International Journal on Software Tools for Technology Transfer (2025) 27:411–414
https://doi.org/10.1007/s10009-025-00808-y

COMPETITIONS AND CHALLENGES

Special Issue: RRRR 2022

Replicable theory

Benjamin Lucien Kaminski[1,2]

Accepted: 29 April 2025 / Published online: 13 May 2025

Abstract
In addition to the difficulties with replicating experiments or systems from some given theoretical description, we discuss *the possibility that already the theory itself is poorly replicable*. After explaining what we understand by theory replicability, we propose to scientifically evaluate whether or not the broader field of *Logic, Semantics, and Verification in Computer Science* suffers from systematic theory replicability problems.

Keywords Replicability · Theory · Intuition

If you can't explain it simply,
you don't understand it
well enough.
Albert Einstein [9]
(misattributed and misquoted)

1 Introduction

Replicability is a cornerstone of the scientific method. Even theory-driven fields like theoretical physics ultimately depend on replicability: Trust in a theory is time and again established by conducting many different experiments that repeatedly confirm the theory with ever increasing precision. To conduct experiments that faithfully confirm a theory, the theory has to be well understood by the experiment designers.

In many experiment-driven sciences, replication crises have been discussed, meaning that it has been found that many scientific findings are difficult or impossible to replicate [8]. Recently a possible replication crisis has also been brought forward in a more formal science, namely mathematics [2]. In this article, we want to discuss replicability of theory in computer science. We first explain what we understand by replicability of a theory, showcase possibly detrimental situations arising from poor replicability, and finally propose to conduct an experimental evaluation of how replicable the findings in theoretical computer science are.

✉ B.L. Kaminski
kaminski@cs.uni-saarland.de

[1] Saarland University, Saarland Informatics Campus, Saarbrücken, Germany

[2] University College London, London, UK

2 Reproduction and replication of theories

At least in my own case, understanding mathematics
doesn't come from reading or even listening.
It comes from rethinking what I see or hear.
I must redo the mathematics in the context
of my particular background. [. . .]
When I have reorganized the mathematics in my own
terms, then I feel an understanding,
not before. Stephen Smale [6]

According to the Stanford Encyclopedia of Philosophy, we can distinguish between *reproducibility* and *replicability* as follows [4]:

> *"Reproducibility is the reproducibility of an experiment, given a fixed theoretical description.* [. . .] *Replicability* [. . .] *is where experimental procedures differ to produce the same experimental result."*

In this short paper, however, we do not consider reproducibility or replicability of *experiments* but rather that of the *theoretical descriptions*. And so we ask:

When is a theoretical description
reproducible or replicable?

Untrained in philosophy, we make the following (perhaps simplistic) argument: Consider a theoretical description which is printed on paper. To *reproduce* that description, one could *make a copy* of the paper. This process indeed (re)produces the same description and can be performed *entirely without any understanding* of the theory. We would thus argue that *any theoretical description is reproducible* in practice.

Replicating the theoretical description, on the other hand, is something else entirely. For a true replication that agrees

with the Stanford "definition", we argue that one would need to *read* the description, *understand* the description and *develop an intuition* for the theory, *rethink* the theory, indeed almost *reinvent* the theory, and then *formulate a new description of the same theory*, but in one's own words. Revisiting the epigraph of this section, we coin this procedure *Smalian theory replication*. With Smalian theory replication, the *cognitive procedures to produce an equivalent*[1] *description of the same theory will differ*, if only because the persons conducting the rethinking may have very different scientific backgrounds.

Adopting the Smalian notion of what theory replication constitutes, we claim that *not* every description of a theory is replicable, even if the theory and its description are sound. Indeed, we believe that the *degree of replicability* varies greatly, and can and should be considered *a* suitable and important criterion for judging the description's quality.

3 Ramifications of poor replicability

In this section, we present two pieces of evidence that poor comprehensibility of theoretical contributions has had detrimental consequences.

Inter-universal Teichmüller theory Our first piece of evidence is the notable example of Shinichi Mochizuki's *Inter-universal Teichmüller Theory* (IUT) [5]. Its most striking application would be to provide a proof for many outstanding conjectures in number theory, most centrally the *abc conjecture*. Alas, Mochizuki's theory is considered incomprehensible widely across the mathematical community and thus *abc* remains a conjecture to most mathematicians [7]. Still, as the implications of IUT would be so profound, many mathematicians, among them at least one Fields medalist, have spent significant time trying to understand the theory. The total amount of time dedicated to understanding IUT is estimated to have already exceeded 30 researcher years [2], and efforts continue to this day.

