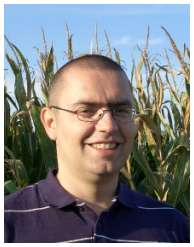
Views on Internal and External Validity in Software Engineering

Janet Siegmund

Norbert Siegmund



Sven Apel

TECHNISCHE UNIVERSITÄT CHEMNITZ

UNIVERSITÄT LEIPZIG

SE

SIC Saarland Informatics Campus

Let‘s start with a little experiment

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

An Experiment About Static and Dynamic Type Systems

Doubts About the Positive Impact of Static Type Systems on Development Time

Stefan Hanenberg
Institute for Computer Science and Business Information Systems
University of Duisburg-Essen
Schützenbahn 70, D-45117 Essen, Germany
stefan.hanenberg@icb.uni-due.de

Abstract

Although static type systems are an essential part in teaching and research in software engineering and computer science, there is hardly any knowledge about what the impact of static type systems on the development time or the resulting quality for a piece of software is. On the one hand there are authors that state that static type systems decrease an application's complexity and hence its development time (which means that the quality must be improved since developers have more time left in their projects). On the other hand there are authors that argue that static type systems increase development time (and hence decrease the code quality) since they restrict developers to express themselves in a desired way. This paper presents an empirical study with 49 subjects that studies the impact of a static type system for the

introduction of Generics in Java) or new programming languages are constructed that provide a new static type system.

In teaching, students are educated in the formal notation of static type systems as well as in proofs on static type systems (see for example [1, 19]). In industry, type systems become important for different reasons. Possibly, a programming language in use evolves by introducing a new static type system. If this new static type system should be applied, developers need to be educated, which causes additional costs. Maybe existing libraries or products should be adapted to match the new type system which also causes additional costs. Finally, additional tools might be required due to the new type system (such as tools that measure the current state of the software product) which potentially also cause additional costs.

Do Static Type Systems Improve the Maintainability of Software Systems? An Empirical Study

Sebastian Kleinschmager, Stefan Hanenberg
University of Duisburg-Essen
Essen, Germany
sebastian.kleinschmager@stud.uni-due.de
stefan.hanenberg@icb.uni-due.de

Romain Robbes, Éric Tanter
Computer Science Dept (DCC)
University of Chile, Chile
rrobbes@dcc.uchile.cl
etanter@dcc.uchile.cl

Andreas Stefik
Department of Computer Science
Southern Illinois University Edwardsville
Edwardsville, IL
astefik@siue.edu

***Abstract*—Static type systems play an essential role in contemporary programming languages. Despite their importance, whether static type systems influence human software development capabilities remains an open question. One frequently mentioned argument for static type systems is that they improve the maintainability of software systems—an often used claim for which there is no empirical evidence. This paper describes an experiment which tests whether static type systems improve the maintainability of software systems. The results show rigorous empirical evidence that static type are indeed beneficial to these activities, except for fixing semantic errors.**

I. Introduction

The debate regarding the pros and cons of static or dynamic type systems is ongoing in both academia and the software industry. While statically typed programming languages such as C, C++ and Java dominated the software market for many years, dynamically typed programming languages such as Ruby or JavaScript are increasingly gaining ground—especially in the domain of software development for the web. This paper contributes to this discussion with a controlled experiment (see [11], [23], [19], [16] for introductions on controlled ex
benefits o

How would you try to answer this research question?

TECHNISCHE UNIVERSITÄT CHEMNITZ

In the wild (realistic)

In the lab (controlled)

