The Software Engineering Research Method Project at Simula

The overall purpose is to support the private and public services and industries in developing better IT-systems using fewer resources by:

1. advancing the state of the art of empirical software engineering research by developing new, or improving existing, methods and infrastructures for conducting empirical studies in software engineering, and

2. proposing and validating theories on the basis of the studies primarily carried out in software development organizations.
Purpose of doing research on research

• Help a research community to define the goals of its research
• Help a research community to define quality standards for the way the research is carried out
• Help a research community to achieve these quality standards

Purpose of SE research: Improve the way software systems are built and maintained
Software Engineering Research

SE research is about

(1) the development of new, or modification of existing, technologies to support SE activities

(2) the evaluation and comparison of the effect of using such technology in the often very complex interaction of individuals, teams, projects and organisations, and various types of tasks and software systems. This may be referred to as evidence-based or empirical software engineering, whose goal is to:

- to enable the development of scientific knowledge or evidence about how useful different SE technologies are in which situations. Such knowledge or evidence should guide the development of new SE technologies and be a major input to important SE decisions in industry.

Scientific Evaluation of SE Technology

- Today: Mostly based on anecdotal evidence, personal opinion, arbitrary tests, etc.

- Sciences that study real-world phenomena use empirical methods by necessity, which involve systematic observation and experimenting, rather than deductive logic or mathematics

- If SE research is to be a science, it must include the use of empirical methods
Examples

- Pair programming, and its interaction with individual skills and system complexity
- Effect of UML in the context of software evolution
- Regression testing: test selection, minimization, prioritization

Challenges

- More empirical studies
- Higher quality studies
  - More relevant studies
  - More valid studies
    - Internal validity*
    - Construct validity*
    - External validity*
    - Statistical conclusion validity*
    - Reliability (experiments, case studies, etc.)
- More focus on synthesizing evidence
- Theory building

# Yin, 2003
Current SE research literature

- Percentage of articles that report empirical studies:
  - Tichy: 17%
  - Glass et al.: 14%
  - Sjøberg et al.: 12-17%

- Primary studies
  - Controlled experiments 1.9% (Sjøberg et al.)
  - (Personal opinion) Surveys 1.6% (Glass et al.)
  - Case studies 12% (Holt)
  - Action research 0% (Glass)

- Reviews and meta-analysis: 1-3% of papers

- Rough estimate: 180 studies a year

Need: ~2000 studies

- Assume there are 1000 research questions of high industrial importance that are meaningful to decide empirically, and
- assume that a research question requires at least 20 high quality studies, conducted over the last 10 years.
- This requires that we conduct at least 2000 high-quality empirical studies every year.

More empirical studies

<table>
<thead>
<tr>
<th>State of Practice</th>
<th>Target (2020-2025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively few empirical studies in SE research. Focus on developing new technology</td>
<td>Large number of studies covering all important fields of SE and using different empirical methods. Most research that leads to new or modified technology is subject to empirical evaluation</td>
</tr>
<tr>
<td>Empirical methods not part of industrial practice</td>
<td>Most large software development organizations conduct empirical studies as part of decisions making and process improvement</td>
</tr>
</tbody>
</table>

More relevant studies

- Relevance on topic
- Relevance of applicability of the results, see external validity later
Why is relevant topic important?

“Currently, research priorities in the IS field seem to be driven more by the interests of researchers rather than by the needs of practice or society. Hirschheim et al. (1996) see this as a good thing, in that it encourages diversity and promotes academic freedom. However, in an applied discipline, it also reflects a lack of social accountability. For example, there would be a public outcry if medical researchers spent their time researching health problems that interested them while ignoring the major health problems in society.”


