Rerum Causae
Student journal of Philosophy

Salvador De Viterbo Pitta Borba Gouveia (2021) *The Limits of Randomised Controlled Trials – Lessons from the Replication Crisis*
Rerum Causae, 12(1)

The limits of randomised controlled trials - lessons from the replication crisis

Salvador De Viterbo Pitta Borba Gouveia[1]

[1] London School of Economics and Political Science

Abstract

In recent years, there has been a rising focus on the fact that the biomedical, social, and behavioural sciences might be going through a replication crisis - the fact that a lot of the findings published in the literature fail to replicate. After discussing what the term 'replication' actually means, I turn my attention to a major experimental method in these sciences - the randomised controlled trial (RCT), which is by many considered 'the gold standard' of experimental research. In this essay, I analyse the methodological features of RCTs, and, in light of the factors that are considered to be mostly responsible for the replication crisis, I discuss whether a wider use of RCTs is a viable (if only partial) solution to the replication crisis. I conclude answering this question negatively, arguing that some of the features which make RCTs methodologically advantageous are also those which make them inadequate as a response to the replication crisis.

1 Introduction

The worry that many of the findings that are published in the biomedical, social, and behavioural sciences are false has been subject to increasing attention over the past recent years. For that reason, there has been rising focus on the notion of replication, and the associated 'Replication Crisis'. Probably most notably, the results in social psychology have been particularly hit by this replication crisis. Some estimates point to a failure in replication of around 50% (Open Science, 2015). But the crisis seems to extend to other sciences. Experimental economics has been hit as well, with reports indicating around 40% of replication failure (Camerer et al., 2016). In medical research, there is a similar concern that many published findings are later refuted by follow-up research evidence (Ioannidis, 2005).

A central method that can often be used in these sciences is the randomised controlled trial (RCT), considered by the Evidence Based Movement the 'gold standard', and the 'favourite' method of the most recent Economics Nobel Laureates. But RCTs are not immune to criticism. And, indeed, an exploration of the replication crisis must look at a major method employed by the sciences which are undergoing this crisis. In this essay, I will therefore explore the link between the use of RCTs and the replication crisis. Specifically, I aim to investigate whether a wider use of RCTs can mitigate the replication crisis, and I shall argue not.

With that end in mind, in section II I will briefly present the concept of a randomised controlled trial and present the features which are taken to be RCTs' methodological advantages, and are hence used to justify their gold-standard status, thus grounding the suggestion that the RCT can be a good antidote against replication problems. In section III, I turn my attention away from RCTs to the issue of the replication crisis, and clarify what it means for a certain study to be a replication of another. In section IV, I identify two main accounts of what the factors that give rise to the replication crisis are. In sections V and VI, I will be in a good position to return to the topic of RCTs and what role, if any, they can play in mitigating the crisis. I will do this by looking at how RCTs' alleged methodological advantages interact with the diagnoses presented in section IV. For each of the two identified diagnoses, I will argue that the very features of RCTs that can be considered methodologically advantageous are also problematic once the issue of replicability is considered, and are thus an inadequate response to the crisis.

As such, I conclude in section VII answering negatively the question of whether a wider use of RCTs can mitigate the replication crisis.

II. What are randomised controlled trials?

In order to address the question of whether randomised controlled trials (RCTs) can be a solution to the replication crisis, I first need to present three main concepts — that of a randomised controlled trial, a replication, and the replication crisis. In this section, I shall introduce the former, and in the next one I will introduce the latter two.

An epithet commonly attributed to RCTs among the Evidence-Based Medicine movement is that of 'the gold-standard', which signals that RCTs are superior to the other evidence-gathering methods. In economics, too, RCTs grow popular. Although RCTs do not so commonly hold such a noble status in the social science disciplines as they do in biomedicine (Deaton and Cartwright, 2018), RCT-practitioners in economics are rising stars. Take the 2019 Economics Nobel prize laureates, who, with their introduction of RCTs in the study of global poverty, are credited with having revolutionised the field of development economics, thus being distinguished for "their new experiment-based approach" (The Prize in Economic Sciences, 2019).