Maximize external validity

Reveals generally occurring effects

Causes of effects unclear

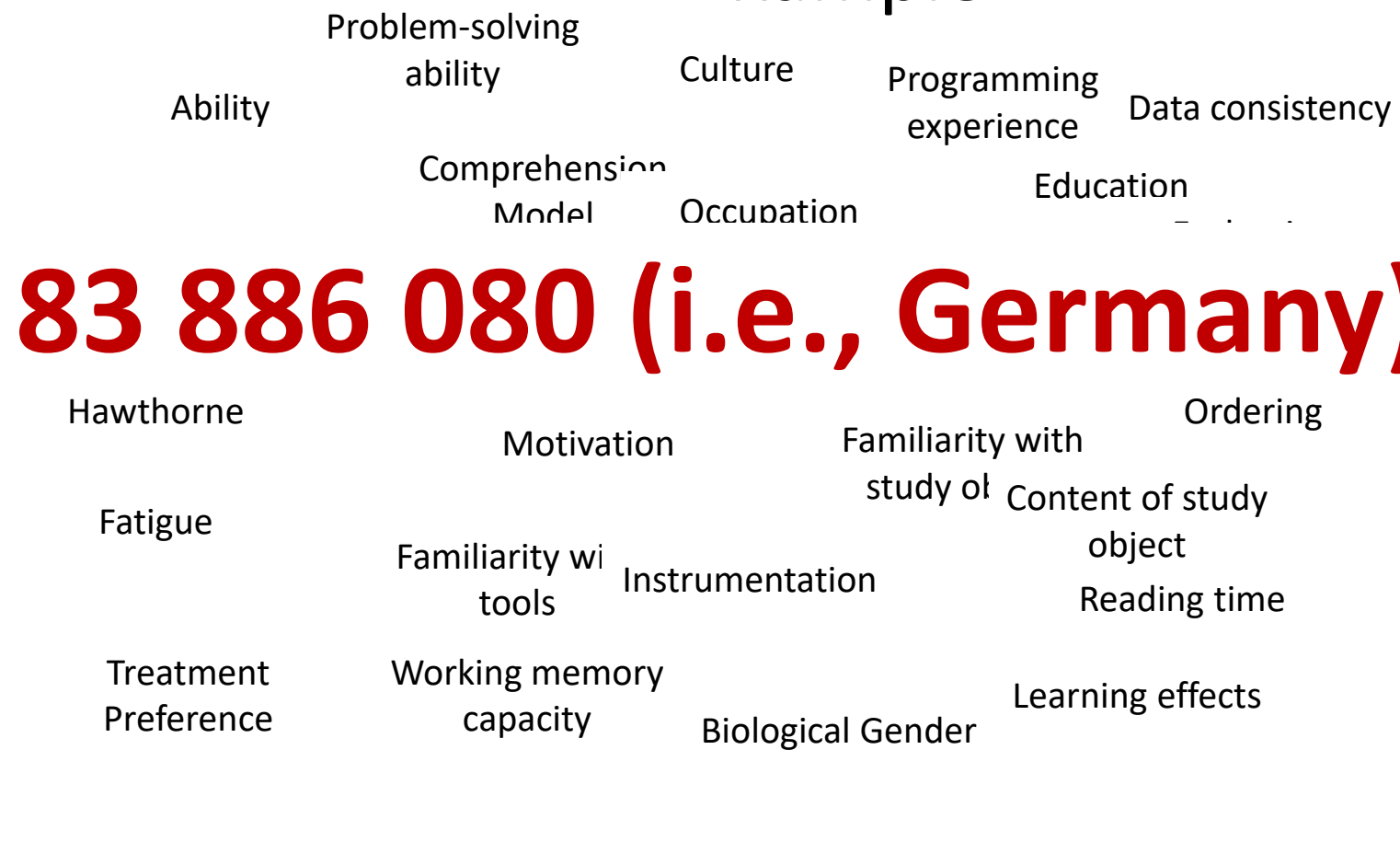
Maximize internal validity

Reliably explains the causes of effects

Hard to generalize

A fundamental tradeoff!

Example


Problem-solving ability
Culture
Ability
Programming experience
Data consistency
Comprehension
Model
Occupation
Education
83 886 080 (i.e., Germany)
Hawthorne
Motivation
Familiarity with study ob
Ordering
Fatigue
Content of study object
Familiarity wi tools
Instrumentation
Reading time
Treatment Preference
Working memory capacity
Biological Gender
Learning effects

Literature Survey

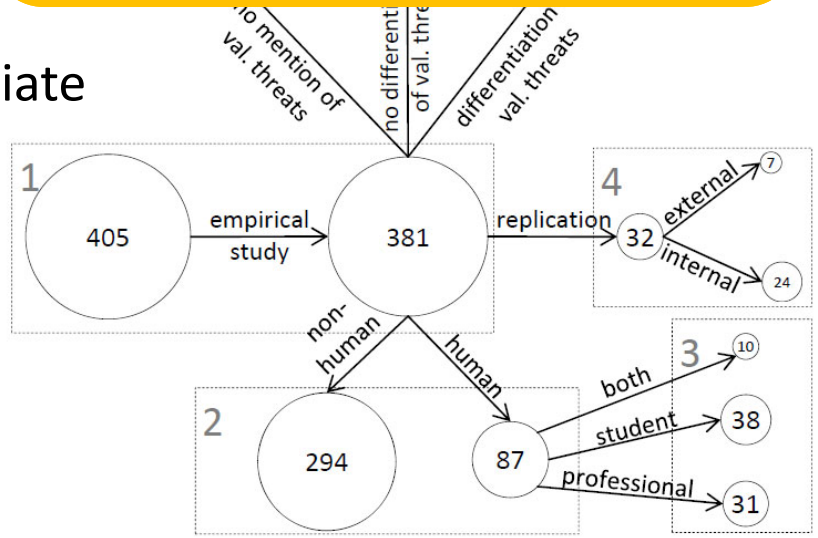
- Goal: get an overview of the **awareness** of and the **choices** regarding this tradeoff

