Replication Is for Meta-Analysis

Samuel C. Fletcher

University of Minnesota, Twin Cities

scfletch@umn.edu

Abstract: The role or function of experimental and observational replication within empirical science has implications for how replication should be measured. Broadly, there seems to be consensus that replication's central goal is to confirm or vouchsafe the reliability of scientific findings. I argue that if this consensus is correct, then most of the measures of replication used in the scientific literature are actually poor indicators of this reliability or confirmation. Only meta-analytic measures of replication align functionally with the goals of replication. I conclude by addressing some objections to meta-analysis.

Acknowledgements: Thanks to discussants at Wuppertal (2020), Castelveccana (PSE4, 2020), St. Louis (PSX6, 2021), and Baltimore (PSA 2020/1) for their comments, which have molded the final form of this essay.

1. Introduction

Over the last decade, in scientific disciplines such as cancer biology (Begley & Ellis 2012; Mobley et al. 2013), social and cognitive psychology (OSC 2015; Klein et. al. 2018), experimental economics (Camerer et al. 2016), and the social and human sciences more generally (Camerer et al. 2018), researchers have found discrepancies between the results of a wide range of studies and the results of replication efforts directed towards them. Although varied in their particular details, these efforts in general seek to repeat past studies using as close to the same methodology as is reasonably possible. These repetitions, or replication attempts, were understood to succeed if and only if their results were sufficiently similar to the originals to be regarded as "the same." Researchers in various disciplines interpret the fact that so many of these replication attempts have *not* succeeded as indicating a widespread methodological problem (Baker 2016). They aver that this "crisis of replicability" (Spelman 2015) is a problem because such successful replication is necessary for scientific results to be confirmed, reliable, or trustworthy.[1] Probing these features is in fact the goal or function of replication efforts in the first place.

However, much of this scientific attention tacitly assumes that the way replication is measured in these studies comports with this function for replication efforts. My goal in this

[1] The (perceived) failure of widespread replication efforts is only one of the indicators of the crisis. See Fidler and Wilcox (2018) and Romero (2019) for more on these and its multifarious hypothesized causes.

essay is to scrutinize this connection. I argue that there is in fact a mismatch between the function of replication and the form it presently takes in much (though not all) of the scientific literature. That form distills statistical properties of the results of pairs of scientific studies to arrive at one of two outcomes: either one study replicates another study or it does not, simpliciter. Such binary conclusions are too coarse-grained to support replication's function of evidence amalgamation and can mislead researchers into seeing replication efforts as attacks on the validity of particular experimental studies. Adopting the rich techniques of meta-analysis, I suggest, would ameliorate these problems while better serving replication's functional goals.

To establish this, first, in section 2, I review how various accounts of what replication is accord regarding replication's function. So, although there is not yet agreement on how exactly replication should be defined, the seeming consensus on its function can serve as a robust starting point for investigating how that function should be measured. Second, in section 3, I show how most criteria or measures of replication in fact poorly represent how well particular scientific results are reliable or well-confirmed. This is because they give a very coarse-grained representation of how different experimental results relate to each other. Third, in section 4, I describe how meta-analysis does not suffer from this problem and how its technical goals align with the epistemological goals of replication in empirical science. Finally, I defend meta-analysis from common criticisms in section 5.

2. Replication's Function

It is not difficult to find scientists in print asseverating the importance of replication. For example, both Moonsinghe et al. (2007, 218) and Simons (2014, 76) declare that replication "is the cornerstone of science."[2] But what part of the edifice of science does this cornerstone support? Moonsinghe et al. allude to its role in confirming causal hypotheses, and Simons to its ability to bolster the reliability of scientific descriptions of phenomena. In referring to their "cornerstone" declaration, Maxwell et al. (2015, 487) emphasize that replication undergirds the trustworthiness of scientific results, allowing us to distinguish "a true finding" from a false positive one.

Philosophers of science seem broadly to concur. As Romero (2019, 1) affirms,"We trust scientific findings because experiments repeated under the same conditions produce the same results." Popper (1959, 24-5) emphasizes that replication is a criterion for the reality and objectivity of a phenomenon, which is the object of scientific inquiry: "the scientifically significant physical effect may be defined as that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed." He goes on:

---

[2] Simons uses the term "reproducibility," but in context intends nothing different from Moonsinghe et al. Some authors use these terms and their derivatives interchangeably, while others use them to denote different activities (Fidler and Wilcox 2018, §1). In this essay, I will use "replication" but intend to draw no particular distinction in doing so. So, when I quote other authors using these other terms I will not editorially alter them.