An understanding of the methodological features of RCTs which grant them this high praise is pertinent here. An RCT is used to estimate the average effect that a certain treatment X has on an outcome variable of interest Y. The treatment is randomly assigned to some individuals (the treatment group) while other (also randomly assigned) individuals are not treated (the control group). The variable Y is then measured in both groups, and the average difference between them yields the estimated causal effect of X on Y.

The great appeal of RCTs is thus that they offer a 'clean', transparent, and rigorous solution to problems of selection bias and confounding variables which are ubiquitous in empirical work. In randomising the treatment to individuals, we are, in expectation, making sure that the treated and control groups are similar with respect to all relevant factors — observable and

unobservable — to the outcome of interest, but for the fact that one group was treated and the other was not. If a difference in the outcome of interest is registered, then it must be due to the only factor that systematically varies between the two groups — the treatment. In other words, randomisation warrants a *ceteris paribus* assumption when evaluating the (average) effect a certain treatment has on a population of interest.

More detailed formal presentations of the inferential features of RCTs (see, for instance, Deaton (2010a:438-442) or Cartwright (2007)) show how 'positive results in an ideal RCT deductively imply that the treatment causes the outcome' (Cartwright, 2007:15). The fact that the conclusion follows from the experiment *deductively*, rather than inductively, confers it great reliability. Moreover, RCTs do away with the need to explicitly control for all relevant variables, thus overcoming the difficulty often present in empirical studies that some relevant variables cannot be controlled for because they are unobservable. That is why, according to Pearl and Mackenzie (2018:132), the RCT is the one situation in which orthodox statisticians find it acceptable to talk about causality, a notion which has traditionally been regarded as suspicious and obscure within the profession. RCTs are thus seen as extremely reliable tools to find the average causal effect of a certain variable on another.

It is this reliability which might warrant claims that RCTs can offer a solution to the replication crisis — if RCTs offer this clean, reliable solution to problems associated with finding the real causal effect of a treatment on an outcome variable, this might warrant the claim that a wider use of RCTs would result in the attainment of more replicable results. However, we are not yet ready to be precise about how this claim can be warranted, as a more detailed look into the concept of 'replication' is needed. For that reason, it is important to introduce the next two central concepts, so that we are then able to investigate how the replication crisis can be an interesting lens through which the advantages and limitations of RCTs can be explored.

III - Replications, crises, and the replication crisis

Talking of a replication crisis requires clarity about the term 'replication'. Currently, however, a whole variety of studies are referred to as replications - from mere verifications of the code

used to analyse a certain data set, to experiments run in different contexts and in a different time period from the original study (Clemens, 2017). In this section, I will argue for the abandonment of this vague usage in favour of the adoption of a narrow sense of the concept of replication. This narrow sense, I will argue, is more consistent with the notion that there is a crisis of replication in the social, behavioural and biomedical sciences, and allows for a better understanding of the causes the problem.

An often drawn distinction is that between direct and conceptual replication (see Schmidt, 2009). Although there might be disagreements over the precise definition of each, for the purposes of this essay, a rough distinction - which can be borrowed from current literature - will suffice. A direct replication is 'an experiment whose design is identical to an original experiment's design in all factors that are supposedly causally responsible for the effect' (Romero, 2019: 2). Direct replications, as studies which estimate *the same* population parameter (Clemens, 2017) therefore test the *reliability* of the inference presented in the original study, since it is expected that both studies should yield the same result regarding the estimated parameter.

A conceptual replication, on the other hand, goes beyond the scope of the original experiment in some relevant way - 'it attempts to establish the same theoretical conclusion as an original experiment with different experimental manipulations or measures' (Machery, 2019). In a conceptual replication, then, the use of different methods and measures casts doubt on the assumption that there are no good reasons to expect the each experiment (original and follow-up) to yield an estimate of the same population parameter (Clemens, 2017). As Doyen et al. (2014: 28) put it, '[t]he problem with conceptual replication (...) is that there is no such thing as a "conceptual failure to replicate"'. In this sense, follow-up studies which change the methods and/or measurements used are better interpreted as extensions (Clemens, 2017; Machery, 2019), and, as such, robustness tests (Clemens, 2017) - not replications - of the original study. Extensions test the *validity* or *generalisability* (and not reliability) of the inference of the original study (Machery, 2019).