- Data set: 405 full papers
  - ICSE (2012, 2013)
  - ESE/FSE (2011, 2012, 2013)
  - EMSE (2011, 2012, 2013)

Results of the Literature Survey

- 94% of the papers used an empirical metho
- 8% reported on a replication study
- 46% did not mention explicitly threats to va
- 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

RO	Questions	Answer options
1, 2	Which option would you prefer for an evaluation?	○ Max. internal validity, ○ Max. external validity
	[We asked this question two times, for human and non-human studies]	○ No preference
1	Would it be a reason to reject a paper that does not choose your favorite option?	○ Yes, ○ No
1, 2	In your opinion, what is the ideal way to address research questions like the one outlined above?	Open
1	Did you recommend to reject a paper in the past mainly for the following reasons?	□ Int. validity too low, □ Ext. validity too low
1, 2	For research questions like the one presented above (FP vs. OOP), do you prefer more practically relevant research or more theoretical (ground) research?	○ Applied, ○ Basic, ○ No preference
1	Have you changed how you judged a paper regarding internal and external validity?	○ Yes, ○ No
1, 3	What do you think about a reviewing format with several rounds, but with publication guarantees?	Open
1, 2	Do you have any suggestions on how empirical researchers can solve the dilemma of internal vs. external validity of empirical work in computer science?	Open
3	During your activity as a reviewer, how often have you reviewed a replicated study?	○ Never, ○ Sometimes, ○ Regularly
3	In general, how were the replications rated by you... by your fellow reviewers?	○ Accept, ○ Borderline, ○ Reject
3	During your activity as a reviewer, did you notice a change in the number of replicated studies?	○ Yes, increase, ○ Yes, decrease, ○ No
3	Do you think we need to publish more experimental replications in computer science?	○ Yes, ○ No
3	As a reviewer of a top-ranked conference, would you accept a paper that, as the main contribution,...	
	...exactly replicates a previously published experiment of *the same group*?	○ Yes, ○ No, ○ I do not know
	...exactly replicates a previously published experiment of *another group*?	○ Yes, ○ No, ○ I do not know
	...replicates a previously published experiment of the *same group*, but *increases external* validity?	○ Yes, ○ No, ○ I do not know
	...replicates a previously published experiment of *another group*, but *increases external* validity?	○ Yes, ○ No, ○ I do not know
	...replicates a previously published experiment of the *same group*, but increases *internal* validity?	○ Yes, ○ No, ○ I do not know
	...replicates a previously published experiment of *another group*, but *increases internal* validity?	○ Yes, ○ No, ○ I do not know

○ Maximize **internal** validity ○ Maximize **external** validity ○ …

Which option would you prefer for an evaluation?

□ **Internal** validity too low □ **External** validity too low

Did you recommend to reject a paper in the past mainly for the following reasons?

○ Yes ○ No

Do you think we need to publish more experimental replications?

