Views on
Internal and External Validity in Software Engineering

Janet Siegmund  Norbert Siegmund  Sven Apel
Let’s start with a little experiment

What is more productive to use, a statically or dynamically typed language?

How would you try to answer this research question?
In the wild (realistic)

Maximize external validity

Reveals generally occurring effects

Causes of effects unclear

In the lab (controlled)

Maximize internal validity

Reliably explains the causes of effects

Hard to generalize

A fundamental tradeoff!
Example

83 886 080 (i.e., Germany)
Views on Internal and External Validity – Janet Siegmund

Literature Survey

• Goal: get an overview of the awareness of and the choices regarding this tradeoff

• Data set: 405 full papers
  – ICSE (2012, 2013)
Results of the Literature Survey

- 94% of the papers used an empirical method
- 8% reported on a replication study
- 46% did not mention explicitly threats to validity
- 28% discussed threats, but did not differentiate

Replication: a repetition of an experiment under similar conditions, but with specified variation [Wohlin et al.]
Who cares?

The NIPS experiment (consistency in reviewing)
   PC split into two independent committees
   166 submissions have been reviewed by both

Close to random acceptance!

What can cause such a randomness?

Are there different expectations by reviewers?

http://blog.mrtz.org/2014/12/15/the-nips-experiment.html
Survey

Goal: What does the community think?

Research questions:

- Assess the **awareness of** the community of the **tradeoff** between external and internal validity
- Assess the opinion of the community regarding **how to address this tradeoff**
- Assess the opinion of the community regarding the **role of replication**
Survey Setup: Participants

PC Members from 2010 to 2013 → key players

- ASE (Automated Software Engineering)
- EASE (Evaluation and Assessment in Software Engineering)
- ECOOP (Object-Oriented Programming)
- EMSE (Empirical Software Engineering)
- ESEC/FSE (Foundations of Software Engineering)
- ESEM (Empirical Software Engineering and Measurement)
- GPCE (Generative Programming)
- ICPC (Program Comprehension)
- ICSE (Software Engineering)
- ICSM (Software Maintenance)
- OOPSLA (Object-Oriented Programming)
- TOSEM (Software Engineering and Methodology)
- TSE (Software Engineering)

807 people contacted
94 completed the survey (typical 10% response rate)

Online questionnaire (May 2014)
**Questionnaire**

1. **Which option would you prefer for an evaluation?**
   - Maximize **internal** validity
   - Maximize **external** validity
   - ... (options may be missing or not visible in the image)

2. **Did you recommend to reject a paper in the past mainly for the following reasons?**
   - Internal validity too low
   - External validity too low
   - Yes
   - No

3. **Do you think we need to publish more experimental replications?**
   - Yes
   - No

---

**Table**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer options</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Which option would you prefer for an evaluation?</td>
<td>Max. internal validity, Max. external validity</td>
</tr>
<tr>
<td>2. Would it be a reason to reject a paper that does not choose your favorite option?</td>
<td>No preference</td>
</tr>
<tr>
<td>3. Did you recommend to reject a paper in the past mainly for the following reasons?</td>
<td>Int. validity too low, Ext. validity too low</td>
</tr>
<tr>
<td>4. For research questions like the one presented above (FP vs. OOP), do you prefer more practically relevant research or more theoretical (ground) research?</td>
<td>Yes, No, No preference</td>
</tr>
<tr>
<td>5. Do you have any suggestions on how empirical researchers can solve the dilemma of internal vs. external validity of empirical work in computer science?</td>
<td>Yes, No, Open</td>
</tr>
<tr>
<td>6. During your activity as a reviewer, how often have you reviewed a replicated study?</td>
<td>Never, Sometimes, Regularly</td>
</tr>
<tr>
<td>7. In general, how were the replications rated by you... by your fellow reviewers?</td>
<td>Accept, Borderline, Reject</td>
</tr>
<tr>
<td>8. As a reviewer of a top-ranked conference, would you accept a paper that...</td>
<td>Yes, No, I do not know</td>
</tr>
</tbody>
</table>

---

**Thank you...**
Analysis Method: Open Card Sorting

Looking for higher-order themes in open answers using cards

19 open questions × 2h per question = 38h for 776 answers
Results RQ1

Mixed degree of awareness of the tradeoff!

“[maximizing internal validity] [w]ould show no value at all to SE community”

“Without internal validity, the results cannot be trusted”

Opinions differ when human subjects are involved

“Removing humans from the exercise reduces the challenges for internal validity. In that context, knowing how general the approach was would seem a more important issue to address.”

“It makes no difference with or without humans! We are talking about software technologies”

Awareness of community regarding tradeoff between external and internal validity
Results RQ1

Assess the awareness of the community of the tradeoff between external and internal validity

Which option would you prefer for an evaluation?

