A Survey on Controlled Experiments in Software Engineering

Lutz Prechelt
prechelt@inf.fu-berlin.de
Institut für Informatik, Freie Universität Berlin
2006-01-05
Source

Scope

• Controlled experiments published in nine journals and three conferences during 1993 to 2002
  • but excluding pure HCI or IS studies

• Journals:
  • ACM Transactions on Software Engineering Methodology (TOSEM)
  • Empirical Software Engineering (EMSE)
  • IEEE Computer
  • IEEE Software
  • IEEE Transactions on Software Engineering (TSE)

• Conferences
  • Intl. Conf. on Software Engineering (ICSE),
  • IEEE Intl. Symp. on Empirical Software Engineering (ISESE)
  • IEEE Intl. Symp. on Software Metrics (METRICS)

• Information and Software Technology (IST)
• Journal of Systems and Software (JSS)
• Software Maintenance and Evolution (SME)
• Software: Practice and Experience (SP&E)
# Prior Surveys

<table>
<thead>
<tr>
<th>Purpose</th>
<th><strong>Our survey</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Surveys topics, subjects, tasks, environments, and internal and external validity of controlled experiments in SE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Scope</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Journals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
</tr>
<tr>
<td></td>
<td>EASE, EMSE, ICSE, IEEE Computer, IEEE Software, ISESE, IST, JSME, JSS, METRICS, SP&amp;E, TOSEM, TSE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Sampling of papers</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All papers in the period 1993-2002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Number of investigated papers</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5453 papers scanned, 103 papers analysed in depth</td>
</tr>
</tbody>
</table>

Lutz Prechelt, prechelt@inf.fu-berlin.de
Prior Surveys (2)

There are a few other surveys

- on narrower topics, e.g.
  - object-oriented technology (Deligiannis, Shepperd, Webster, Roumeliotis 2002)
  - testing techniques (Juristo, Moreno, Vegas 2004)
  - software effort estimation (Jørgensen, Teigen, Moløkken 2004)

- on only one single conference
  - Shaw on ICSE 2002

- or with theory-formulating intent
  - Zendler on 31 controlled experiments (2001)
Definition: Controlled Experiment

• **Experiment:**
  • A study in which an intervention is deliberately introduced to observe its effects

• **Controlled Experiment:**
  • Comparing the effects of two (or more) different interventions, while keeping all other conditions constant
  • Note: If people are involved, constancy usually requires
    • multiple trials for each condition and using averaging,
    • random assignment of subjects to interventions
  • However, the survey also includes *quasi-experiments*, in which random assignment is missing
Number of articles per source

<table>
<thead>
<tr>
<th>Journal/Conference</th>
<th>articles investigated</th>
<th>reporting controlled experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>EMSE</td>
<td>124</td>
<td>22</td>
</tr>
<tr>
<td>ISESE</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>METRICS</td>
<td>177</td>
<td>10</td>
</tr>
<tr>
<td>JSS</td>
<td>886</td>
<td>24</td>
</tr>
<tr>
<td>TSE</td>
<td>687</td>
<td>17</td>
</tr>
<tr>
<td>ICSE</td>
<td>520</td>
<td>12</td>
</tr>
<tr>
<td>IST</td>
<td>745</td>
<td>8</td>
</tr>
<tr>
<td>SME</td>
<td>186</td>
<td>2</td>
</tr>
<tr>
<td>IEEE SW</td>
<td>532</td>
<td>4</td>
</tr>
<tr>
<td>TOSEM</td>
<td>125</td>
<td>1</td>
</tr>
<tr>
<td>IEEE Comp</td>
<td>780</td>
<td>0</td>
</tr>
<tr>
<td>SP&amp;E</td>
<td>671</td>
<td>0</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>5453</strong></td>
<td><strong>103</strong></td>
</tr>
</tbody>
</table>
Number of experiments

- 103 articles describing experiments
  - 207 scientists from 109 institutions in 19 countries are involved as authors
- 12 of the articles report more than one experiment
- 4 experiments occur in more than one article

Density of experiments across articles:
- Tichy et al. survey: 14% of articles in SE journals describe empirical work of some kind
- Glass et al. survey: 3-4% describe controlled experiments
- Zelkowitz/Wallace survey: 3% describe controlled exp.
- Here: 2% describe controlled experiments
Who did them: Individuals

- article with n authors ➔ weighted by 1/n
- V. Basili on rank 19,
  H.-D. Rombach not among top 20
- average author
Who did them: Institutions