TECHNISCHE UNIVERSITÄT CHEMNITZ

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

Awareness of community regarding tradeoff between external and internal validity

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

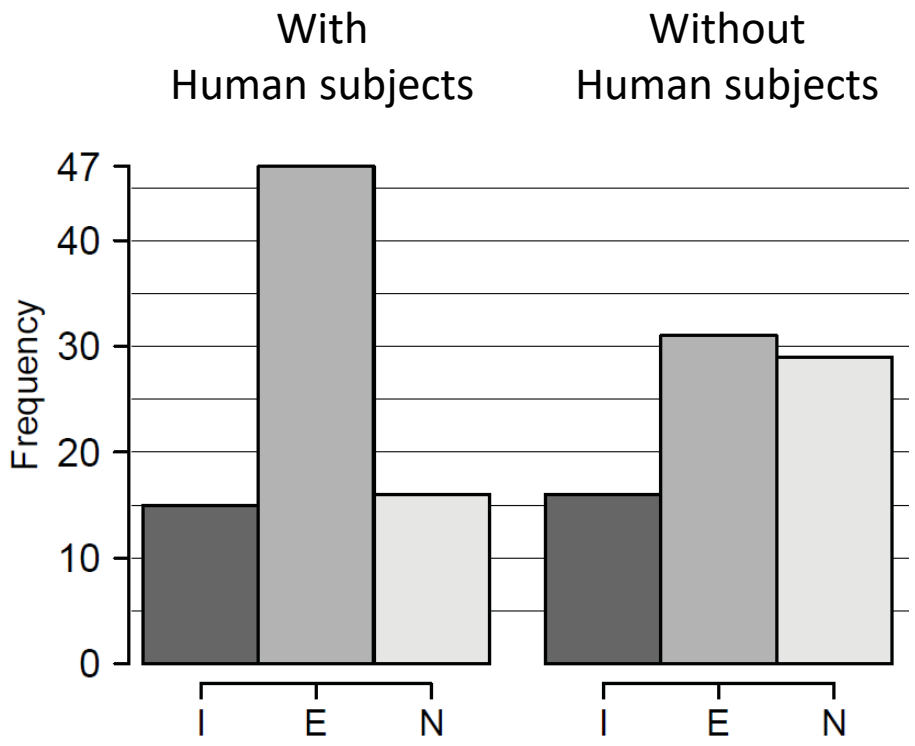
"Removing humans from the exercise reduces the challenges for internal validity. In that context, knowing how general the approach was would seem a more important issue to address."

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

Results RQ1

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

Which option would you prefer for an evaluation?



Internal

External

No preference

Results RQ2

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

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

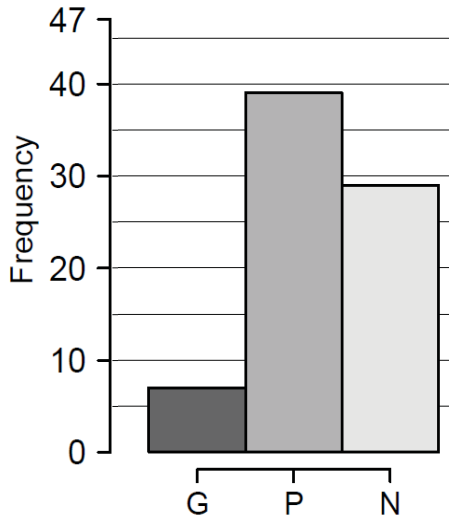
"I am worried that maximizing internal validity easily creates overly academic papers that provide little impact.[...]"

"[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'."

Results RQ2

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

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



Ground

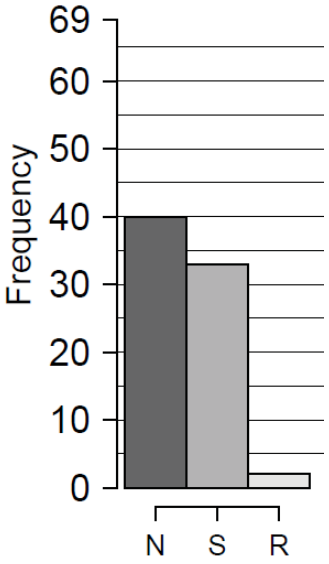
Practical

No preference

Results RQ3

Assess the opinion of the community regarding the role of replication

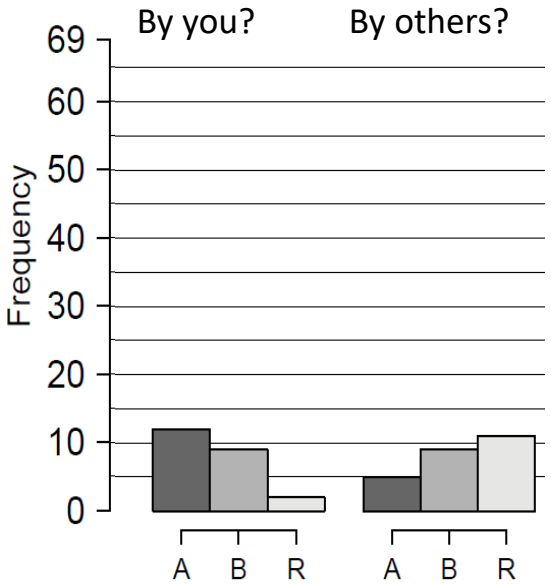
How often have you reviewed a replication?