More relevant studies

<table>
<thead>
<tr>
<th>State of Practice</th>
<th>Target (2020-2025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few results answer questions posed by industrial users, e.g., “Which method should we use in our context?” Current focus is on comparing mean values of technologies without a proper understanding of individual differences or the studied population</td>
<td>More focus on individualized results, individual differences, and better descriptions of populations and contexts; why, when and how technology X is better than technology</td>
</tr>
<tr>
<td>Reference points for comparisons of technologies are frequently not stated, or not relevant</td>
<td>New technology is compared with relevant alternative technology used in the software industry</td>
</tr>
<tr>
<td>One may question the industrial relevance of many SE studies</td>
<td>More case studies and action research. Experiments should show more realism regarding subjects, technology, tasks, and software systems</td>
</tr>
</tbody>
</table>
More Valid Studies

Internal validity

The *internal validity* of an experiment is “the validity of inferences about whether observed co-variation between $A$ (the presumed treatment) and $B$ (the presumed outcome) reflects a causal relationship from $A$ to $B$ as those variables were manipulated or measured” [Shadish, 2002]. Changes in $B$ may have alternative causes than the manipulation of $A$. An alternative cause for the outcome is a *confounding factor*.

---

Systematic review

- Randomized allocation of treatments to subjects is one way of handling threats to internal validity. Done in 58% of experiments [Kampenes et al. 2007].
- Randomization is not always desirable or possible in SE, hence 35% quasi-experiments.
- Only half of the quasi-experiments measured a pretest score to control for selection bias, and only 8% reported a threat of selection bias.

<table>
<thead>
<tr>
<th>State of Practice</th>
<th>Target (2020-2025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results are often not robust due to lack of replications and reliance on only one type of research design.</td>
<td>Replications and triangulation of research designs are frequently used.</td>
</tr>
</tbody>
</table>
More Valid Studies

Construct validity

- We need to measure something to understand it, but just as importantly, we need to understand something in order to measure it.
- For example: Quality = number of errors? What about functionality, usability, maintainability, etc. And what kind of errors, found where, found when? Compared with what?
- In general, low construct validity in SE studies, although little systematic investigation on this issue. Simula plans to carry out a systematic review in this area.

More Valid Studies

External validity – Generalisation

The validity of inference about whether the cause-effect relationship holds over variation in:
- Actors: individual, teams, project, organisation or industry
- Technology: process model, method, technique, tool or language
- Activities: plan, create, modify or analyze (a software system)
- Software systems: many dimensions, such as size, complexity, application domain, business/scientific/student project or administrative/embedded/real time, etc.
### Dimensions of Generalization

<table>
<thead>
<tr>
<th></th>
<th>Statistical generalization</th>
<th>Analytical generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual studies</td>
<td>Statistical hypothesis testing</td>
<td>Generalization through theory or analogy</td>
</tr>
<tr>
<td>Collection of studies</td>
<td>Meta analysis</td>
<td>Research synthesis, aggregation of evidence, and theory</td>
</tr>
</tbody>
</table>

### Generalization

<table>
<thead>
<tr>
<th>State of Practice</th>
<th>Target (2020-2025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The scope of validity of empirical studies is rarely defined explicitly</td>
<td>The scope is systematically and explicitly defined and reported</td>
</tr>
<tr>
<td>Statistics-based generalization is the dominant means of generalization</td>
<td>Studies include a diverse and reflected view on how to generalize, particularly through the use of theory</td>
</tr>
</tbody>
</table>
Challenge

- The applicability of the experimental results to industrial practices is in most cases hampered by the experiments’ lack of realism and scale regarding, that is, the challenge of achieving external validity.

- Many aspects of the complexity of software engineering only manifest themselves in controlled experiments if the experiments involve a sufficiently large number of subjects, tasks and systems, for example, differences among subgroups of subjects.

- How to generalise from SE experiments. How do we convince practitioners and managers in industry that the results are relevant to them?