Distorted, because researchers switch institutions
Topics of articles

- The experiment topics cover a broad range
- But most topic areas are represented by only 1 or 2 articles
  - 34 areas are mentioned overall
- Notable exceptions:
  - Inspections and reviews: 35 articles
  - OO design methods: 8 articles
  - Process models: 5 articles
Subjects

<table>
<thead>
<tr>
<th>Subject Category</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduates</td>
<td>2969</td>
<td>54.1</td>
</tr>
<tr>
<td>Graduates</td>
<td>594</td>
<td>10.8</td>
</tr>
<tr>
<td>Students, type unknown</td>
<td>1203</td>
<td>21.9</td>
</tr>
<tr>
<td>Professionals</td>
<td>517</td>
<td>9.4</td>
</tr>
<tr>
<td>Scientists</td>
<td>74</td>
<td>1.3</td>
</tr>
<tr>
<td>Unknown</td>
<td>131</td>
<td>2.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>5488</td>
<td>100</td>
</tr>
</tbody>
</table>

- 7 articles are on experiments where both students and professionals were present
- Only 3 of them assess the differences.

Findings:
1. no difference
2. professionals better
3. 3 tasks: no difference in two, professionals better in the third
Subject mortality

• Mortality is the dropping-out of subjects during the experiment
  • No or no complete results can be reported for these subjects
  • If mortality depends on the intervention (group membership), the effect damages the random assignment and threatens the experiment's validity

• Mortality is reported for only 24 experiments
  • and for them was 2% on average
Information about the subjects

- "In order to generalize from an experiment [...] one needs information about various characteristics and their variation both in the sample and in the group to which the results will be generalized (target population).
- [...] However, there is no generally accepted set of background variables for guiding data collection in a given type of study, simply because the software engineering community does not know which variables are the important ones."
### Information about the subjects (2)

- **91 articles on experiments with students reported**
  - gender: for 7 experiments
  - age: 6
  - grades: 6
  - programming experience (general): 17
  - prog. exp. (years, #langs): 11
  - has industrial work experience: 9
  - years of industrial work exp.: 9
  - task-related experience: 64
  - task-related training: 27

- **27 articles on experiments with professionals reported:**
  - gender: 2
  - age: 3
  - job type: 7
  - degrees: 3
  - programming experience (general): 2
  - prog. exp. (years, #langs): 7
  - prog. exp. self-assess: 2
  - years of industrial work exp.: 5
  - task-related experience: 14
  - task-related training: 12
Information about the subjects (3)

Examples of reports on programming experience:

- **Negative example:**
  - "Some of the students had industrial programming experience."

- **Positive example:**
  - "On average, subjects’ previous programming experience was 7.5 years, using 4.6 different programming languages with a largest program of 3510 LOC. Before the course, 69 percent of the subjects had some previous experience with object-oriented programming, 58 percent with programming GUIs"
Subject recruitment and compensation

- **Subjects participation was**
  - mandatory: 12 student experiments
  - voluntary: 29 experiments
  - not reported: 72 experiments

- **Student subject compensation was**
  - better grades: 10 experiments
  - extra credits: 9 experiments
  - money: 3 experiments
  - exhibition trip: 1 experiments
  - no reward: 16 experiments
  - nothing reported: 74 experiments
Subject sampling

• Only 1 article explicitly reported random sampling from a defined target population
• A few claimed random sampling
  • but did not report population or procedure
• By far most studies used convenience sampling
  • which may introduce bias
Task type and duration

Tasks are classified into four broad categories:

- Plan (plan, estimate, negotiate) 10%
- Create (design, code) 20%
- Modify a design or code 16%
- Analyze (inspect (37%), test/debug, comprehend) 54%

- Work time per subject is reported for only 41 exp.
  - See figure for distribution of median work times:
  - Two thirds of all tasks have a median duration of two hours or less
Task duration by type

- the 41 experim. with detailed time data only
- 4 missing: 18, 55 h Coding, 25 h change code, 18 h design
Programs/materials under study

- Only 16 experiments expressly use 'real' applications
  - 10 for inspection tasks
  - 2 for design tasks

- Only 67 experiments report on size of materials
  - range from 1 page to 67 pages for documents
  - range from 25 LOC to 3955 LOC for code
  - partially dependent on task type, e.g. larger for inspections