Never **R**egularly

Sometimes

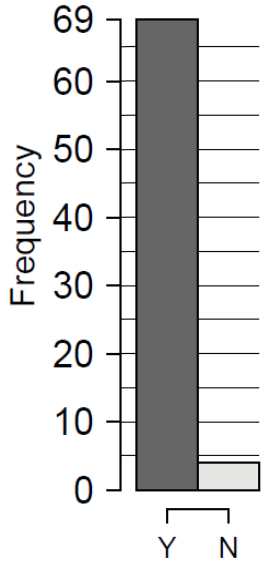
How were the replications rated?



Accept **R**eject

Boderline

Do we need more replications?



Yes **N**o

TECHNISCHE UNIVERSITÄT CHEMNITZ

Results RQ3

Assess the opinion of the community regarding the role of replication

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."

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

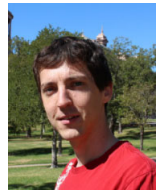
- Appreciate internally valid studies and don‘t confuse external validity with practicality
- Don‘t pay lip service to proper replications, but view them as an important piece of the puzzle
- Develop standards on how to assess (replication) studies

Authors

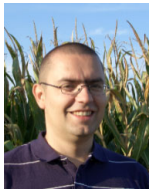
- Conduct multiple studies (internally and externally valid)
- Do not necessarily map an experiment 1:1 to a paper
- Report on validity issues and be concrete

Views on Internal and External Validity in Software Engineering

Janet Siegmund



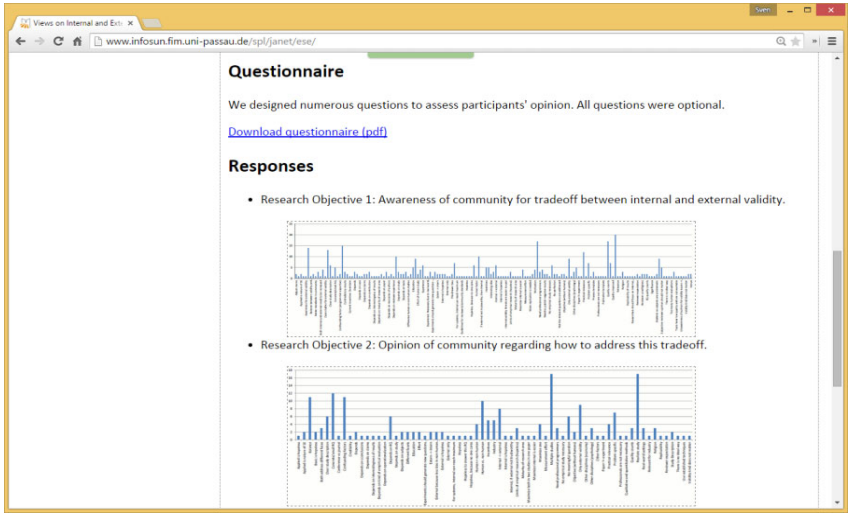
Norbert Siegmund



Sven Apel

See supplementary web site:

https://www.tu-chemnitz.de/informatik/ST/research/material/ese/



TECHNISCHE UNIVERSITÄT CHEMNITZ

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)
- …

Threats to External Validity

- Only small and specific sample of the SE community
- Do the results translate to other subcommunities (RE, MODELS, MSR, PLDI, …)?
- …

Studies on Replication

Empir Software Eng (2014) 19:501–557
DOI 10.1007/s10664-012-9227-7

Replication of empirical studies in software engineering research: a systematic mapping study

Fabio Q. B. da Silva • Marcos Suassuna • A. César C.
Tatiana B. Gouveia • Cleviton V. F. Monteiro • Igor Eb

Published online: 1 September 2012

Editor: Natalia Juristo

Abstract In this article, we present a systematic mappin
engineering. The goal is to plot the landscape of current p
studies in software engineering research. We applied the s
and select published articles, and to extract and synthesize
reported replications. Our search retrieved more than
selected 96 articles, reporting 133 replications performe
original studies. Nearly 70 % of the replications were publ
studies were internal replications. The topics of software
tion, and software quality concentrated over 55 % of the r
configuration management, and software tools and method

Information and Software Technology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

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 Pernambuco, Recife 50.740-560, Brazil

ARTICLE INFO

Article history:
Received 11 September 2014
Received in revised form 28 January 2015
Accepted 2 February 2015
Available online xxxx

Keywords:

ABSTRACT

Context: Two recent mapping studies which were intended to verify the current state of replication of empirical studies in Software Engineering (SE) 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 1996 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 in