I am henceforth adopting the use of the word 'replication' which matches the narrower sense in which it can be used, that is, a use which excludes extensions from the category of

replication. I thus follow closely Clemens' (2017: 327) consideration that '[a] "replication" test is distinguished by strong reasons to believe that the follow-up test should give, in expectation, materially the same quantitative result as the original study.' The adoption of this narrow sense is useful for philosophers who are particularly interested in studying the replication crisis, because talking of a replication *crisis* has a clear normative connotation. If there is a crisis of replication, this denotes a time of difficulty as far as replication is concerned. The replication crisis therefore has associated with it the normative idea that there is something that should be corrected, changed, in order to address frequent replication failures. This normative connotation of the term 'crisis' perfectly matches the narrow sense of the term 'replication', whereas the same cannot be said if the broader sense is employed. That is because if a replication test is used to refer only to experiments which are expected to yield the same estimate as the original study, then it is potentially problematic for the reliability of the inference presented in the original experiment that they do not. If, however, other follow-up studies which do not offer sufficient reasons to expect the same result as in the original studies are also considered replications, then this very fact should cast doubt as to whether the replication crisis is a *crisis*: if one does not expect the follow-up studies to yield the same results as the original, then it is not clear, without further arguments, why it is a problem that follow-up studies do not replicate (in the broader, rejected sense of the term).

IV - Identifying the causes of the crisis

The vague way in which the term 'replication' is used might suggest that perhaps there is no such thing as a replication crisis, where replication is understood narrowly, as it is possible that the cases used to infer that we are in a replication crisis are all instances of robustness, rather than replication, failures. There are, however, strong reasons to believe that replication failures are indeed more frequent than desirable. Indeed, existing articles have reported low replication rates, where the term replication is understood narrowly (see, for instance, Open Science Collaboration, 2015; Camerer et al., 2016; Romero, 2019).

Having established the plausibility that the mentioned sciences are indeed going through a replication crisis in the narrow sense, it is now pertinent to identify some of the main diagnoses of the causes of the crisis, which I group in two main categories, labelled as follows:

Diagnosis (1) *Incentives in science*

Diagnosis (2) *Lack of theory.*

After presenting these two diagnoses, in section V I will return to the particular experimental method I want to evaluate — the randomised controlled trial —, and I will consider how the insights from these diagnoses can be relevant to exploring the role that RCTs can play in the replication crisis — in particular, whether a wider use of RCTs can mitigate the crisis.

(1) Incentives in Science

In this diagnosis, I include a variety of reasonings which see scientists as actors in a structure with a set of incentives, some of which might be conflicting when some goals - in particular replicability - are considered. In this diagnosis, a lot of emphasis is given to the fact that scientists seek to be published, and that a lot of their career- and reputation-incentives point them towards publishability. Heesen (2018) presents scientists as credit seekers who strive to publish their findings - they 'rush to print'. In taking into account the fact that usually only the first scientist to make a discovery is rewarded - the priority rule (Merton, 1957) -, which incentivises the publication of novel results, Heesen shows how there can be a trade-off between the publishability (which is highly dependent on speed of research) and reproducibility of research (increasing with higher-power studies). Indeed, Nosek et al. (2012) point towards the same conflict between speedy and replicable work. Other structural factors include how 'hot' a certain research topic is - the reasoning behind this being that the greater the number of independent studies on a certain topic is, the higher the probability of finding a false positive (Ioannidis, 2005) - or whether there is direct influence of financial interests on research (Ioannidis, 2005; Howick, 2019; Als-nielson et al., 2003; Lesser et al., 2007; Yaphe et al., 2001). Finally, still related to publishing, is the fact that the criteria that indicate what is publishable bias publications towards a certain kind of type - namely those that report statistically significant (rather than null) results (Nosek et. al, 2012; Romero, 2019). This, in turn, creates incentives for scientists to get this type of results, thus incentivising what are called questionable research practices (QRPs), which allow scientists, by means of exploiting some flexibility inherent to empirical work (Romero, 2019), to achieve statistical significance.