"Only when certain events recur in accordance with rules or regularities, as in the case of repeatable experiments, can our observations be tested—in principle—by anyone. … Only by such repetition can we convince ourselves that we are not dealing with a mere isolated 'coincidence,' but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable." (Popper 1959, 45). While Popper describes this criterion in terms of testing, which is not in his view a matter of potential confirmation, what's essential for present purposes is that he sees replication as a means to rule out alternative hypotheses (e.g., "coincidences") that would account for the observed regularity of the data.[3] This aligns with Moonsinghe et al.'s view, summarized above, that replication allows one to confirm hypotheses of interest, i.e., disconfirm alternatives.

The *functional account* of replication (Schmidt 2009, 2017; Fletcher 2021a) elevates this observation to a definition and a typology of replication in terms of its confirming or falsifying functions. According to it, the outcomes of individual scientific studies do not typically confirm or falsify scientific hypotheses by themselves, in line with the underdetermination of theory testing by experiment. In order for them to confirm or falsify, they must assume various auxiliary hypotheses, many of which amount to denials that certain other hypotheses would explain the data observed. The goal of replication is to accumulate

[3] This "ruling out" of course must operate within the constraints of testing holism and the theory-ladenness of observation (Popper 1959, sec. 30).

evidence *against* these alternative explanatory hypotheses, which come in a variety of types concerning an original scientific study's results (Fletcher 2021a):[4]

1. They are due to mistakes in the data analysis.
2. They are due to sampling error.
3. They depend on contextual factors that the theory or hypothesis tested claims they do not.
4. They arise from fraud or questionable research practices.
5. They do not generalize beyond the original sample to a larger or different population which the theory or hypothesis tested claims they should.
6. The results do not generalize to other operationalizations or different types of tests of the same hypothesis.

A replication directed towards disconfirming or falsifying one of these alternatives repeats an original study by keeping fixed the latter's procedure, inasmuch as is possible and according to a background's theory's specification of what differences are and are not relevant. This may include features that should not make a difference to the results if the theory or

[4] Providing evidence against the first of these is sometimes called establishing the "reproducibility" of the results, narrowly construed (cf. fn. 2). New studies that provide evidence against 2–5 are sometimes called "direct" replications, and those against 6, "conceptual" replications. Nothing in what follows requires adopting this terminology, or demanding that they neatly partition all the relevant replications.

hypothesis is true, but which would make a difference if one of the alternatives is true. In this way, the logic of replications' implications for scientific hypotheses often make use of Mill's methods.

Now, it is important to emphasize that the functional account of replication asserts neither that replication is the only way to rule out these alternative hypotheses, nor that all these alternatives always arise in the context of any scientific investigation. In this sense, the account does not support any universal claims about replication being a "cornerstone" of science simpliciter.[5] Rather, whether replication is possible and important depends on the particular details of the scientific investigation at hand; in this sense, its importance is "bottom-up," rather than "top-down" (Fletcher 2021a). But the functional account does describe what successful replications establish and what failed replications leave open, thereby grounding most of the other claims that scientists make about the importance of replication. For instance, the sense in which scientific results are more reliable or trustworthy when they have been replicated, according to the functional account, is simply that replications eliminate or disconfirm alternative explanations of the results, thereby confirming the original or target explanatory hypothesis. Such alternatives would be the "false positives" described by Maxwell et al., whose truth would undermine the existence of

[5] Importantly, the tradition of analyses skeptical of this universality (e.g., Cartwright 1991; Leonelli 2018; Feest 2019) have not objected to the goals of replications when they are pursuitworthy.

the "real" effect or phenomenon that the results purportedly support. In particular, in reference to alternative 2 listed above, the results are not likely due to random variation or coincidences in the data.

Although the functional account of replication is not universally held, proponents of alternative accounts seem to agree that at least one central function of replication is to confirm hypotheses of interest and thereby bolster their reliability. For example, while Machery (2020) takes pains to contrast his "resampling" account of replication from the functional account, he nevertheless affirms that "Once a scientist has collected experimental data, she can assess whether a phenomenon is genuine and attempt to characterize it on the basis of these data. The inference from data to the reality and nature of phenomena would be unjustified if the token experiment having produced the data were unreliable … A token experiment is reliable if and only if, if one repeatedly sampled new values for the experimental components that are treated as random factors …, everything else being kept constant, the same experimental outcome would be found with high frequency" (Machery 2020, 554-5).[6] The purpose of a replication, on the resampling account, is "to assess the

[6] Machery (2020, 554-5) notes another defeater for the inference from data to phenomena, that of the "validity" of an experiment: "A token experiment is valid just in case it actually supports the conclusion it claims to establish." In what follows, I will focus on the reliability aspect.

reliability of the original experiment" (Machery 2020, 556), in line with the purposes of replication that the functional account identifies (especially in ruling out sampling error).