Development tasks in industry usually take longer and are more complex than in most experiments:

Duration of experiments with time measurements
Example of difference among subgroups

Mean Effort (minutes)

Undergraduate  Graduate  Junior  Intermediate  Senior

Total Effect of PP

Difference from individuals

84 %

Duration  Effort  Correctness

-8 %  7 %
Moderating Effect of System Complexity on PP

- Duration: CC (easy) -20%, DC (complex) 6%
- Effort: CC (easy) 60%, DC (complex) 112%
- Correctness: CC (easy) -16%, DC (complex) 48%

Moderating Effect of System Complexity for Juniors

- Duration: CC (easy) 4%, DC (complex) 6%
- Effort: CC (easy) 109%, DC (complex) 112%
- Correctness: CC (easy) 32%
Is a helicopter better than a bike?
The effect of PP “depends on”

<table>
<thead>
<tr>
<th>Programmer expertise</th>
<th>Task complexity</th>
<th>Use PP?</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junior</td>
<td>Easy</td>
<td>Yes</td>
<td>Provided that increased quality is the main goal</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Easy</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>Yes</td>
<td>Provided that increased quality is the main goal</td>
</tr>
<tr>
<td>Senior</td>
<td>Easy</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>No*</td>
<td></td>
</tr>
</tbody>
</table>

* Unless you are sure that the task is too complex to be solved satisfactorily even by solo seniors

- The performance of the various categories may depend on their relevant education, work experience, the actual task and system, development technology, etc.
- In the survey of 113 experiments, 7 involved both students and professionals. Only 3 measured difference in performance: partly no difference, partly professionals better.

© Institutt for informatikk - Dag Sjøberg 15.10.2008

Subjects

<table>
<thead>
<tr>
<th>Subject Category</th>
<th>Reported Subject Types</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduates</td>
<td>Undergraduates, Bachelors, Third and fourth-year students, Last-year students, Honors and Majors.</td>
<td>2969</td>
<td>54.1</td>
</tr>
<tr>
<td>Graduates</td>
<td>Graduate students, Students following graduate courses or Master’s programs, MSc and PhD students.</td>
<td>594</td>
<td>10.8</td>
</tr>
<tr>
<td>Students, type unknown</td>
<td>Students in computer science, Students</td>
<td>1203</td>
<td>21.9</td>
</tr>
<tr>
<td>Professionals</td>
<td>Developers, Practitioners, Software engineers, Analysts, Domain experts, Business managers, Facilitators, Professionals.</td>
<td>517</td>
<td>9.4</td>
</tr>
<tr>
<td>Scientists</td>
<td>Professors, Post-doctorates, Staff members of educational institutions</td>
<td>74</td>
<td>1.3</td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td>131</td>
<td>2.3</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>5488</td>
<td>100</td>
</tr>
</tbody>
</table>


© Institutt for informatikk - Dag Sjøberg 15.10.2008
More Valid Studies

Statistical conclusion validity

The validity of inferences about the correlation (covariation) between treatment and outcome.

- Statistical power is the probability that a statistical test will correctly reject the null hypothesis. A test without sufficient statistical power will not provide enough information to draw conclusions regarding the acceptance or rejection of the null hypothesis.
- An effect size quantifies the effects of an experimental treatment. Whereas $p$-values reveal whether a finding is statistically significant, effect size indicates practical significance, importance, or meaningfulness.

Why that many subjects? Power analysis

Research question:

What is the effect regarding duration, effort and correctness of pair programming for various levels of system complexity and programmer expertise when performing change tasks?

- 2x2x3 fixed-effect analysis of covariance:
  - pair programming (two levels), control style (two levels) and expertise (three levels), resulting in twelve levels/groups
  - $N = 170$ (85 individuals and 85 pairs)
  - $N = 14$ in each of the 12 groups
Statistical Conclusion Validity

<table>
<thead>
<tr>
<th>State of Practice</th>
<th>Target (2020-2025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stat. methods are used mechanically, with little focus on limitations and assumptions. Populations not defined, and for experiments, lack of power analysis and effect size estimation.</td>
<td>The use of statistical methods is mature. Populations are well defined, and power analysis and effect size estimation are conducted when appropriate.</td>
</tr>
</tbody>
</table>

More Repeatable Studies

Can the study be repeated with the same results?

Repeatability of a study represents the ability of other investigators to follow the same procedures, to perform exactly the same study, and to arrive at the same findings and conclusions (Yin 2003).

