Empirical validation of the Invariant Refinement Method – a controlled experiment

Ilias Gerostathopoulos
iliasg@d3s.mff.cuni.cz
Useful material


Matthias Galster

April 8, 2014
Main result and its interpretation

“We observed a (statistically valid) increase in the effectiveness of the designers when they were using the Invariant Refinement Method compared to using design intuition and ingenuity.”
Empirical research: what, why, types of studies

What software engineers can learn from social scientists ...
Empirical research – what and why

Scientific method
The world is observed and a model is built based on the observation (e.g., a simulation model)

Engineering method
The current solutions are studied and changes are proposed, and then evaluated
A model is proposed and evaluated through empirical studies, e.g. case studies, experiments

Empirical method
A formal theory is proposed and then compared with empirical observations

Analytical method

So, why?
- Anecdotal evidence or “common-sense” often not good enough
- Evidence important for successful technology transfer
- Scientists are searching for “truth”
Why? ...examples

- Identify problems with the current state of practice / art
- Characterize properties of new tools / techniques
- Obtain evidence that approach A is better than approach B

“studying objects in their natural setting and letting the findings emerge from the observations”

Empirical research can be *exploratory* and/or *explanatory*

“quantifying a relationship or to compare two or more groups with the aim to identify a cause-effect relationship”
Types of empirical studies

Primary studies

**Academia**

- Phenomena studied by probing reality using questionnaires or interviews

**Industry**

- Phenomena studied in a lab set up to mimic reality, often using students

**Controlled Experiments**

- Phenomena studied in controlled environments, often using students

**Surveys**

- Phenomena studied through intervention observation and reflection

**Field Experiments**

- Phenomena studied in context using open-ended interviews and questionnaires

**Fieldwork**

- Phenomena studied in their "real" environment using practitioners

**Off-line research**

- Do not generate any data from direct measurements, but analyze a set of primary studies and usually seek to aggregate the results from these in order to provide stronger forms of evidence about a phenomenon

**Secondary studies**

- Systematic literature reviews
  - Do not generate any data from direct measurements, but analyze a set of primary studies and usually seek to aggregate the results from these in order to provide stronger forms of evidence about a phenomenon

- Mapping studies

What will you accept as empirical truth?

It depends on which philosophical stance you adopt.

**Positivism** (reductionism): all knowledge must be based on logical inference from a set of basic observable facts
- **Control experiment**, survey, case study

**Constructivism** (interpretivism): knowledge cannot be separated from its human context: interpretations of what a theory means are just as important in judging its truth as the empirical observations on which it is based
- **Ethnography**, exploratory case study, survey

**Critical Theory**: knowledge is judged by its ability to free people from restrictive systems of thought
- **Action research**, case study

**Pragmatism**: all knowledge is approximate and incomplete, and its value depends on the methods by which it was obtained; knowledge is judged by how useful it is for solving practical problems
- **Mixed** methods: different methods to shed light to the issue under study

Controlled experiment: process and illustration
Experiment process

Experiment idea

Experiment scoping → Experiment planning → Experiment operation → Analysis & interpretation → Presentation & package → Experiment report

Figure source: Wohlin, C., Runeson, P., Host, M., Ohlsson, M.C., Regnell, B., Wesslen, A.: Experimentation in Software Engineering. Springer (2012).
Scoping – what is the goal of the experiment?

<table>
<thead>
<tr>
<th>Analyze</th>
<th>products, processes, resources, models, metrics, theories</th>
<th>Object of study</th>
</tr>
</thead>
<tbody>
<tr>
<td>for the purpose of</td>
<td>evaluate, characterize, evaluate, predict, control, change</td>
<td>Purpose</td>
</tr>
<tr>
<td>with respect to</td>
<td>effectiveness, cost, reliability, maintainability, portability</td>
<td>Quality focus</td>
</tr>
<tr>
<td>from the point of view of</td>
<td>developer, maintainer, manager, customer, researcher</td>
<td>Perspective</td>
</tr>
<tr>
<td>in the context of</td>
<td>subjects, objects, treatments, etc.</td>
<td>Context</td>
</tr>
</tbody>
</table>

Analyze the IRM modeling process for the purpose of evaluation with respect to effectiveness from the point of view of the researcher in the context of MS and PhD students producing DEECo artifacts from requirement artifacts products, processes, resources, models, metrics, theories, evaluate, characterize, evaluate, predict, control, change, developer, maintainer, manager, customer, researcher, subjects, objects, treatments, etc.
Planning

Figure source: Wohlin, C., Runeson, P., Host, M., Ohlsson, M.C., Regnell, B., Wesslen, A.: Experimentation in Software Engineering. Springer (2012).
Planning

Context Selection
We will conduct an off-line (vs on-line) experiment with students (vs professionals), addressing toy (vs real) problems, in a general (vs specific, e.g., bound to a specific company/environment) frame.

Hypothesis(es) formulation
• Null Hypothesis: “Using IRM to design a DEECO-based system is not more effective than using design intuition and ingenuity.”
• Alternative Hypothesis: “Using IRM to design DEECO-based system is more effective than using design intuition and ingenuity.”

Variables selection
• Independent variable
  • design method for DEECO-based systems, values={IRM, no-systematic-method}, scale: nominal
• Dependent variables
  • Score of final DEECO design artifact, range=0..100, scale: ratio
  • Increase in the confidence about the correctness of final DEECO design artifact, range: 1-5 Likert, scale:
Variables and scales

Nominal (Categorical): maps attribute of entity to name/symbol

Ordinal (Rank-order): Ranks entities after an ordering criterion

Interval: Difference between two measures are meaningful, but not the value itself

Ratio: There exists a meaningful zero value and ratio between two measures is meaningful
Variables selection
Target population: Junior Software Engineers
Convenience sampling: the nearest and most convenient persons are chosen.
Other options: simple random, systematic, stratified random, quota
Sample size: 10 x 10 (both MS and PhD students)

Experiment design
Multi-test within object study → one object, many subjects
Balancing: control and treatment group have the same number of students

Pseudo-randomized design (vs paired comparison/crossover)
→ Standard design for “one factor with two treatments”
Validity evaluation

- **Conclusion validity**: threats that limit our ability to draw the correct conclusion about relations between the treatment and the outcome
- **Internal validity**: threats that can affect the independent variable with respect to causality, without the researcher’s knowledge
- **Construct validity**: threats that limit our ability to generalize the result of our experiment to the concept or theory behind the experiment
- **External validity**: threats that limit our ability to generalize of our experiment to industrial practice
Lessons learned

- Plan for outliers
- Wrong to put both groups in the same room
- Inspect the actual outcome of the experiment
- New study with more than one objects?