What is Worth Replicating? - SIPS 2020 Unconference session

Peder M. Isager & Anna van 't Veer

2020-06-30

Given that most research is original and we have limited resources available for replication, we need guidelines for study selection in replication research. But what makes a study worth replicating? In our unconference session at SIPS this year, we had the pleasure of discussing this problem with fifty-some enthusiastic scholars. Here is a quick summary of what they taught us.

First off, we went into this session with our own prior beliefs about what makes a study worth replicating. Namely, we strongly believe that studies are worth replicating (1) if their results are uncertain, (2) if we can reduce uncertainty about the results by replicating, and (3) the results are worth becoming less uncertain about.

In part, this session was an opportunity for us to see if our beliefs are shared by others. In part, it gave us an opportunity to fish for blind spots in our own beliefs; perhaps there are important factors for determining what is worth replicating that we have not thought of?

To get the discussion going, we asked participants to focus on one of three questions:

1) Which specific studies/claims in your field of interest do you believe currently are the most in need of replication? Why these?
2) What criteria (general or specific to your field of interest) would you use to determine which studies in your field to spend resources replicating?
3) Have you ever conducted a replication? If so, how did you go about selecting a study to replicate? Did you need to select one of multiple candidates? If so, how did you choose among them?

Factors mentioned that jive with our prior beliefs:

The discussion was mainly focused on general factors that contribute to determine replication worthiness (question 2). From the notes and summaries provided by the groups at the end of the session, we find that many of the factors that people brought up are related to the framework we had in mind going into the session. I.e. the uncertainty about existing knowledge, the value of having knowledge, and the possibility of gaining knowledge through replication, all seem to be factors that people brought up as important. It is also abundantly clear from the discussion that these factors are complicated, and cannot easily be tied down to any specific set of operationalizations. Here are just some of the factors that people brought up as potentially important for considering if something is worth replicating (*italicized indented bullet points indicate our post-session replies to comments*):

Uncertain about existing knowledge

- Distrust in how the original study was run
- Statistical uncertainty (e.g. as measured by confidence intervals)
- Distrust in the researcher/lab that conducted the study
- Generalizability of results
- Replicability

- Small sample size/low statistical power
- Suspicious data/prevalence of p-hacking
- Results only replicable in a single lab
- Methodological details missing from the report
- "Fragile" findings
- HARKed conclusions
- Multiple explanations for the finding exist
  - *Reply: Here I think it is worth considering if a replication can actually help us mitigate the problem. If a study design leaves open multiple explanations for the results, and we simply replicate the design with more data, we may still not be able to separate between different explanations when we see our results. Thus, this kind of uncertainty might actually call for novel study designs rather than replication (e.g., in an adversarial collaboration).*
- Degree of available evidence supporting the phenomenon
- Certainty in result

Our ability to reduce uncertainty

- Feasibility
- Expense/Resource cost
- Availability of materials
- Replication estimates the efficacy and generalisability of specific stimuli
- Replication estimates the degree to which an effect exists 'in the wild'
- Availability of stimuli, resources, full explanation of methods, etc.
  - *Reply: Although most replicators will of course strive for a rigorous replication, some replication studies might not add much in terms of reducing uncertainty (e.g. if no resources are available to achieve adequate power). It seems most of these bullets assume the replication itself is 'good' but we should stay critical there as well. For selecting a study to replicate, the question of how much uncertainty the replication can reduce may for instance be raised when phase 1 reviewers judge a replication proposal.*

Value of becoming certain

- Scientific impact
- "Real world" consequences
- Number of citations
  - *Reply: However, as many pointed out, citation count is not a perfect indicator of impact, and some citations have little to do with the impact of the paper on a field.*
- Impactfulness
- Interesting study
- Study design one intends to extend
- Societal impact. (e.g. on public consumption, policy, interest/awareness/attention in media)
- Number of scientists working on the topic? To avoid lots of wasted resources.
- Number of people who could benefit from the knowledge.

- International impact.
- Potential benefit/harm for vulnerable populations.
- Relevance (personal/theoretical/political)

There was also a group that discussed study selection strategies in past replication research (question 3). The criteria mentioned by this group as having been important for actual study selection decisions largely overlapped with the factors listed above.

Factors mentioned that do not obviously jive with our prior beliefs.

It is always useful to understand when a normative (what should be replicated) framework does not coincide with descriptive reality (what is replicated), or when different normative frameworks collide. Here are a few examples of factors people brought up, that do not obviously fit within our three-category system for deciding what is worth replicating:

- Focus on replicating old studies vs new studies
  - *Reply: It is perhaps reasonable to assume that age of the study is correlated with the importance of the study for a field of research. Older studies have been in circulation for longer, and have had more time to become embedded in the research canon. However, some study designs and results become outdated with age (e.g. through improved study designs, or by having conclusions falsified by subsequent research). Thus, you could also argue that more recent studies are more likely to be influencing current theorizing and should be prioritized for replication. Both are perhaps equally sound conclusions in different scenarios, which would entail that there is no straight-forward relationship between age and replication worthiness of a study.*
- Finding is surprising/violates common sense
  - *Reply: It is not completely clear to me whether this factor is mentioned because it entails uncertainty, because findings that clash with our intuitions are more valuable, or for some other reason. I agree that it is likely important, and certainly has been the motivation for several actual replication efforts. It would be interesting to dig further into why/when a surprising finding ought to be more worth replicating than an unsurprising one.*
- Surveying the 'health' of a discipline as a whole
  - *Reply: A good example of a replication project with this goal in mind is the Reproducibility Project Psychology. In this project, even though authors were still motivated by things like impact and feasibility (see their reported study inclusion criteria), it is clear that their motivation was partly to get a representative sample of findings from the field. In this case, the desire to identify studies worth replicating must naturally be weighed against the need for unbiased sampling from the population of original studies. Thus, whether a study - in isolation - is worth replicating is not always the guiding principle for study selection.*
- Arbitrariness (i.e. picking a random study)
  - *Reply: This goal makes sense in scenarios like the one described in the previous point. When the goal is to estimate average replicability in a field, random/arbitrary selection is necessary to ensure valid inferences about replicability to the population of psychology research. It is an interesting factor since it essentially prevents you from selecting on any other criteria. For a broader discussion of the benefits of random study selection, see Kuehberger, & Schulte-Mecklenbeck, (2018).*
- Teaching/pedagogy
  - *Reply: Several people brought this up during discussion. It nicely illustrates a point that one discussion group put a lot of emphasis on: the goal of the replication study will partially determine what is worth replicating. I.e. you cannot adequately answer what should be replicated without first defining what you hope the replication will achieve. For example, if our goal is to reduce*

uncertainty, then we likely want to replicate original studies that are highly uncertain. However, if our goal is pedagogical, we might want to replicate studies with highly certain outcomes (e.g. to demonstrate a principle, or to know when a study design was successfully implemented), even if replicating them does not reduce our uncertainty much.

- Some findings are reliable enough to be used as manipulations in later experiments
  - *Reply: This reminds me of replications that are done because the authors are motivated, not to replicate a finding, but to change or extend a study procedure (this was also mentioned by several people in our discussion). It is important to note that in cases where replication is just a minor step in reaching a larger goal, we may not be replicating to reduce our uncertainty about the results of prior research. In fact, we may actually prefer the results we are replicating to already be highly replicable (e.g. when we reuse a previously validated measurement tool).*

Other interesting comments, suggestions, and questions (and our replies).

- When do we say "there is no need of further replication" of a study or theoretical construct?
  - *Peder: This is a good and difficult question, and it relates to a few other comments made in the session. In general, you probably always need to consider this question as relative to other replication studies you could do. I.e. once you have conducted a replication of study A, is it worth replicating A again, or is there now a study B that has become more important to replicate than A, given that A is corroborated/falsified by the replication you just conducted?*
  - *Anna: this also reminds me that we as researchers are generally bad at knowing how to weigh the evidence we are looking at, perhaps because we are not used to calibrating our conclusions with the evidence, or because we are still catching up when it comes to reporting and expressing our certainty in terms of confidence intervals. For a good read on planning your sample size on the basis of how accurate you want your estimate to be, see Maxwell, Kelley & Rausch, 2008.*
- Most replications are probably conducted by students interested in the subject matter at hand (which might incidentally include some of these factors like influence, importance, likelihood etc.).
  - *Reply: An important point. An implicit assumption behind our work is that researchers often have multiple replication candidates that would be equally interesting to replicate. Alternatively, you could perhaps assume that personal interest is not as important as replicating research with societal impact, highly uncertain results, etc., so personal interest should not factor into the equation. But these assumptions could of course be false, and it is certainly the case that many researchers select studies based on personal interest (see e.g. https://pedermisager.netlify.app/post/what-to-replicate/).*
- Retrospective registered report: If you get a paper results blind, would the method and rationale convince you that you want to run a replication?
  - *Reply: A very intriguing idea! To offer one caveat, I do think that results are sometimes important for knowing if a study is worth replicating. For example, if it turns out that the results of a study are extremely accurate because of very low variance in the estimates then one might want to conclude that this study does not need replication. However, one would need the results to know whether this is the case. On the other hand, one could perhaps assume that variance is mainly determined by factors such as sample size, which are known before seeing the results.*
- Should we really buy into the idea of 'not everything can be replicated'? If a research question was worth asking, worth funding, worth using participants/animals/resources to answer–isn't it worth building in replication from the start? That is, instead of selecting completed projects to replicate, shouldn't we design projects with internal and/or cross-lab replications from the start?
  - *Reply: Of course we should! The question is, whether we still want existing literature to factor into what we study. If we do, we need to know its robustness before we implement build-in replications/validations.*

Links to materials.

PDF versions of all documents and notes pertaining to our session can be found on this OSF project page: https://osf.io/ufea4/. If you'd like a more detailed introduction to our, the session hosts', beliefs about what is worth replicating, I recommend checking out the youtube recording in a previous post: https://pedermisager.netlify.app/post/choosing-what-to-replicate-talk/.

We want to thank everyone involved in our unconference session for their time and contributions!