(2) Lack of theory

Some literature focuses instead on the fact that recent empirical work has not been led nor backed up by theory. Muthukrishna and Henrich (2019), for instance, argue for greater cumulative theoretical understanding in social psychology, emphasising the importance of developing sound, formal 'overarching theoretical frameworks'. With a Popperian flavour (see Popper, 1963), Muthukrishna and Henrich (2019) see theory as a framework which informs on which empirical research should be conducted, depending on which theoretical predictions are derived, and then evaluate that theory in light of the evidence gathered, and vice-versa. In particular, having a developed theoretical framework about the phenomena being studied informs on which empirical results are in line with theory, and which aren't. This is relevant as far as replication is concerned, since theory informs on how surprising an empirical result is. The more surprising a result, the more in need of being replicated in order to be confirmed it is. Replications thus serve as a mechanism for scientists to identify which results really hold and which ones were false-positives, and theory is the guide based on which we decide how confident we are that a certain result is a false-positive and thus in greater need of replication.

Oberauer and Lewandowsky (2019) too identify the disconnect between theory and empirical research as a major factor behind replication failures. Yet while Muthukrishna and Henrich focus on the role of theory in guiding which replication work should be prioritised, Oberauer and Lewandowsky highlight the importance of the role that formal theoretical modelling plays in preventing replication problems in the first place. They argue that formal modelling, in requiring clearly identified assumptions for the derivation of empirical hypotheses, constrains the 'researcher degrees of freedom' (RDFs) (Simmons, 2011). The RDF is a measure of the extent to which the researcher can bias results through different (often a posteriori) decisions about which observations to include in the data, what contexts to analyse, what specifications and functional forms to include, etc. - decisions which affect the rate of false-positive findings, and thus their replicability. This is because, without a clearly identified set of assumptions before data is analysed and results are derived, it is easy to achieve statistically significant results conditional on the data that one has collected. As Gelman and Loken (2013:13) put it,

> '[t]here are many roads to statistical significance, and if data are gathered with no preconceptions at all, it is obvious that statistical significance can be obtained from pure

> noise, just by repeatedly performing comparisons, excluding data in different ways, examining different interactions and controlling for different predictors, and so forth.'

Theory development, then, can serve the purpose of introducing some 'preconceptions', so as to make more rigorous and less exploitable the path statistical significance.[1]

V - RCTs and the problem of theory

In the two following sections, I will draw on current debates about the particular method of RCTs. RCTs are used in much of the current experimental research in the biomedical, social, and behavioural sciences, and an inquiry into the replication crisis that these disciplines face should therefore look at the role played by RCTs. Debates about RCTs have so far focused on methodological issues which will be addressed in this essay too. However, the acknowledgement that these sciences are going through a replication crisis provides an opportunity to think about the role played by RCTs in these sciences from another perspective. In particular, it is of special relevance for the evaluation of this method the role it can play in the replication crisis - can a wider use of RCTs mitigate the replication crisis? In this section, I will focus particularly on the problem of theory, addressed in Diagnosis (2), and how it is similar to some current discussions about RCTs and their external validity. I will show how these discussions present a tension between the fact that RCTs can achieve knowledge of what causal relations there are - of 'what works' - fairly independently of theory and the fact that this is bad for assessing an experiment's expected replicability.

[1] Clearly, this point interacts a lot with the worries outlined in diagnosis (1), in particular with the issue of QRPs and the exploitation of the flexibility inherent to empirical work. However, insofar as they put the emphasis on lack of theory as the source of the problem, I chose to include it in diagnosis (2).

RCTs, due to their internal validity - which is achievable under minimal assumptions (Deaton, 2010a) -, are very reliable (if adequately executed) in finding causal effects.[2] Due to methodological advantages such as this, the Evidence-Based Movement considers RCTs to be the gold standard of experimental research in biomedicine, and although such status is not so commonly held in social science disciplines like economics (Deaton and Cartwright, 2018), it has recently become a more and more common method. Its salience is such that the 2019 Economics Nobel prize laureates - who are credited for having revolutionised the field of development economics with their introduction of RCTs in the study of global poverty - were awarded the distinction for "their new experiment-based approach" (The Prize in Economic Sciences, 2019).