The BITA conference The following is anecdotal evidence based on true events; all names were anonymized.[2]

Alice served on the program committee of BITA and she was assigned to review a paper about progress on the ALAM framework. The theoretical development in this paper was mostly inaccessible to Alice and there was, by her judgement, no way she could have replicated this paper, not even within an unreasonable amount of time. It emerged from the PC discussion that the paper was also rather inaccessible to the other reviewers. The reviewers agreed that it would need an ALAM expert to properly judge this paper. But finding an expert reviewer proved to be virtually impossible, because all ALAM experts ended up being conflicted with one another, and in particular conflicted with the authors.

How should the reviewers have decided in such a situation? Accept the paper in the spirit of *"didn't fully understand, but looks fine to me"*? Or reject the paper in the spirit of *"I'll reject whatever I don't understand"*?

This whole predicament could have been mitigated, had the paper been accessible to the broad BITA audience, not solely to ALAM experts. What is more: Were the paper accessible to the whole BITA audience, then

(1) non-ALAM experts could still properly and fairly judge the paper, even if only from the perspective of an ALAM outsider, and
(2) – much more importantly – once the paper is published (be it at BITA or elsewhere), more people have access to the knowledge that the ALAM-paper authors produced.

Presuming that ALAM is any good, more people being able to replicate the ALAM theory will likely increase the number of ALAM experts over time, which would ultimately benefit the ALAM *and* the BITA community.

4 Proposed experimental evaluation

True to the motto *"The first step in solving any problem is recognizing there is one"*, we propose to experimentally evaluate whether research in theoretical computer science, in particular in the field of *logic, semantics, and verification*, suffers from poor replicability or not. Such an experiment could be conducted with the program committee (PC) members of a major theory-driven, yet broad, conference like CAV, CSL, ESOP, FoSSaCS, ICALP, LICS, OOPSLA, POPL, or TACAS, to name only a few. If anyone, the PC members should be considered experts in the respective field and furthermore they should more or less resemble the breadth of the audience of the respective conference.

Experiments on PCs are not unprecedented. In 2014, the program chairs of NeurIPS, a top-tier conference in machine learning, conducted an experiment on their PC members [3]: About 10% of the 1,678 submissions to NeurIPS 2014 were randomly selected to be reviewed by *two* independent PCs. A particularly striking outcome of that experiment was that, regarding which papers to accept and which not, the two PCs were only in agreement for about half of the papers. The experiment was repeated for NeurIPS 2021 [1].

Experiment design sketch Our experiment on theory replicability could look roughly as follows: We randomly

[1] It is of course virtually impossible that two persons replicating a theory would arrive at exactly the *same* wording.

[2] Alice's gender, the conference name, and the framework name were randomly chosen/generated using `random.org`.

select a chunk of N papers from the list of *accepted papers* at conference ABC and ask PC members to evaluate the replicability (or rather the *potential for replicability*, according to their assessment) of all N papers. A questionnaire for replicability assessment could be along the following lines:

1. How would you rate your expertise on the presented theoretical contribution?
2. How well did you understand the theoretical contribution?
   - If rather well, how many hours did you need to understand the material?
   - If not so well, how far did you get? (pages, percentage, etc.)
3. Do you feel confident that you could reformulate/ replicate the theoretical contributions (at least the key results) in your own words?
   - How many hours would you expect to need for the replication?
4. What level of expertise do you believe is required to perform such replication?

It is also conceivable to ask authors of accepted papers to create short questionnaires about the key points of their theoretical contributions and then rate how well the PC members actually understood the papers. Going further, one could even perform actual replication experiments where PC members or other external trial participants try to replicate the theory and (other) PC members evaluate the replications.[3] Indeed, this would come closer to a faithful empirical evaluation of the replicability of a theoretical contribution.

Obviously, the ideas above are just *rough sketches*, nowhere close to a full-fledged study design. Our intent is to demonstrate along which lines such a study could be designed and what it is supposed to evaluate.

It is also important to note that such a study should not be about praising or shaming the replicability of individual papers. The results would have to be appropriately anonymized, but it should be possible to draw conclusions like:

At least X% of the PC members
(think they) are able to replicate at least Y%
of the papers accepted at ABC.