<table>
<thead>
<tr>
<th></th>
<th>With Human subjects</th>
<th>Without Human subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I</strong> (Internal)</td>
<td>47</td>
<td>37</td>
</tr>
<tr>
<td><strong>E</strong> (External)</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td><strong>N</strong> (No preference)</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

**Internal**

**External**

**No preference**
Results RQ2

A single study is not seen as piece of the puzzle, but requires immediate practical impact

“I am worried that maximizing internal validity easily creates overly academic papers that provide little impact. [...]”

“[studies in medicine or biology] have hundreds/thousands of participants, over several years, and address very narrow issues [...]. We don’t see there studies that use 20 participants, are done in 2 months, and attempt to answer questions of the caliber ‘is CT better than MRI’.”

Assess the opinion of the community regarding how to address this tradeoff
Results RQ2

Do you prefer more practically relevant or more theoretical (ground) research?

Assess the opinion of the community regarding how to address this tradeoff

- Ground
- Practical
- No preference
Results RQ3

How often have you reviewed a replication?

- Never
- Regularly
- Sometimes

How were the replications rated?

- By you?
- By others?

- Accept
- Reject
- Borderline

Do we need more replications?

- Yes
- No
Results RQ3

Most participants appreciate replications, but see that they are difficult to conduct and publish.

“I have seen few replications (and perform myself a few) because they are too difficult to publish: there will always be a (dumb) reviewer to say ‘this is not novel!’...”

“Getting a publication accepted that doesn’t contribute anything but a new experiment while assessing the same question (not even adding artifacts) is a good example of hunting for publications just for the sake of publishing. Come on.”

Assess the opinion of the community regarding the role of replication
Results RQ3

Assess the opinion of the community regarding the role of replication

Replications are appropriate to address the tradeoff, but:

**Missing incentives**

“It seems that replication is rarely done since it is costly, hard to do (often not all details, tools, software, or datasets involved in an earlier study are available), and it carries a low-impact factor (at least, in certain venues).”

**No standards on how to conduct replications**

“It depends [...] whether the findings contradict the previous ones [...]”.
Further Insights

External validity = practicality?

“[...] external validity is very important since it provides indications about the potential for industrial adoption.”

Empirical study = paper?

“Excuse me, but are we discussing science and the way it should be done, or how to prepare papers to be accepted?”

Empirical research not for its own sake

“[...] a good example of hunting for publications just for the sake of publishing. Come on.”
Bottom Line?

Reviewer: “We do not know what we are doing.”
So, what can we do?

**Reviewers**

- Appreciate internally valid studies and don’t confuse external validity with practicality
- Don’t pay lip service to proper replications, but view them as an important piece of the puzzle
- Develop standards on how to assess (replication) studies

**Authors**

- Conduct multiple studies (internally and externally valid)
- Do not necessarily map an experiment 1:1 to a paper
- Report on validity issues and be concrete
Views on Internal and External Validity in Software Engineering

See supplementary web site:
https://www.tu-chemnitz.de/informatik/ST/research/material/ese/
Threats to Internal Validity

- Are PC and EB members the key players?
- Which subset of PC and EB members responded?
- Rosenthal effect (wording of questions)
- ...

Views on Internal and External Validity - Janet Siegmund
Threats to External Validity

• Only small and specific sample of the SE community
• Do the results translate to other subcommunities (RE, MODELS, MSR, PLDI, ...)?
• ...
Replication of empirical studies in software engineering research: a systematic mapping study

Fabio Q. B. da Silva • Marcos Suassuna • A. César C. Tatiara B. Gouveia • Cleiton V. F. Monteiro • Igor Eben

Published online: 1 September 2012
© Springer Science+Business Media, LLC 2012
Editor: Natália Justo

Abstract In this article, we present a systematic mapping of empirical studies in software engineering research. The goal is to plot the landscape of current empirical studies and identify gaps in the research. We applied the study to identify the research that has been published and select published articles, and to extract and synthesize the reported replications. Our search retrieved more than 2,000 articles, reporting 133 replications performed on original studies. Nearly 70% of the replications were published in research articles, which were internal replications. The topics of software configuration management, software quality assurance, and software quality were the most frequent.

Investigations about replication of empirical studies in software engineering: A systematic mapping study

Cleyton V.C. de Magalhães *, Fabio Q.B. da Silva, Ronnie E.S. Santos, Marcos Suassuna
Centre for Informatics, Federal University of Paraná, Brazil

Keywords:

ARTICLE INFO

Article Notes:
Received 11 September 2012
Revised in revised form 28 January 2015
Accepted 2 February 2015
Available online xxxx

ARTICLE ABSTRACT

Context: Two recent mapping studies which were intended to verify the current state of replication of empirical studies in software engineering (8) identified two sets of studies: empirical studies actually reporting replications (published between 1994 and 2012) and a second group of studies that are concerned with definitions, classifications, processes, guidelines, and other research topics or themes about replication work in empirical software engineering research (published between 1998 and 2012). Objective: In this current article, our goal is to analyze and discuss the contents of the second set of studies about replications to increase our understanding of the current state of the work on replication.