Nosek and Errington (2020) also contrast their account with the functional account. They state that a "Replication is a study for which any outcome would be considered diagnostic evidence about a claim from prior research. … To be a replication, 2 things must be true: outcomes consistent with a prior claim would increase confidence in the claim, and outcomes inconsistent with a prior claim would decrease confidence in the claim" (2020, 2). Despite the contrast, this account is in a sense even more functional than the functional account: all that it takes for a study to be a replication is that it fulfills a confirmatory function for an original study's hypothesis; it doesn't also need to have a similar protocol, as the functional and resampling account hold. Nosek and Errington's proposal accordingly has many counterintuitive consequences. For instance, as long as one study is seen not to be completely irrelevant to a hypothesis tested by another, it would count as a replication. But all that matters for present purposes is that even on such an extreme view, there is broad agreement on replication's function. Indeed, this agreement seems to be generic. When Gómez, Juristo, and Vega (2010) canvassed scholarly literature on replication across 18 disciplines, they identified many replication typologies, but found commonality among the functions of replication that align with those of the functional account presented. Thus there seems to be a substantial consensus about the functions of replication in science for confirming hypotheses that explain the data of a study and eliminating or disconfirming alternative explanatory hypotheses.

3. Binary Criteria of Replication

Given that there seems to be broad consensus on the (dis)confirmatory functions of replication, it stands to reason that the best way to measure the extent to which one study replicates another would be in terms of how well the replication study confirms the hypothesis of interest or disconfirms alternative hypotheses. This is especially so when the target of replication is any of the alternative hypotheses, 2-5, e.g., that the original results were due to sampling error (alternative 2). In this case, the replication is most often very similar to the original study in design (a so-called "direct" replication).

However, such measures are not what one finds, generically, when one surveys the sorts of replication criteria employed by the large-scale replication efforts described at the beginning of this essay, or in most methodological discussions of the replication crisis. Instead, one usually finds the use of one or more of several *binary* criteria for replication. In other words, these criteria represent successful replication as a binary relation on pairs of sufficiently (and relevantly) similar studies. They intend to capture a sense in which the results of the studies are the same, or at least sufficiently similar. Judgements of scientific reliability based on them then derive from whether a replication of the original study of interest was successful, or from the proportion of replications that were successful.

There are a variety of binary replication criteria. Here are four that OSC (2015) employed:

1. Compare whether the original and replication studies arrive at the same conclusion in the same test of statistical significance of the same hypothesis at the same

significance level. This criterion, which is probably the most commonly employed in the replication literature, assumes the framework of null hypothesis significance testing (NHST). For instance, in the test of whether a treatment has any effect on a measurable variable of interest, one calculates the p-value of a statistic summarizing the data assuming a null hypothesis of no effect; if the value is below a pre-specified significance level—often 0.05—then the test rejects the null hypothesis, and does not reject otherwise. These are the two possible conclusions of a study using NHST.

2. Also in the framework of NHST, report the result of a significance test of the null hypothesis that the data for the original and replication studies were drawn from the same population. The replication is successful if and only if the test does not reject.
3. Check whether the original study's effect size point estimate lies within the replication study's confidence interval at a certain confidence level—often 0.95—for the effect size (or, alternately, with the role of the original and replication reversed). A particular confidence interval construction procedure selects a set of point hypotheses based on the data—often an interval, if point hypotheses are represented by real numbers—such that at a probability at least equal to the confidence level, the set includes the "true" hypothesis. The replication is successful if and only if the point estimate does lie in the confidence interval.
4. Query the replication team as to whether their results successfully replicate the original's (or with the roles of the replication and original teams reversed).

While there are other binary replication criteria that could be employed, these illustrate both their representative commonality and variability.