(Yin uses the term “reliability” instead of “repeatability”)
Synthesizing Evidence

• **Primary**: collection and analysis of data
  – Experiments, surveys, case studies, action research, and others

• **Secondary**: research synthesis, summary, integration and combination of the findings of different primary research studies on a certain topic
  – Systematic reviews (see lecture 16.10.2007), meta-analysis

Reviews on research methods (= PhD of Vigdis By Kampenes):
• A survey of controlled experiments in software engineering
• A systematic review of statistical power in software engineering experiments
• A Systematic review of effect size in software engineering experiments
• A systematic review of quasi-experiments in software engineering

### Table: Effect size and 95% CI

<table>
<thead>
<tr>
<th>Study name</th>
<th>Effect size</th>
<th>Lower limit</th>
<th>Upper limit</th>
<th>Effect size and 95% CI</th>
<th>Relative weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P07a</td>
<td>0.11</td>
<td>-0.24</td>
<td>0.46</td>
<td></td>
<td>23.24</td>
</tr>
<tr>
<td>S06a</td>
<td>0.08</td>
<td>-0.28</td>
<td>0.43</td>
<td></td>
<td>22.85</td>
</tr>
<tr>
<td>S00</td>
<td>1.04</td>
<td>0.65</td>
<td>1.43</td>
<td></td>
<td>18.64</td>
</tr>
<tr>
<td>S03</td>
<td>0.10</td>
<td>-0.44</td>
<td>0.64</td>
<td></td>
<td>9.80</td>
</tr>
<tr>
<td>S06b</td>
<td>0.28</td>
<td>-0.32</td>
<td>0.88</td>
<td></td>
<td>7.95</td>
</tr>
<tr>
<td>P07b</td>
<td>0.69</td>
<td>-0.09</td>
<td>1.46</td>
<td></td>
<td>4.73</td>
</tr>
<tr>
<td>S02</td>
<td>0.30</td>
<td>-0.50</td>
<td>1.09</td>
<td></td>
<td>4.53</td>
</tr>
<tr>
<td>S06c</td>
<td>0.32</td>
<td>-0.69</td>
<td>1.32</td>
<td></td>
<td>2.81</td>
</tr>
<tr>
<td>S06b</td>
<td>0.51</td>
<td>-0.59</td>
<td>1.62</td>
<td></td>
<td>2.32</td>
</tr>
<tr>
<td>P98</td>
<td>0.91</td>
<td>-0.28</td>
<td>2.10</td>
<td></td>
<td>2.02</td>
</tr>
<tr>
<td>Overall effect</td>
<td>0.38</td>
<td>0.21</td>
<td>0.55</td>
<td></td>
<td>1.09</td>
</tr>
</tbody>
</table>

| b) Duration |
|-------------|-----------|-------------|-------------|------------------------|-----------------|
| P07a        | 0.21      | -0.13       | 0.54        |                        | 33.33           |
| S06a        | 0.57      | 0.07        | 1.07        |                        | 16.24           |
| S00         | 0.16      | -0.40       | 0.73        |                        | 11.83           |
| S03         | 0.55      | -0.03       | 1.14        |                        | 10.97           |
| S05b        | 1.30      | 0.64        | 1.96        |                        | 8.66            |
| P07b        | -0.59     | -1.37       | 0.20        |                        | 6.09            |
| S02         | -0.05     | -0.84       | 0.74        |                        | 6.01            |
| P02         | 0.96      | -1.08       | 2.11        |                        | 2.87            |
| S06b        | 0.98      | -0.24       | 2.40        |                        | 2.61            |
| S06d        | 0.85      | -0.54       | 3.29        |                        | 1.63            |
| P98         | 4.09      | 1.98        | 6.20        |                        | 0.84            |
| Overall effect | 0.46   | 0.21        | 0.59        |                        |                 |

| c) Effort |
|-----------|-----------|-------------|-------------|------------------------|-----------------|
| P07a      | -0.68     | -1.03       | -0.34       |                        | 47.66           |
| S06a      | -1.09     | -1.61       | -0.58       |                        | 21.17           |
| S05b      | -0.25     | -0.85       | 0.35        |                        | 19.73           |
| S05c      | -0.49     | -1.62       | 0.63        |                        | 4.46            |
| S06b      | -0.64     | -1.83       | 0.54        |                        | 4.04            |
| S06c      | -0.11     | -1.31       | 1.10        |                        | 3.88            |
| S06d      | 2.52      | 1.17        | 3.88        |                        | 3.07            |
| Overall effect | -0.57 | -0.81       | -0.33       |                        |                 |

