The Future of Empirical Research in Software Engineering

Introduction

Page description:

ID 2

In software engineering, empirical studies have become more and more important, especially over the past few years. Empirical studies face several obstacles. In this questionnaire, we are interested in the **validity** of empirical studies. Typically, two kinds of validity are of primary concern: internal and external validity.

Internal validity is the degree to which the influence of confounding factors on the results are controlled. This allows experimenters to observe the results without bias. For example, when recruiting novice programmers, results are not biased by different levels of programming experience.

In contrast, **external validity** is the degree to which results of one experiment can be generalized. For example, when recruiting programmers with different levels of programming experience, according experimental results apply to these different levels of programming experience.

There is a trade-off between internal and external validity; only one at a time can be maximized. There are different ways to address this trade-off in empirical research, and we would like your thoughts on this.

Of course, we will anonymize your data.

Reviewer Activity

Page description:

ID 20

1. Since you are an expert in software engineering, we highly value your opinion. To help us better understand your answers, we would like to know for which conferences and journals you served as a reviewer (technical and research papers). We will anonymize your data.

	2014	2013	2012	2011	2010
ASE International Conference on Automated Software Engineering	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
EASE International Conference on Evaluation and Assessment in Software Engineering	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
ECOOP European Conference on Object-Oriented Programming	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
EMSE Empirical Software Engineering	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
ESEC/FSE European Software Engineering Conference/Symposium on the Foundations of Software Engineering	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
ESEM International Symposium on Empirical Software Engineering and Measurement	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
GPCE International Conference on Generative Programming: Concepts & Experiences	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
ICPC International Conference on Program Comprehension	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
ICSE International Conference on Software Engineering	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
ICSM International Conference on Software Maintenance	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
OOPSLA International Conference on Object-Oriented Programming, Systems, Languages, and Applications	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010
TOSEM ACM Transactions on Software Engineering and Methodology	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010

TSE IEEE Transactions on Software Engineering	☐ 2014	☐ 2013	☐ 2012	☐ 2011	☐ 2010

ID 34

2. Did we forget some important venue? Let us know which venue and in which years (startig from 2010) you reviewed papers:

Scenario

Page description:

ID 36

Consider the following scenario:

Suppose you are a reviewer of a submission in which the authors want to determine, based on an empirical study with human participants, whether functional or object-oriented programming (FP vs. OOP) is more comprehensible for programmers. So far, this question has never been addressed. Now, there are two options to design the study:

Option 1: maximize internal validity
The authors develop an artificial language that has a very similar design in the functional and object-oriented version. They leave out special features of existing languages (e.g., generics, classes, ...), recruit students as participants, use a stripped-down IDE, use artificial tasks to measure program comprehension, etc. In short, authors control the influence of all possible confounding factors.

Option 2: maximize external validity
Instead of creating an artificial set-up, the authors use existing languages, IDEs, and tasks to conduct the experiment. Authors recruit professional programmers as participants, use real projects from SourceForge, Eclipse as IDE, and let participants fix real bugs. In other words, authors create a practical, everyday setting.

ID 37

3. Which option would you prefer for an evaluation?

- Maximize internal validity
- Maximize external validity
- No preference

ID 38

Please, elaborate:

ID 39

4. Would it be a reason to reject a paper that does not choose your favorite option?

- Yes
- No

ID 41

Please, elaborate:

ID 42

5. In your opinion, what is the ideal way to address research questions like the one outlined above (FP vs. OOP)?

ID 44

Consider a different research question, not one in which human participants are observed, but, say, a **new approach that promises faster response times for database systems**.

Again, there are the two options for evaluation:
maximize internal (e.g., look at one database system in detail) or
maximize external (e.g., look at as many systems as possible, neglecting system-specific details) validity.

ID 45

6. Assuming that both options would be realized in the best possible way, which option would you prefer for evaluation like the one outlined above?

- Maximize internal validity
- Maximize external validity
- No preference

ID 46

Please, elaborate why you did or did not select a different option than for Question 3:

Research Direction

Page description:

ID 48

For the following questions, please look back on your activity as a reviewer.

ID 47

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

- [ ] Internal validity too low
- [ ] External validity too low

ID 49

Please, elaborate:

ID 51

8. For research questions like the one presented above (FP vs. OOP), do you prefer more practically relevant research or more theoretical (ground) research?

- Applied research (focus on practicability)
- Basic research (focus on sound scientific foundations)
- No preference

ID 52

Please, elaborate:

LOGIC Show/hide trigger exists.

ID 16

9. During your reviewer career, have you changed how you judged a paper regarding internal and external validity?

- Yes
- No

ID 53

Please, elaborate:

LOGIC Dynamically shown if "During your reviewer career, have you changed how you judged a paper regarding internal and external validity?" = Yes

ID 18

Please, specify:

- [ ] Yes, I now appreciate papers with high **internal** validity more.
- [ ] Yes, I now appreciate papers with high **external** validity more.

Validity

Page description:

ID 56

The following questions are related to the representation of empirical research in the software-engineering literature.

ID 67

10. In your opinion, do you think that in the literature, empirical evaluations with human participants **are needed more or less often**?

Considerably less often	Less often	Fine as is	More	Considerably more	I do not know
○	○	○	○	○	○

Considerably less often
Less often
Fine as is
More
Considerably more
I do not know

ID 68

11. In your opinion, do you think that in the literature, empirical evaluations with human participants **are accepted/rejected too often**?