But RCTs are not immune to criticism. Objections to both the internal validity of RCTs - ie, objections to the plausibility of the assumptions under which an RCT estimate is likely to be a good estimate of the average treatment effect, even within the context in which the RCT was run - and to their external validity - ie, objections which attack the generalisability of the results achieved by an RCT experiment - are well developed in the literature (see, for example, Worrall, 2007; Cartwright, 2007; Deaton, 2010a; Cartwright, 2012; Deaton and Cartwright, 2018).[3] What a lot of these criticisms have in common, as far as external validity is concerned, is their focus on the lack of understanding of the mechanisms behind the estimated effects - the lack of theory behind RCTs. This is very congruent with Diagnosis (2) in the previous section. I am not saying, however, that they serve the same purpose. Indeed, note how, in section II, I prescribed the use of a narrow sense of the term replication, one which distinguishes replications from extensions, reliability from validity/generalisability. In this sense, the problem of the replication crisis is distinct from the external validity problem. What this

[2] For a presentation of how RCTs achieve this internal validity and are used to find causal relations, see Deaton (2010a: 438-442).

[3] However, see reasons to be sceptical of criticisms to both internal and external validity objec- tions in Howick and Mebius (2016); Backmann (2017).

congruence highlights, however, is how problematic lack of theory can be - not only does lack of theory compromise claims that a certain RCT result is externally valid, it also creates replicability issues. Moreover, as we will see, the existing literature on the problem of lack of theory for the externally validity of RCTs is extremely useful to think about the problem of replicability.

The answer to this problem then appears to be fairly intuitive - if more theory is needed, let us have more theory! However, the problem might not be so easily solvable. Indeed, the usual empirical work done with RCTs has distanced itself from the hypothetico-deductive model of scientific inquiry (Deaton, 2010b; Deaton and Cartwright, 2018), where deductive consequences of theory are tested empirically, precisely because RCTs provide good conditions for that detachment - namely, the previously stated fact that RCTs' internal validity is achievable under minimal assumptions; that we can learn from RCTs 'without overreliance on questionable theory or statistical methods' (Deaton, 2010: 424). RCTs' detachment from theory is thus seen as both an advantage and a source of problems.

However puzzling this conclusion may be, it reflects the depth of the problem of theory. RCTs are a method to find 'what works', and they are able to do that regardless of the background theory. But theory is not totally irrelevant. For one thing, it is needed in order to evaluate whether the repetition of a certain RCT in a different context counts as a replication, because only the understanding of the mechanisms behind a certain effect can inform us on whether those mechanisms hold in a different context (Rodrik, 2009) - in which case we should expect the same estimate in the follow-up study. Secondly, as we have seen in Diagnosis (2), experimentation with little or no theory guidance is not likely to be very informative, and unlikely to suggest which experimental results are more surprising and in greater need of being confirmed by means of a replication. Given an original RCT and an RCT which follows up on it, Box 1 summarises the questions which require theory in order to be answered and are relevant to both the replicability and external validity of RCTs.

- Is the follow-up RCT a replication or an extension?
- If it is a replication, how confident are we that it should replicate (ie, how surprised are we at the original RCT's result)?
- If it is an extension, should we expect the result of the original result to be externally valid in the context of the follow-up RCT?

Box 1

I am not claiming that there is no 'sweet spot' in this tension between the advantages and disadvantages of the methodological features of RCTs. However, the discussion above serves to show that the problem of theory is not so easily solvable. A perhaps even more problematic trade-off, however, is the one which I will present in the next section, and which follows from the *Incentives in Science* diagnosis.

VI - RCTs and the problem of incentives

That RCTs can be very theory-independent in generating 'what works' knowledge - and the problems which that entails for their replicability - has been shown to be serious enough a problem for it not to be dismissed. I would argue, however, that the problems do not end here, and an analysis of RCTs in light of Diagnosis (1) should make this clear.

I introduce three central concepts:

(a) – an experiment's replicability (*R*);

(b) – a scientist's incentive to replicate a certain experiment (*I*);

(c) – an experiment's flexibility of design (*F*).

We can interpret *R* as the probability that a replication will yield the same result as the original replication.

I will depend on considerations such as how likely the scientist thinks her replication is to be published, and not simply how rigorous and well-done the replication is.

F takes into account that different methods allow for different 'researcher degrees of freedom' (Simmons, 2011), ie different (often *a posteriori*) decisions about which observations to include in the data, what contexts to analyse, what specifications and functional forms to

include, etc. - decisions which affect the rate of false-positive findings, and thus their replicability.