5 Towards replicable theory

In more experimentally-driven research in the broader scope of logic, semantics, and verification, awareness for replicability is increasingly finding its way into the mainstream through *artifact evaluations*, e.g., at conferences like CAV, ESOP, OOPSLA, PLDI, POPL, or TACAS. For more theoretical contributions, replicability is certainly more difficult to assess and *"presentation"* is often already a (perhaps too secondary) evaluation criterion. But *theory replicability in the Smalian sense* is at least a more tangible — and perhaps a more purposeful — criterion than *"good presentation."* We believe that if theory replicability became a *core evaluation criterion* in the reviewing process, theoretical contributions would become more replicable (thus increasing in quality) on a broader scale, from which our community could only benefit.

A first step, however, would be to find out whether or not our field suffers from poor theory replicability; and if so, to which degree.

6 Addendum

Should our scientific community, perhaps after conducting above-mentioned experiments, arrive at the consensus that we indeed suffer from a systematic (even if latent) replicability problem, we should investigate root causes and mitigation strategies. The following aspects which might affect paper accessibility and replicability come to mind[4] (in no particular order):

Page limits Are the customary page limits sufficient? Do we need more pages? Fewer pages? Should we have page limits at all? Allowing for more pages might render papers more accessible to non-experts. On the other hand, a slight nudge towards brevity often even increases quality of exposition.

"Publish or perish!" and the associated ever-growing number of publications. Would the average quality of exposition per paper increase if there was less pressure to publish as quickly (and as much) as possible?

Size & structure of program committees For evaluating theory replicability seriously during the review process of conferences, do we need larger or differently structured program committees? Do we perhaps also need another layer of hierarchy in program committees like *area chairs* or similar? Do we need a lot more lower level reviewers, who are assigned fewer papers (if more than one at all) but can review them in much more detail? If yes, how do we scale up to the necessary number of reviewers? Does raising the barrier for asking subreviewers[5] affect review quality? For the better or

[3] This was suggested by Reviewer 1 of an earlier version of this paper.

[4] In fact, not to the mind of the author of this paper but to the mind of Reviewer 3 of an earlier version of this paper. We present those aspects here in our own words and sometimes add our own thoughts on these aspects.

[5] As is currently becoming more and more customary in programming languages conferences as a result of introducing double blind reviewing.

for the worse? Does it affect the replicability and quality of the ultimately published papers?

Desired degree of replicability How much needs to be understood and replicated? Do we really need 100% understanding? Do we really need 100% replicability for a paper to be worth publishing?[6]

Reviewer & author guidelines What are good reviewer guidelines that in effect gear papers towards theory replicability? Should our aim be to maximize the number of expert or knowledgeable reviewers per paper? Should we instead purposefully have each paper be reviewed by reviewers of *varying* expertise? Should we limit the time that is required or expected to understand a paper, depending on the level of expertise?

What are good author guidelines for gearing papers towards theory replicability? What can be done to help authors make their papers more accessible?

Acknowledgements The author would like to sincerely thank the anonymous reviewers of an earlier version of this paper, as well as the participants of the 1st *Workshop on Reproducibility and Replication of Research Results,* for their valuable and constructive feedback and for treating this – admittedly different – contribution with genuine interest and a truly open mind.

Funding information Open Access funding enabled and organized by Projekt DEAL.

References

1. Beygelzimer, A., Dauphin, Y., Percy, L., Wortman Vaughan, J.: The NeurIPS 2021 Consistency Experiment (2021). https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/ [Accessed online 24 February 2022]
2. Bordg, A.: A replication crisis in mathematics? Math. Intell. **43**(4), 48–52 (2021)
3. Cortes, C., Lawrence, N.D.: Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. CoRR (2021). arXiv:2109.09774

[6] In the unsurprising opinion of the author of this paper, a scientific paper that is not replicable also need not be published in its current state.

4. Fidler, F., Wilcox, J.: Reproducibility of scientific results. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Summer Metaphysics Research Lab, 2021th edn. Stanford University Press, Stanford (2021)
5. Mochizuki, S.: A Panoramic Overview of Inter-Universal Teichmüller Theory (2014)
6. Smale, S.: Finding a Horseshoe on the Beaches of Rio. Math. Intell. **20**(1), 39–44 (1998)
7. Wikipedia: Inter-universal Teichmüller theory (2022). [Accessed online 23 February 2022]
8. Wikipedia: Replication crisis (2022). [Accessed online 24 February 2022]
9. Wikiquote: Albert Einstein (2022). [Accessed online 23 February 2022]