Considerably too often rejected	Too often rejected	Fine as is	Too often accepted	Considerably too often accepted	I do not know
Considerably too often rejected	Too often rejected	Fine as is	Too often accepted	Considerably too often accepted	I do not know

ID 69

12. In your opinion, do you think that in the literature, empirical evaluations with human participants **need higher internal/external validity**?

Considerably higher internal validity	Higher internal validity	Fine as is	Higher external validity	Considerably higher external validity	I do not know
Considerably higher internal validity	Higher internal validity	Fine as is	Higher external validity	Considerably higher external validity	I do not know

ID 70

Please, elaborate

ID 71

13. In your opinion, do you think experiments without human participants (e.g., performance evaluation, code measurement) **are needed more or less often**?

Considerably less often	Less often	Fine as is	More	Considerably more	I do not know
○	○	○	○	○	○
Considerably less often	Less often	Fine as is	More	Considerably more	I do not know

ID 73

14. In your opinion, do you think experiments without human participants (e.g., performance evaluation, code measurement) **are accepted/rejected too often**?

Considerably too often rejected	Too often rejected	Fine as is	Too often accepted	Considerably too often accepted	I do not know
○	○	○	○	○	○
Considerably too often rejected	Too often rejected	Fine as is	Too often accepted	Considerably too often accepted	I do not know

ID 74

15. In your opinion, do you think experiments without human participants (e.g., performance evaluation, code measurement) **need higher internal/external validity**?

Considerably higher internal validity	Higher internal validity	Fine as is	Higher external validity	Considerably higher external validity	I do not know
○	○	○	○	○	○
Considerably higher internal validity	Higher internal validity	Fine as is	Higher external validity	Considerably higher external validity	I do not know

ID 77
Please, elaborate

Replication

Page description:

ID 78

To increase validity of empirical studies, researchers replicate experiments. That is, the same or other researchers conduct the experiment again, either exactly as it took place, or with some modifications.

LOGIC Show/hide trigger exists.
ID 79
16. During your activity as a reviewer, how often have you reviewed a replicated study?

- Never
- Sometimes
- Regularly

ID 88
Please, elaborate:

LOGIC Dynamically shown if "During your activity as a reviewer, how often have you reviewed a replicated study?" = Sometimes or "During your activity as a reviewer, how often have you reviewed a replicated study?" = Regularly

ID 85

17. In general, how were the replications rated...

	Accept	Boderline	Reject	Not applicable
...by you?	○Accept	○Boderline	○Reject	○Not applicable
...by your fellow reviewers?	○Accept	○Boderline	○Reject	○Not applicable

LOGIC Dynamically shown if "During your activity as a reviewer, how often have you reviewed a replicated study?" = Sometimes or "During your activity as a reviewer, how often have you reviewed a replicated study?" = Regularly

ID 80

Please, elaborate:

ID 81

18. During your activity as a reviewer, did you notice a change in the number of replicated studies?

- ○ Yes, it increased.
- ○ Yes, it decreased.
- ○ No

ID 82

Please, elaborate:

ID 83

19. Do you think we need to publish more experimental replications in computer science?

- Yes
- No

ID 96

Please, elaborate:

ID 89

20. As a reviewer of a top-ranked conference, would you accept a paper that, as the main contribution,... (assuming authors realized it in the best possible way)

	Yes	No	I do not know
...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 (e.g., by recruiting expert programmers or using another programming language)?	Yes	No	I do not know
...replicates a previously published experiment of **another group**, but **increases external** validity (e.g., by recruiting expert programmers or using another programming language)?	Yes	No	I do not know
...replicates a previously published experiment of the **same group**, but **increases internal** validity (e.g., by recruiting expert programmers or using another programming language)?	Yes	No	I do not know
...replicates a previously published experiment of **another group**, but **increases internal** validity (e.g., by recruiting expert programmers or using another programming language)?	Yes	No	I do not know

ID 84

Please, elaborate:

Concluding Remarks

Page description:

ID 98

A huge problem for authors is that empirical evaluations require a high effort in terms of time and cost (e.g., recruiting participants, designing tasks, selecting/designing languages). Since the outcome of an experiment is not clear and might be biased (e.g., due to deviations), the effort is at high risk.

ID 99

21. What do you think about a reviewing format with several rounds, but with publication guarantees? That is, the paper is guaranteed to be published (independent of the results), if the authors conduct a further, sound empirical evaluation that improves either internal or external validity.

ID 100

22. Do you have any suggestions on how empirical researchers can solve the dilemma of internal vs. external validity of empirical work in computer science?

ID 101

23. There are several factors that influence how researchers balance internal and external validity, such as maturity of research area, availability of experimental data, or effort of recruiting and preparing participants/subject systems. In your opinion, what are possible influencing factors for balancing internal and external validity?

ID 102

24. Do you have any additional comments to this survey, questionnaire, or empirical research in general?

ID 103

Thank you very much for your time. If you have any questions regarding this survey, please contact:

Janet Siegmund: siegmunj@fim.uni-passau.de
Norbert Siegmund: siegmunn@fim.uni-passau.de
Sven Apel: apel@fim.uni-passau.de

Thank You!

ID 1

Thank you for taking our survey. Your response is very important to us.