RCT's *F* is relatively low. Indeed, randomised trials, because of their low flexibility, are pointed as partial answers to the problem that 'most published research findings are false' (Ioannidis, 2005). That should be good for RCT's replicability: less flexibility, fewer 'researcher degrees of freedom', less room for questionable research practices, lower rate of false-positives. Once incentives are considered, however, one might reconsider the claim that RCTs are viable as a solution to the replication crisis. The very fact that an original study's *F* is low influences *I*. In particular, an inflexible research design may perhaps allow us to be more confident that a follow-up study really is a replication. But this inflexibility also reduces the incentives of scientists to replicate studies with low *F*, because the follow-up and the original studies would be too similar. Going back to the priority rule, what journal would be willing to rehash an already-published study, if the dimensions along which it is allowed to vary, in virtue of its low *F*, are so very limited (Rodrik, 2009)?

We now have seen that what makes RCTs replicable in principle is also what lowers the incentives for their replication. More generally, the relation between *R*, *F* and *I* can be represented in Figure 1, where blue arrows are causal arrows, and the black, double-headed arrow represents correlation.

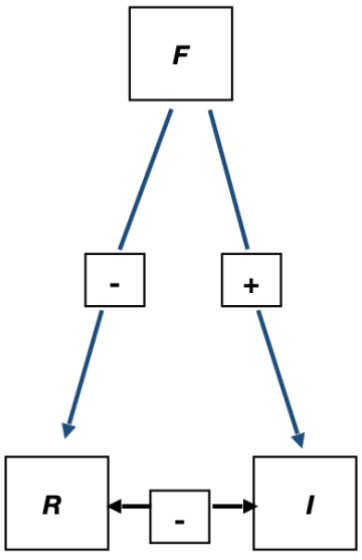

Figure 1

What we see is that, in virtue of its design flexibility, when a certain experimental method allows for higher replicability in principle, all else equal, it is also less likely to generate the right incentives for it to be replicated.

It might be counterargued that the fact that RCT's *R* is relatively high makes the problem of incentives less serious. After all, since *R* is high, we don't expect that many false-positives in the first place, and therefore it is not that problematic that we do not in fact perform RCT replications. This reply, however, is faulty because it ignores the importance of identifying what studies *actually* replicate. Moreover, as seen in the previous subsection, the RCT is a method which promotes evidence-gathering procedures fairly independently from theory. Taking up Muthukrishna and Henrich's (2019) point again, this in turn entails that there is little idea about how surprising RCT-results are, and hence little guidance on which results are in theory more likely to be replicable. This ultimately renders epistemic scenarios where evidence is almost exclusively gathered by RCTs unappealing as far as worries about replicability are concerned — not only are incentives to replicate especially low, we also cannot rely on informed guesses as to which RCTs would *in theory* replicate, because, as we have seen, RCTs promote little theory development.

VII - Conclusion

Perhaps counterintuitively, what Figure 1 reveals is that it might be more beneficial to have not-so-stringent research designs, provided they significantly increase the incentives to replicate. Of course, this is only a *ceteris paribus* analysis. Moreover, a greater *F* might increase the chances that a follow-up study might not in fact be a replication, because the contexts in which the estimates are obtained might be allowed to change too significantly from the original study's. And, indeed, this is one if the reasons why *I* increases with F, since innovation in design increases the likelihood of being published. Once more, theory becomes crucial here, as it informs when a follow-up study really counts as a replication. Finally, it should be acknowledged that the parameters *R*, *F* and *I* are defined somewhat vaguely. I believe, however, that, for the purposes of this essay, the degree of precision employed is sufficient to uncover the tensions which underly the methodolo- gical features of RCTs.

If, on the one hand, RCTs are praised for being able to generate knowledge without needing to rely on contentious theory and models, and for decreasing the possibility for QRPs, they also raise serious doubts as to whether they can be the basis of cumulative and replicable knowledge

in the biomedical, social and behavioural sciences. In light of external validity considerations, RCTs' status as the gold standard has been seriously questioned. The study of the replication crisis serves to further doubt the adequateness of such 'title'. As such, while not claiming that RCTs are behind the replication crisis, I have shown that it is misguided to argue that in them lies the solution to it.