There are many apt criticisms of binary criteria for replication. Here I focus on two that apply to all of them in virtue of the features they share, namely, that they are binary relations on studies or experiments.[7] The first is that the division of data into distinct studies or experiments has a pragmatic, conventional aspect that renders replication criteria based on it objectionably conventional. To see this, imagine an experimenter who has collected data using three different instruments, the results of which they publish with a joint data analysis. How many studies have they conducted? Delineating by data analysis, one; delineating by instruments, three. If two of the instruments were directed towards the same spatiotemporal events, then this could count as two studies, delineated by spatiotemporal contiguity. Which of these (or others) the experimenter presents as being the case depends on the conventions of their discipline. These conventions for distinguishing which sets of data belong to different studies allow scientists to coordinate their collective research enterprises and distribute credit, but they do not carry any evidential import. Adopting a different convention does not change the confirmatory import of the total data on hypotheses of interest. Yet different conventions permit different claims about replication.

Consider the case, in the above thought experiment, in which the experimenter delineates by instruments. There are then three studies. Using some binary replication

[7] For further criticisms based on particular details of these replication criteria, see Fletcher (2021b).

criterion, it is possible that one study provides evidence for an effect but the other two do not. This would be a double failure of replication. If the experimenter then delineates by spatiotemporal contiguity, thereby combining what were the two failed replications into one, it's possible for the resulting study to be a successful replication. Yet if the experimenter delineates by data analysis, there is no question of replication success or failure; no replication has yet been attempted. This variability does not permit any objective, convention-free description of how the data bear on replication's confirmatory function.

The second criticism of binary replication criteria that I would like to propound here is that these criteria are too coarse-grained to be as informative measures of replication as one would desire, given that one is committed to the evidential goals of replication. Recall from section 2 that those goals, or functions, are to confirm hypotheses of interest by ruling out or disconfirming alternative explanatory hypotheses for the original study's data. Although there is no agreement on what the correct account of scientific confirmation or evidence is, there is agreement that it comes in varied degrees. Confirmation by instances, hypothetico-deductivism, and Bayesian confirmation theories affirm this (Crupi 2021), as do evidential or epistemic interpretations of classical statistical testing (Mayo 1996, 2018; Fletcher and Mayo-Wilson forthcoming). Even falsificationists, such as Popper, demand high "corroboration" of a theory in order for it to be retained as reliable. But by definition, a binary criterion for replication does not come in varied degrees. Given a replication, any binary criterion is mute about just how well the resulting evidence supports the hypothesis of

interest if the replication is successful, and about just how discordant the resulting evidence is if the replication fails.

To be clear, binary replication criteria do provide *some* information about the strength of our evidence for hypotheses of interest. Given a fixed convention for delineating studies, counting the proportion of replications that agree with an original study, according to some criterion for sameness, is more informative than nothing. But that information is unhelpfully partial and coarse. It does not entail, for instance, that the evidential basis for a hypothesis, hence its reliability, is the same as another even if they have the same number of successful replications. The evidential basis is in finer-grained details of the data. Binary replication criteria may also well serve the practical purpose of drawing the research community's attention to replication problems, despite the partial information that they convey, but that purpose is different than their evidential purpose.

4. Replication and Meta-Analysis

If the goal of measuring replication is really to assess the reliability or confirmation of scientific hypotheses, then a good measure should come in the same degrees that confirmation does. Meta-analysis, the class of statistical techniques that amalgamates the evidence for a hypothesis, such as an effect size, across multiple studies, does exactly this. It is "meta" in the sense that it takes studies as input rather than a single data set. Thus it

combines evidence from multiple studies in order to better understand the strength of the total evidence for or against hypotheses.[8]

Some approaches to meta-analysis with replications simply amalgamate the evidence provided by the original and replicating experiments, then use one of the binary replication criteria outlined in section 3, with the meta-analytic study in place of the replication. But that is not the use of meta-analysis that aligns best with the confirmatory goals of replication described in section 2. Rather, the best use is what Braver et al. (2014, 334) call *continuously cumulating meta-analysis* (CCMA):

> In CCMA, instead of misleadingly noting simply whether each replication attempt did or did not reach significance [in the manner of NHST-based replication measures], we *combine* the data from all the studies completed so far and compute various meta-analytic indexes to index the degree of confidence we can have that a bona fide phenomenon is being investigated. In other words, the individual effect sizes of the entirety of completed studies are

[8] I leave discussion of the formal details of meta-analytic techniques for another occasion; see, for instance, Borenstein et al. (2021) for a contemporary textbook treatment. My argument in this section only depends on the fact that meta-analysis encompasses the general process of using statistical methods to combine the data and evidence about hypotheses, especially regarding the sizes of effects.

> pooled into a single estimate. … The CCMA approach therefore shifts the question from whether or not a single study provided evidential weight for a phenomenon to the question of how well all studies conducted thus far support conclusions in regards to a phenomenon of interest.