-2.00  -1.00   0.00  1.00  2.00
Favors solo programming  Favors pair programming
## Synthesis of evidence

<table>
<thead>
<tr>
<th>State of Practice</th>
<th>Target (2020-2025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narrative, biased reviews and little appreciation of the value of systematic reviews</td>
<td>Scientific methods are used to undertake integrative and interpretive reviews to inform research and practice</td>
</tr>
<tr>
<td>The number and coverage of systematic reviews is very limited</td>
<td>Policy-makers, practitioners, and the general public have up-to-date and relevant systematic reviews and evidence-based guidelines and checklists at their disposal</td>
</tr>
<tr>
<td>Lack of common terminology and appropriate descriptors and keywords</td>
<td>The SE community is mature regarding understanding and use of basic terminology, descriptors and keywords. The electronic resources have high quality in their support of information about SE research</td>
</tr>
<tr>
<td>No common understanding of SE phenomena</td>
<td>Agreed-upon conceptual and operational definitions of key SE constructs and variables</td>
</tr>
<tr>
<td>Limited advice on how to combine data from diverse study types</td>
<td>Methods are available for synthesizing evidence from a variety of perspectives and approaches to research and practice</td>
</tr>
</tbody>
</table>

## Theory building

<table>
<thead>
<tr>
<th>State of Practice</th>
<th>Target (2020-2025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generally, little use of theories. The theories used mainly justify research questions and hypotheses; some explain results; very few test or modify theory</td>
<td>Most SE studies involve theories. Considering using, testing, modifying or formulating theory is part of any empirical work</td>
</tr>
<tr>
<td>Almost no SE-specific theories are proposed</td>
<td>Many SE theories are proposed and tested</td>
</tr>
<tr>
<td>Theories are generally poorly documented</td>
<td>There are widely used standards for describing theories in a clear and precise way</td>
</tr>
<tr>
<td>Difficult to identify the theories that actually are used or have been proposed</td>
<td>For each SE sub-discipline, there are web-sites and systematic reviews that systematize and characterise relevant theories</td>
</tr>
</tbody>
</table>
How to improve the quality of SE research?

- Increasing competence regarding how to conduct empirical studies
  - Guidelines and empirical methods included in SE curricula
  - Develop infrastructures to support the conducting of studies
- Improving the links between academia and industry
  - Get involved in SPI work in companies
  - Give seminars and courses where studies are included
- Developing common research agendas
  - More concentrated effort – SE researchers should work on common research programs
- Increasing the resources available

Increasing competence

<table>
<thead>
<tr>
<th>State of Practice</th>
<th>Target (2020-2025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers often do not build sufficiently on previous research results</td>
<td>There is a strong emphasis on building on previous research results, including those from other disciplines</td>
</tr>
<tr>
<td>Skills in conducting controlled experiments and reviews have improved, but not skills in conducting surveys, case-studies and action research</td>
<td>Research method and design elements are carefully selected and combined, based on an in-depth understanding of their strengths and weaknesses</td>
</tr>
</tbody>
</table>
Consulting related disciplines

Software engineering is typically performed by humans in organisations. Hence, Simula has established research collaborations with disciplines such as psychology, sociology and management, in addition to statistics.

Need for infrastructures, f.ex.:
The logistics of controlled experiments is work intensive and error prone

- Personal information and background information of subjects must be collected
- General information and specific task documents must be printed and distributed
- Solution documents must be collected
**Web-based tool support (SESE)**

- **Administrator**
  1: Define experiment
  2: Add participants
  3: Questionnaires
     - Task descriptions
     - Source code, design documents, etc.