<table>
<thead>
<tr>
<th>Application type</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constructed</td>
<td>80</td>
<td>70.8</td>
</tr>
<tr>
<td>Commercial</td>
<td>16</td>
<td>14.2</td>
</tr>
<tr>
<td>Student project</td>
<td>5</td>
<td>4.4</td>
</tr>
<tr>
<td>Open source</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Unclear</td>
<td>12</td>
<td>10.6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>113</td>
<td>100</td>
</tr>
</tbody>
</table>
Experiment location

- classroom setting 10
- laboratory 20
- "academic" setting 9
- real office 1
- not reported 73

- Only about half report the name of the institution (university, company)
Experiment equipment

- Almost half of the experiments do not report on the tools used by the participants!
  - Most of these are probably pen-and-paper
- Pen-and-paper may be OK for inspections, but is dubious for most other types of task

<table>
<thead>
<tr>
<th>Tool</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC or workstation (only)</td>
<td>32</td>
<td>28.3</td>
</tr>
<tr>
<td>Pen and paper (only)</td>
<td>25</td>
<td>22.1</td>
</tr>
<tr>
<td>Combination</td>
<td>5</td>
<td>4.4</td>
</tr>
<tr>
<td>Unknown</td>
<td>51</td>
<td>45.1</td>
</tr>
<tr>
<td>Total</td>
<td>113</td>
<td>100</td>
</tr>
</tbody>
</table>
Replication

- 20 of the exp. call themselves replications of others
  - 7 of them in inspections, 5 in maintenance
- 5 are close replications, 15 are differentiated
  - 4 with different programs, 3 with different tasks,
    many with different kinds of subjects
- All 5 close replications confirm the original results
- Of 7 differentiated repls. performed by other authors, only 1 confirms the original results
- Of 8 differentiated repls. performed by other authors, 7 confirm the original results
Threats to internal validity

<table>
<thead>
<tr>
<th>Category</th>
<th>Threat not handled</th>
<th>Threat reduced</th>
<th>Threat eliminated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>10</td>
<td>35</td>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td>Instrumentation</td>
<td>9</td>
<td>30</td>
<td>6</td>
<td>45</td>
</tr>
<tr>
<td>Maturation</td>
<td>3</td>
<td>14</td>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>Testing</td>
<td>2</td>
<td>22</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>History</td>
<td>3</td>
<td>9</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>Attrition</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Regression</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Ambiguous Temporal Precedence</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Additive and Interactive Effects</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No of threats*</td>
<td>32 (17.8%)</td>
<td>114 (63.3%)</td>
<td>34 (18.9%)</td>
<td>180 (100%)</td>
</tr>
<tr>
<td>No of Experiments</td>
<td>26 (23.0%)</td>
<td>55 (48.7%)</td>
<td>19 (16.8%)</td>
<td>71† (62.8%)</td>
</tr>
</tbody>
</table>

- Many articles fail to discuss internal validity
- About half of all experiments appear to be quasi-experiments only (selection effects)
Threats to external validity

<table>
<thead>
<tr>
<th>Factors addressed as threats to external validity</th>
<th>Experiments</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject (only)</td>
<td>14</td>
<td>12.4</td>
</tr>
<tr>
<td>Task (only)</td>
<td>10</td>
<td>8.8</td>
</tr>
<tr>
<td>Environment (only)</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td>Subject and environment</td>
<td>2</td>
<td>1.8</td>
</tr>
<tr>
<td>Subject and task</td>
<td>31</td>
<td>27.4</td>
</tr>
<tr>
<td>Subject, environment and task</td>
<td>14</td>
<td>12.4</td>
</tr>
<tr>
<td>Treatment and subject, task or environment</td>
<td>6</td>
<td>5.3</td>
</tr>
<tr>
<td>Threats to external validity not addressed</td>
<td>35</td>
<td>31.0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>113</td>
<td>100</td>
</tr>
</tbody>
</table>

- More than half do not contain adequate discussion
  - none at all or one area of concern only
- Few consider the environment or the treatment
- Reporting is rather unsystematic
Conclusions

- The analyses performed in such a survey cannot guarantee that an article is good
  - "Good" meaning credible and reliable
- However, many of its aspects can point out where an article is not good
- The average quality in these respects is not very good
  - so it is not really difficult to be in the upper half
  - but there are quite a number of issues to think of!

- Want to learn how to do good experiments?
- Come to "Empirische Bewertung in der Informatik"
  (2V+2Ü, 6 LP, SoSe)
Thank you!