for interchange and partnership between academia and industry. Alex Wolf pointed out that the best ideas come from practitioners, but they have trouble generalizing, which is the strength of academicians. Larry Votta pointed out that the NASA SEL (Software Engineering Laboratory) is a classic and highly successful example of industry/academic cooperation in experimentation. However, several problems with industry collaboration were raised. Kathleen pointed out that academic collaboration is very costly for practitioners in terms of time, risk, unclear benefits, and dealing with the "attitudes" of academic researchers, many of whom have little respect for real problems. Alex pointed out that there is sometimes an attitude problem on the side of practitioners, who feel
that their problems are not only unique, but uniquely horrible. Both Selma Arbauoi and Carolyn Seaman felt that purely academic experiments were highly suspect, in part because the motivations of the subjects (students) are very different from those of practitioners. In general, academic experiments are not sufficiently convincing. But both Larry and Dewayne Perry pointed out the value of prototyping and refining an experimental design with a set of student subjects.

Issue #3: What kind of process tailorability do project managers need?

Dewayne Perry made a proposal of a way to experimentally study this question by merging two different development groups using different methods and tools, but the same set of generic processes, and then studying how methods and tools were chosen and integrated into the joint environment. Another approach to studying the problem, raised by Prasanta Bose, is the Xerox PARC approach, which is an ethnographic observational approach. Larry Votta said that there are other studies out there using this approach, but there are some real cost/benefit tradeoffs associated with it.

Watts Humphrey pointed out that we usually think of "tailoring" as happening infrequently, but it actually happens all the time. It usually takes place at project planning, when a process is built out of parts, or "building blocks", but we rarely think about it that way. We should start thinking in terms of an "organization's process assets", not an "organization's process".

Nazim Madhavji shared results from a study he had done which identified a number of factors influencing how processes were tailored: culture, geography, size of system, safety issues, customer population, and politics. Nazim also mentioned his previous work on congruency (fitness of a process to a context), in which he interviewed 70 people in 30-40 organizations, built a causal model, and validated it. This model predicts the effect of different methods on "success", measured in various ways.

The issue of tailoring a process "on the fly", or during process execution, also came up. Watts pointed out that it's risky to do this, but sometimes it is necessary in order to make "repairs".

Recommended Readings

Listed below are a set of recommended readings to create a base for discussing empirical work in organizational settings. A heading and brief description are provided to help direct the reader. The bibliography contains the actual reference.

• Case Study Methods -- Allen Lee's paper describes the logic and rationale of case studies and discusses ways to improve their credibility [11].

• Data Analysis and Statistics -- There are two excellent references in this area. The first is George Box et al. book on statistics for experimenters [4]. The second is Sidney Siegel and John Castellan's book specializing to the behavioral sciences [17].

• Large Engineering Organizations -- There is a lot of work on large engineering organizations. Thomas Allen has written an excellent book on his studies of paired developments [1]. Jeffery Liker and Walton Hancock describe their case study of an automobile manufacturer's engineering division and do an excellent job of describing the kinds of demoralizing forces that can occur in large organizations [12].

• Social Experimental Design -- Although Samuel Conte et al. book is about software engineering and not about social experimental design, he discusses many basic social experimental designs for use in software engineering in chapter 3 [5]. The work is unique since it also has a good discussion of the analysis techniques to use for the different designs. Charles Judd et al. is an excellent general reference for social experimental design [8]. The authors cover many areas and provide many references for further reading on many topics.

• Understanding People Effects in Experiments -- There are many papers in this area. However, for the person trying to just get started there are three that give the reader different perspectives of people and experiments. George Miller's paper really discusses how well people can measure from memory and under time pressure [14]. H. Parson discusses in his paper probably one of the most misunderstood experiments of the twentieth century -- the Hawthorne experiments [15]. Finally, the paper by A. Dennis and J. Valacich discusses the task of brainstorming and what the social effects of groups are on the number and kind of ideas generated [6].

References