The indexes to which Braver et al. refer describe effect sizes and variability, thus quantify the evidence against the sorts of alternative hypotheses described in section 2 whose disconfirmation is the primary goal of replication.

Importantly, meta-analysis uses statistical models in the same way as any other (non-"meta") study does, so it supports the same progressive methodology of starting with simple models, statistically checking the validity of its assumptions, then adding complexity as needed, according to whichever simple assumptions are disconfirmed. For instance, in the presence of heterogeneity indicating evidence *for* certain of the alternative hypotheses described in section 2, there are meta-analytic techniques for correcting the estimate of the total evidence *in spite* of this heterogeneity. (See, e.g., Schmidt and Hunter (2015, Ch. 13) for more on this.) This shows that the effectiveness of meta-analysis does not depend on any prior (dis)confirmation of the hypotheses whose support it is supposed to measure.

5. Objections to Meta-Analysis

Although the tools of meta-analysis, and CCMA in particular, provide non-binary measures of replication that actually serve replication's function, they have not been without

criticism. Here I briefly review and rebut two of these. The first sort of criticism, which is very common in the scientific literature, is that meta-analysis is susceptible to publication and reporting bias (Fidler and Wilcox 2018, §2.2). "Publication bias" names the fact that the publication of scientific results is not independent of the contents and goals of those results; it is typically biased towards positive, novel results, and away from negative ones and replications. "Reporting bias" labels the fact that not all data that scientists collect are reported in their published studies; the data published are biased towards those that support the aforementioned sorts of results and away from those that do not or are ambiguous. Since meta-analyses must operate on available studies, these procedural biases lead to biased conclusions about the existence and sizes of effects, stymying the self-correcting features that cumulative scientific evidence ought to afford (Romero 2016). Recent empirical studies of these biases moreover show that these biases can be large and systematic in practice, not just in theory (Kvarven et al. 2020).

I highlight two sorts of responses to this criticism. First, the mentioned biases affect the ability of *all* replication criteria to fulfill their confirmatory goal. It does not affect meta-analysis uniquely or especially. Second, in contrast with binary replication, the quantitative nature of meta-analytic methods permits tools for detecting and correcting these biases, both through the techniques described in section 4 and through novel methods continuously being developed (Carter et al. 2019). Through computer simulation, Bruner and Holman (2019) found that a hypothetical meta-analyst employing some of these techniques

in the situations for which they are designed restored self-correction to the synthesis of the hypothetical studies' evidence.

The second sort of criticism of meta-analytic methods mirrors my first criticism of binary methods, namely that they involve a kind of objectionable conventionality. Stegenga (2011), for instance, has pointed out that in practice meta-analysts have many choices to make in their studies, several options for which seem to be equally legitimate but which lead to different conclusions. However, I follow Holman (2019) in observing that this seeming underdetermination is transient, and can be resolved through the continual improvement of meta-analytic methods and the consensus-seeking procedures of mutual internal criticism within a scientific community (cf. Jukola 2015).

References

Baker, Monya. 2016. “1,500 scientists lift the lid on reproducibility.” *Nature* 533 (7604): 452–54.

Begley, C. Glenn, and Lee M. Ellis. 2012. “Raise Standards for Preclinical Cancer Research: Drug Development.” *Nature* 483 (7391): 531–33.

Borenstein, Micahel, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to Meta-Analysis*. 2nd edn. Oxford: John Wiley & Sons.

Braver, Sanford L., Felix J. Thoemmes, and Robert Rosenthal. 2014. “Continuously Cumulating Meta-Analysis and Replicability.” *Perspectives on Psychological Science* 9 (3): 333–42.

Bruner, Justin P., and Bennett Holman. 2019. “Self-Correction in Science: Meta-Analysis, Bias and Social Structure.” *Studies in History and Philosophy of Science* 78:93–97.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. “Evaluating Replicability of Laboratory Experiments in Economics.” *Science* 351 (6280): 1433–36.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. “Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015.” *Nature Human Behaviour* 2 (9): 637–44.

Carter, Evan C., Felix D. Schönbrodt, Will M. Gervais, and Joseph Hilgard. 2019. "Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods." *Advances in Methods and Practices in Psychological Science* 2 (2): 115–44.

Cartwright, Nancy. 1991. "Replicability, Reproducibility, and Robustness: Comments on Harry Collins." *History of Political Economy* 23 (1): 143–55.