References

Als-nielson, B., W. Chen, C. Gluud, and L.L. Kjaergard. "Association of funding and conclusions in randomized drug trails: A reflection of treatment effect or adverse events?" *Journal of the American Medical Association* 290 (2003):921-928

Backmann, Marius. "What's in a gold standard? In defence of randomised controlled tri- als." *Medicine, Health Care and Philosophy* 20.4 (2017): 513-523.

Camerer, Colin F. et al. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (2016):1433-36

Cartwright, Nancy. "Are RCTs the gold standard?." *BioSocieties* 2.1 (2007): 11-20.
Cartwright, Nancy. "Presidential address: Will this policy work for you? Predicting effecti veness better: How philosophy helps." *Philosophy of Science* 79.5 (2012): 973-989.

Clemens, Michael A. "The meaning of failed replications: A review and proposal." *Journal of Economic Surveys* 31.1 (2017): 326-342.

Deaton, Angus. "Instruments, Randomization, and Learning About Development." *Journal of Economic Literature* 48 (2010a): 424–455

Deaton, Angus. "Understanding the mechanisms of economic development." *Journal of Economic Perspectives* 24.3 (2010b): 3-16.

Deaton, Angus, and Nancy Cartwright. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210 (2018): 2-21

Doyen, Stéphane, et al. "On the other side of the mirror: Priming in cognitive and social psychology." *Social Cognition* 32.Supplement (2014): 12-32.

Gelman, Andrew, and Eric Loken. "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time." *Department of Statistics, Columbia University* (2013).

Heesen, Remco. "Why the reward structure of science makes reproducibility problems inevitable." *The Journal of Philosophy 115.12* (2018): 661-674.

Howick, Jeremy. "Exploring the Asymmetrical Relationship Between the Power of Finance Bias and Evidence." *Perspectives in biology and medicine* 62.1 (2019): 159-187

Howick, Jeremy, and Alexander Mebius. "Randomized trials and observational studies: the current philosophical controversy." (2016): 873-886.

Ioannidis JPA "Why most published research findings are false". *PLoS Med* 2(8) (2005): e124

Lesser, L.I., C.B. Ebbeling, M. Goozner, D. Wypij, and D.S. Ludwig. "Relationship between funding source and conclusion among nutrition-related scientific articles". *Public Library of Science Medicine* 4 (2007):41-46

Machery, Edouard. "What is a replication?." (2019). (unpublished)

Merton, R. K. "Priorities in scientific discovery: A chapter in the sociology of science". *American Sociological Review*, 22 (1957): 635–659

Muthukrishna, Michael, and Joseph Henrich. "A problem in theory." *Nature Human Behaviour* 3.3 (2019): 221-229.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. "Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability." *Perspectives on Psychologi- cal Science* 7.6 (2012): 615-631.

Oberauer, Klaus, and Stephan Lewandowsky. "Addressing the theory crisis in psychology." *Psychonomic bulletin & review* 26.5 (2019): 1596-1618.

Open Science Collaboration. "Estimating the reproducibility of psychological science". *Science* **349**, aac4716 (2015)

Pearl, Judea, and Dana Mackenzie. The book of why: the new science of cause and ef- fect. Basic Books, 2018.

Popper (1963) 'Science: Conjectures and refutations? in *Conjectures and Refutations* Press release: The Prize in Economic Sciences 2019. NobelPrize.org. Nobel Media AB 2020. Wed. 22 Jan 2020. <https://www.nobelprize.org/prizes/economic-sciences/2019/ press-release/>

Rodrik, Dani. "The new development economics: we shall experiment, but how shall we learn?." (2009).

Romero, Felipe. "Philosophy of science and the replicability crisis." *Philosophy Compass* 14.11 (2019): e12633.

Schmidt, S. "Shall we really do it again? The powerful concept of replication is neglected in the social sciences". *Review of General Psychology*, 13(2) (2009): 90–100

Simmons, J. P., Nelson, L. D., & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science,* 22 (2011): 1359-1366

Worrall, John. "Evidence in medicine and evidence-based medicine." *Philosophy Com- pass* 2.6 (2007): 981-1022.

Yaphe, J., et al. The Association Between Funding by Commercial Interests and Study Outcome in Randomized Controlled Drug Trials. *Fam Pract* 18 (6) (2001): 565–68