Experiments in SE
Promises & Perils

Natalia Juristo
Universidad Politécnica de Madrid (Spain)
Experiment?

Some challenges

Replication

Statistics

Industry

Content
Experiment?
From Beliefs to Scientific Knowledge

The information used to build artefacts goes through stages, ranging from beliefs, speculations and lucky guesses to scientific knowledge by means of which an engineering discipline achieves predictable results.

Mary Shaw

*Prospects for an engineering discipline of software*

IEEE Software 1990

- Engineering disciplines are grounded on objectivity.

- **Objective data** are helpful for finding out more about reality, whereas subjective opinions can lead to mistaken perceptions of reality.

Empiricism gets information directly from reality.
Several Types of Empirical Studies

- Observational Studies
- Interventional Studies

Historical Archives
Case Studies
Surveys
Quality of Evidence
Experiment Definition

Empirical procedure where key variables of a reality are manipulated in a controlled environment to investigate the impact of such variations.
How is a SE Experiment

Using A

Software process

Using B

Software process

\[ MA = \text{Average} (m_{ia}, m_{ib}, m_{ic}) \]

\[ MB = \text{Average} (m_{id}, m_{ie}, m_{if}) \]
Experiment Definition

Empirical procedure where key variables of a reality are manipulated in a controlled environment to investigate the impact of such variations.
The controlled environment aims to guarantee that no other variable is impacting results.

Such warranty allows to reveal causality.
Field experiments
Purposely intervention
Natural uncontrolled environment
Observational study
No intervention
In a natural uncontrolled environment
Causality and Data Analysis

- Note that statistical techniques for data analysis can be applied to data collected from any kind of empirical study.

- Cause-effect relationships are revealed thanks to design not to data analysis techniques.
  - An empirical study designed as an experiment reveals causality.
  - Other type of empirical studies reveal only a relationship.
## Misuse of the term experiment in MSR Research

<table>
<thead>
<tr>
<th>Venue 2015</th>
<th>Use of Term Experiment</th>
<th>MSR vs Traditional experiment</th>
<th>MSR Use vs. Misuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESEM</td>
<td>30.5% 11 out of 36</td>
<td>72.7% MSR Works (8 papers)</td>
<td>Observational: 12.5% Data experiments: 87.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27.3% traditional experiments (3 papers)</td>
<td></td>
</tr>
<tr>
<td>MSR</td>
<td>42.8% 18 out of 42</td>
<td>100% MSR Works (18 papers)</td>
<td>Observational: 44.4% Data experiments: 55.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0% traditional experiments</td>
<td></td>
</tr>
<tr>
<td>EMSE</td>
<td>52.72% 29 out of 55</td>
<td>65.5% MSR Works (19 papers)</td>
<td>Observational: 52.6% Data experiments: 47.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34.5% traditional experiments (10 papers)</td>
<td></td>
</tr>
</tbody>
</table>
Challenges of SE experiments
Statistical Shortcomings

To know more…
Rolando Reyes, Oscar Dieste, Efraín Fonseca, Natalia Juristo
*Statistical errors in software engineering experiments: a preliminary literature review*
ICSE (2018)
Statistical Errors in Other Areas

- Papers published in Psychology journals: 15% Wrong, 85% Right
- Papers published in Obstetrics journals: 19% Wrong, 81% Right
- Papers published in Psychiatry journals: 40% Wrong, 60% Right
- Papers published in general medical journals: 60% Wrong, 40% Right
Statistical Errors

Errors are simple (apparently)
✓ definition of statistical hypotheses
✓ interpretation of p-values
✓ sample size calculation
✓ significance levels
✓ confidence intervals
✓ ...

Evaluation instrument

Q1.1 Are null hypotheses explicitly defined?
Q1.2 Are alternative hypotheses explicitly defined?
Q2 Has the required sample size been calculated?
Q3 Have subjects been randomly selected?
Q4 Have subjects been randomly assigned to treatments?
Q5 Have the test assumptions (i.e., normality and heteroskedasticity) been checked or, at least, discussed?
Q6 Has the definition of linear models been discussed?
Q7 Have the analysis results been interpreted by making reference to relevant statistical concepts, such as p-values, confidence intervals, and power?
Q8 Do researchers avoid calculating and discussing post hoc power?
Q9 Is multiple testing, e.g., Bonferroni correction, reported and accounted for?
Q10 Are descriptive statistics, such as means and counts, reported?
Limited awareness

- Dybå et al.'s paper on statistical power.
- Miller's paper on meta-analysis.
- Two papers by Kitchenham and Vegas et al. on within-subject designs.
- Little more...
Review of ICSE papers

Sample

• Between 2006-2015 (10 yrs.)
  • 21 experiments
  • Plenty of empirical studies as part of evaluation sections
    • We chose randomly 3 “experiments as evaluations” per year (30 studies)
Main results

- Standalone experiments contain 3.7 errors in average
  - Range 1-6

- Experiments as evaluations contain 4.2 errors
  - Range 2-6
Likely Causes

- Survey to authors
- Insuficiente statistical training
  - Due to recent adoption of experimentalism in SE
- Improve training of novices
Experiments promise identification of cause-effect relationships among development variables.

