Peer Community In Registered Reports

Bug detection in software engineering: which incentives work best?

*A recommendation by **Chris Chambers** based on peer reviews by **Edson OliveiraJr** of the STAGE 2 REPORT:*

Dmitri Bershadskyy, Jacob Krüger, Gül Çalıklı, Siegmar Otto, Sarah Zabel, Jannik Greif, Robert Heyer (2024) A Laboratory Experiment on Using Different Financial-Incentivization Schemes in Software-Engineering Experimentation. arXiv, ver. 8, peer-reviewed and recommended by Peer Community in Registered Reports.
https://arxiv.org/pdf/2202.10985

Submitted: 25 March 2024, Recommended: 11 September 2024

Cite this recommendation as:
Chambers, C. (2024) Bug detection in software engineering: which incentives work best?. *Peer Community in Registered Reports*, 100746. 10.24072/pci.rr.100746

Published: 11 September 2024

Bug detection is central to software engineering, but what motivates programmers to perform as optimally as possible? Despite a long history of economic experiments on incentivisation, there is surprisingly little research on how different incentives shape software engineering performance. In the current study, Krüger et al. (2024) undertook an experiment to evaluate how the pay-off functions associated with different financial incentives influence the performance of participants in identifying bugs during code review. The authors hypothesised that performance-based incentivisation would result in higher average performance, as defined using the F1-score, and that different incentivisation schemes may also differ in their effectiveness. The results did not support the preregistered predictions, with no statistically significant differences in F1-score observed between groups that received performance-based incentives compared to a control group that received no incentive. Exploratory analyses suggested some potential trends of interest, but the main implication of this work is methodological: that experiments in this field require substantially larger sample sizes to provide definitive tests. The current work is valuable in providing a novel unbiased insight on the magnitude of this challenge, which is now primed for further investigation. The Stage 2 manuscript was evaluated over one round of in-depth review. Based on detailed responses to the recommender and reviewer's comments, the recommender judged that the manuscript met the Stage 2 criteria and awarded a positive recommendation. **URL to the preregistered Stage 1 protocol:** https://osf.io/s36c2 **Level of bias control achieved:** Level 6. *No part of the data or evidence that was used to answer the research question was generated until after IPA.* **List of eligible PCI RR-friendly journals:**

- Peer Community Journal

- PeerJ Computer Science
- Royal Society Open Science

References:

Bershadskyy, D., Krüger, J., Çalıklı, G., Siegmar, O., Zabel, S., Greif, J. and Heyer, R. (2024). A Laboratory Experiment on Using Different Financial-Incentivization Schemes in Software-Engineering Experimentation. Acceptance of Version 8 by Peer Community in Registered Reports. https://arxiv.org/pdf/2202.10985

Reviews

Evaluation round #1

DOI or URL of the preprint: https://arxiv.org/pdf/2202.10985.pdf
Version of the preprint: v6

Authors' reply, 13 August 2024

Download author's reply
Download tracked changes file

Decision by Chris Chambers, posted 31 May 2024, validated 31 May 2024

Minor Revision

Thanks for your patience during this Stage 2 evaluation. I was able to obtain a review of your manuscript from one of the experts who reviewed the Stage 1 proposal. The good news is that the reviewer is very happy with the completed study and recommends acceptance. Based on this reviewer's assessment and my own close reading, I am therefore issuing an interim revision decision. In my own reading I noticed a few areas that would be benefit from minor clarification.

- Abstract. Re the sentence: "Due to the small sample sizes, our results are not statistically significant, but we can still observe clear tendencies." Statistically non-significant results can arise either because the null hypothesis is true or the test is insensitive. When relying on null hypothesis significance testing, we cannot know for certain which is the case (at least not without adding frequentist equivalence testing or Bayesian hypothesis testing), therefore I would like you to consider replacing the sentence above with a less deterministic statement about the potential reasons for non-significance: "Our results are not statistically significant, possibly due to small sample sizes and consequent lack of statistical power, but with some notable trends [that may inspire future hypothesis generation]." (the section in square brackets is a stylistic addition that you are free to omit, but I replaced "clear" with "notable" because non-significant trends are by definition unclear and it is a potential source of interpretative bias to overstate their importance.
- Table 1 (p4). Please add a column to the far right of this table called "observed outcome" that briefly summarises the results and, in particular, states the degree of support for each hypothesis (H1, H2) in the second section (i.e. supported, not supported, with the statement based strictly on the outcomes of the preregistered analyses rather than any additional exploratory analyses). To make room for this

table, I suggest moving the content in the "disproved theory" column to the Table caption (since it applies generally to all aspects of the study), and then this column can be removed from the table to make space for an "observed outcome" column.

- p15: Explain padjusted in a footnote the first time it is used. I believe it is simply the alpha-corrected value following Holm-Bonferroni correction (?), which case padjusted should actually be reported as >.99 rather than =1 because a p value can never equal exactly 1.
- p15: replace "insignificantly" with "non-significantly" consistent with standard statistical parlence.

Once these issues are addressed, I will issue final Stage 2 acceptance without further peer review. In anticipation of the next version being the final preprint, once you have made these revisions please update the latest version of the preprint to be a **clean** version (rather than tracked changes), but upload a tracked-changes version (that shows only these latest revisions) in the PCI RR system when you resubmit.

Reviewed by Edson OliveiraJr, 16 May 2024

Congratulations on the submitted Stage 2 manuscript rigor and fidelity on the approved RR. I answered the recommended questions regarding such a submission as follows.

==========

Have the authors provided a direct URL to the approved protocol in the Stage 2 manuscript? Did they stay true to their protocol? Are any deviations from protocol clearly justified and fully documented?

R.: The authors provided the DOI of the arXiv, which goes directly to the Stage 2 manuscript rather than the exact URL of the RR, which is https://arxiv.org/abs/2202.10985v4, to avoid misunderstanding by the readers.

==========

Is the Introduction in the Stage 1 manuscript (including hypotheses) the same as in the Stage 2 manuscript? Are any changes transparently flagged?

R.: Both Stage 1 and Stage 2 introduction sections are the same, with no relevant changes.

==========

Did any prespecified data quality checks, positive controls, or tests of intervention fidelity succeed?

R.: The pre-specified analysis was successful as part of the performed survey and experiment.

==========

Are any additional post hoc analyses justified, performed appropriately, and clearly distinguished from the preregistered analyses? Are the conclusions appropriately centered on the outcomes of the preregistered analyses?

R.: Yes. There was an additional iteration of the survey with 8 participants to achieve the stipulated sample size. Another subtle change was merging MAIT and MPIT metrics, which were identical, but this was previously expected in the survey RR design. Regarding the survey, there was one change reflected by the merging metrics. Therefore, the experimental design changed from a 4x1 to a 3x1 experiment, thus modifying the treatment function. No other deviation was detected.

==========

Are the overall conclusions based on the evidence?

R.: The conclusion is solid and based on the evidence provided by the survey and experiment analysis.