- **Researcher**
  1: Define experiment
  4: Answer questions
     - Task solutions
     - Source code, design documents, etc.

During 3 & 4: Monitor Experiment
5: Collect & analyze results
Key functionality of SESE

- real-time monitoring of the experiment
- flexibility of defining new kinds of questions and measurement scales
- automatic recovery of experiment sessions
- automatic backup of experimental data
- multi-platform support for downloading experimental materials and uploading task solutions

SESE is built on top of a commercial human resource management system, and is partly being developed by an external company


Practical organisation of large experiments

- Ask for a local project manager of the company who selects subjects according to the specification of the researchers, ensures that the subjects actually turn up, ensures that the necessary tools are installed on the PCs, and carries out all other logistics, accounting, etc.
- Motivate the experiment up-front: inform the subjects about the purpose of the experiment (at a general level) and the procedure (when to take lunch or breaks, that phone calls and other interruptions should be avoided, etc.).
- Ensure that the subjects do not talk with one another in breaks, lunch, etc.
- Ensure that the subjects that the information about their performance is kept confidential (both within company and outside).
- Ensure the company that its general performance is kept confidential.
- Monitor the experiment, that is, be visible and accessible for questions.
- Give all subjects a small training exercise to ensure that the PC and tool environment are working properly.
- Ensure the company and subjects that they will be informed about the results of the experiment.
- Provide a proper experiment support environment that is used to set up and monitor the experiment, and collect and manage the experimental data.
Developing common research agendas

- Common research agendas should be established to improve empirical work per se, but also for specific SE topics, for example, distributed software development.
- A more ambitious, long-term goal would be to establish a program in software engineering similar to the Human Genome Project and CERN.
The costs of running large experiments with professionals

• “The experimental approach is not without limits. First of all, the costs are high and in some cases may become prohibitive. It is clearly impossible to do an experiment with hundreds of professionals, so smaller groups or case studies will have to suffice.”


• “practitioners are understandably skeptical of results acquired from a study of 18-year-old college freshmen.”

“finding 100 developers willing to participate in such an experiment is neither cheap nor easy. … But even if a researcher has the money, where do they find that many programmers?”


Examples of experiments at Simula

• 99 consultants from 8 companies
  – one-day experiment that compared two different object-oriented control styles
• 295 consultants from 29 companies in Norway, Sweden and the UK
  – one-day experiment that tested the effect of pair programming
• 39 consultants from 11 companies
  – Three-day experiment on design patterns
• 20 programmers from 13 companies
  – worked individually from one to two weeks in an experiment on UML
• 35 companies presented bids for a web-based system that we needed
  – 4 were selected to actually build the system independently of each other.
  – The teams (2-3 developers from each company) spent from 7 to 25 person-weeks each
• 30 companies from 11 countries in Europe and Asia presented their bids.
  – 4 companies built the system
  – each spent from 10 to 20 person-weeks
• 40 companies from 8 countries in Europe and Asia estimated five projects each
  – The spent 2-4 person-weeks each
How do we get the subjects?
– Hire consultants

- The experiments listed above cost between €50,000 and €230,000.
- We paid the companies ordinary consultancy fees for individuals or fixed price for a whole project, like any other ordinary customer.
  - The companies have routines for defining (small) projects with local project management, resource allocation, budgeting, invoicing, providing satisfactory equipment, etc.
- Difficult to find many experiment subjects employed in an in-house software development company because the management will typically prioritize the next release of their product.

Large-scale empirical work requires a great amount of resources

- At Simula we have used about 25% of budget for experiments, mainly at the expense of more researchers.
- In research grants applications, one budgets for money for positions, equipment and travel; why not include money for experiments?
- SE researchers should contribute to making the development of software systems a mature industry. Given the importance of software systems in society, there is no reason why research projects in SE should be less comprehensive and cost less than large projects in other disciplines, such as physics and medicine. The U.S. funding for the Human Genome Project was $437 million over 16 years. If many scientific activities related to genomics are included, the total cost rises to $3 billion! CERN’s annual budget is about $800 million.