But, watch out for statistic!!
Lack of Replications

To know more…
Omar S. Gómez, Natalia Juristo, Sira Vegas
Understanding replication of experiments in software engineering: A classification
Information & Software Technology (2014)

Adrian Santos, Omar S. Gómez, Natalia Juristo
Analyzing Families of Experiments in SE: A Systematic Mapping Study
Transactions on SE (2018)

Adrian Santos and Natalia Juristo
Comparing Techniques for Aggregating Interrelated Replications in Software Engineering
ESEM (2018)
Quality of Evidence
Value of Replication

Without replication it is impossible to distinguish whether results

- occurred by chance
- are artifactual
  the event occurs only in the experiment, not in reality
- really correspond to a regularity
Two Roles of Replication

- Validation

- Learning
  - For immature experimental knowledge, the first step is to find out which experimental conditions should be controlled.
  - As more replications of a baseline experiment are run, different conditions influencing the results of the experiment are identified.
Learning & Validation Process

1. Start with **identical replications**
   - At the beginning of experimental research, equality, even if targeted, will not happen
     There will be either invisible but significant changes in conditions or induced changes due to context adaptation or both
   - Failure to get the expected results should be seen as a step towards the discovery of some new factor

2. Later on **changes in the configuration** will be made purposely
   - to learn more variables and
   - rule out artifactual results
Most SE experiments lack replication

[Da Silva et al. EMSE 2013]
But Improving...

- We conducted a SMS on families of experiments
  - Family = 3 replacements or more
- We found 39 families
  - 11 already identified in Da Silva (before 2010)
  - 28 new (after 2010)
Replications / Family
Changes across Families
Aggregation Techniques

- Aggregation provides a textual summary of results (no p-value or effect size)
- AD does not allow to study moderator variables
- IPD fails to account for heterogeneity of results across experiments
- Megatrial does not allow to study moderator variables
- IPD stratified does not provide any effect size

Only 15% fully right!!
Replications promise more reliable experimental findings.

But, watch out for changes and aggregation!!
Difficulties with Field Experiments
Field experiments
Purposely intervention
Natural uncontrolled environment
Types of Difficulties

• Participants Recruitment
• Participants Volatility
• Participants Behavior
• Technologies Variety
Recruitment

• Hard to sign up a significant number of participants
  – Developer time is money
  – Low numbers in all cases
    • 8 -20

• Training was the only carrot that we found

• Better cases
  – Companies with booked time for training were easier to convince
    • F-Secure
  – Project leader as champion better than innovation managers
    • Mapfre & Playtech
Participants Volatility

• Fewer attendees than signed up

• More drop-outs than in academia
  – Missing data points that threaten validity
    • Paf

• Sometimes different profile than expected
  – Redesign on the fly
    • Ericsson
Participants

- Planned: 14 subjects
- Real: 13 subjects
- Useful: 8 subjects
  - Data removed: 5 subjects
    - 4 attended only 1 session
    - Group 3 had only 1 subject

The importance of staying on to perform all experimental tasks was not well understood

Loosing a group meant that we were unable to compare all treatments for several tasks

<table>
<thead>
<tr>
<th>GROUP</th>
<th>DAY 1</th>
<th>DAY 2</th>
<th>DAY 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BSK</td>
<td>SS</td>
<td>MR</td>
</tr>
<tr>
<td>2</td>
<td>SS</td>
<td>MR</td>
<td>BSK</td>
</tr>
<tr>
<td>3</td>
<td>MR</td>
<td>BSK</td>
<td>SS</td>
</tr>
</tbody>
</table>
### Planned Design

Expected subjects: experienced in C++, Eclipse, Boost and unit testing

<table>
<thead>
<tr>
<th>TRAINING</th>
<th>EXERCISES</th>
<th>TREATMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAY 1</td>
<td>Testing Tool Concepts</td>
<td>2 Tool Exercises 1 Motivational Exercise (ITL)</td>
</tr>
<tr>
<td>DAY 2</td>
<td>Slicing</td>
<td>2 Slicing Exercises</td>
</tr>
<tr>
<td>DAY 3</td>
<td>TDD</td>
<td>3 TDD Exercises</td>
</tr>
</tbody>
</table>

### Real Design

Real subjects: very inexperienced in Boost & unit testing; inexperienced in C++

<table>
<thead>
<tr>
<th>TRAINING</th>
<th>EXERCISES</th>
<th>TREATMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAY 1</td>
<td>Testing Tool Concepts</td>
<td>2 Tool Exercises 1 Motivational Exercise (ITL)</td>
</tr>
<tr>
<td>DAY 2</td>
<td>Slicing</td>
<td>2 Slicing Exercises</td>
</tr>
<tr>
<td>DAY 3</td>
<td>TDD</td>
<td>3 TDD Exercises</td>
</tr>
</tbody>
</table>
Participants Behavior

- Professionals are less motivated than students
  - Attendance of a training course <> grading
  - Preoccupied with work issues
  - Used to flexible schedule
  - Young participants more active and enthusiastic than older ones
  - There might be several other psychological issues

- Treatment compliance is lower than for students
  - Students appear to be more willing to abide by the rules defined by instructors
  - Professionals tend to have their own ideas about what they expect to get from the course/experiment

- Professionals might be afraid of being assessed
  - Some subjects removed their code
Technologies

• Language, IDE, testing framework were different across companies

• Experimental instruments had to be adapted several times for different companies
  – Originally for Java, Eclipse and JUnit (for academic setting)
  – Adapted to C++, C#, Boost, Google Test, IntelliJ

• We missed interesting instruments due to adaptations
  – As treatment adherence (only available for Java)
The concept of field experiment needs more research. Adoption to SE is not that simple.

Wild ideas

Volunteers rather than companies
Not co-located but sequential designs
Remote experiments

Improve understanding on subjects validity

Students might be not as bad as we thought as experimental subjects. Access professionals through volunteers.
Field experiments promise generalizable findings

But, accommodation of the concept is needed!!
Experimental paradigm not yet fully adopted

<table>
<thead>
<tr>
<th>Improve use of concepts</th>
<th>-</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improve use of techniques</td>
<td>-</td>
<td>Data analysis</td>
</tr>
<tr>
<td>More replications</td>
<td>-</td>
<td>Better aggregation</td>
</tr>
<tr>
<td>Develop further the concept of field experiment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Questions