Crupi, Vincenzo. 2021. "Confirmation." In *The Stanford Encyclopedia of Philosophy,* ed. Edward N. Zalta. Spring 2021 edn. Metaphysics Research Lab, Stanford University.

Feest, Uljana. 2019. "Why Replication is Overrated." *Philosophy of Science* 86 (5): 895–905.

Fidler, Fiona, and John Wilcox. 2018. "Reproducibility of Scientific Results." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Winter 2018 edn. Metaphysics Research Lab, Stanford University.

Fletcher, Samuel C. 2021a. "The Role of Replication in Psychological Science." *European Journal for Philosophy of Science* 11 (23): 1–19.

Fletcher, Samuel C. 2021b. "How (Not) to Measure Replication." *European Journal for Philosophy of Science* 11 (57): 1–27.

Fletcher, Samuel C., and Conor Mayo-Wilson. Forthcoming. "Evidence in Classical Statistics." In *Routledge Handbook of the Philosophy of Evidence*, eds. Maria Lasonen-Aarnio and Clayton Littlejohn. London: Routledge.

Gómez, Omar S., Natalia Juristo, and Sira Vegas. 2010. "Replications Types in Experimental Disciplines." In *Proceedings of the 2010 ACM-IEEE international symposium on*

empirical software engineering and measurement, ESEM '10, 1–10. New York: Association for Computing Machinery.

Holman, Bennett. 2019. "In Defense of Meta-Analysis." *Synthese* 196 (8): 3189–211.

Jukola, Saana. 2015. "Meta-Analysis, Ideals of Objectivity, and the Reliability of Medical Knowledge." *Science & Technology Studies* 28 (3): 101–20.

Klein, Richard A. Michaelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams Jr., Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–90.

Kvarven, Amanda, Eirik Strømland, and Magnus Johannesson. 2020. "Comparing Meta-Analyses and Preregistered Multiple-Laboratory Replication Projects." *Nature Human Behaviour* 4 (4): 423–34.

Leonelli, Sabina. 2018. "Rethinking Reproducibility as a Criterion for Research Quality." In *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise (Volume 36B of Research in the History of Economic Thought and Methodology)*, eds. Marcel Boumans and Hsiang-Ke Chao, 129–46. Bingley: Emerald Publishing Ltd.

Machery, Edouard. 2020. "What Is a Replication?" *Philosophy of Science* 87 (4): 545–67.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing*. Cambridge: Cambridge University Press.

Maxwell, Scott E., Micahel Y. Lau, and George S. Howard. 2015. “Is Psychology Suffering from a Replication Crisis? What Does ‘failure to replicate’ Really Mean?” *American Psychologist* 70 (6): 487–98.

Mobley, Aaron, Suzanne K. Linder, Russell Braeuer, Lee M. Ellis, and Leonard Zwelling. 2013. “A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic.” *PLoS ONE* 8 (5): e63221.

Moonesinghe, Ramal, Muin J. Khoury, and A. Cecile. J. W. Janssens. 2007. “Most Published Research Findings Are False—But a Little Replication Goes a Long Way.” *PLoS Medicine* 4 (2): e28.

Nosek, Brian A., and Timothy M. Errington. 2020. What is Replication? *PLoS Biology* 18 (3): e3000691.

Open Science Collaboration (OSC). 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): 943–51.

Popper, Karl R. 1959. *The Logic of Scientific Discovery*. London: Routledge.

Romero, Felipe. 2016. “Can the Behavioral Sciences Self-Correct? A Social Epistemic Study.” *Studies in History and Philosophy of Science*,60:55–69.

Romero, Felipe. 2019. “Philosophy of science and the replicability crisis.” *Philosophy Compass* 14 (11): e12633.

Schmidt, Stefan. 2009. “Shall We Really Do It Again? The Powerful Concept of Replication is Neglected in the Social Sciences.” *Review of General Psychology* 13 (2): 90–100.

Schmidt, Stefan. 2017. “Replication.” In *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research*, eds. Matthew C. Makel and Jonathan A. Plucker, 233–53. Washington, DC: American Psychological Association.

Schmidt, Frank L., and John E. Hunter. 2015. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 3rd edn. Thousand Oaks, CA: Sage.

Simons, Daniel J. 2014. “The Value of Direct Replication.” *Perspectives on Psychological Science* 9 (1): 76–80.

Spellman, Barbara A. (2015). “A Short (Personal) Future History of Revolution 2.0.” *Perspectives on Psychological Science* 10 (6): 886–99.

Stegenga, Jacob. (2011). “Is Meta-Analysis the Platinum Standard of Evidence?” *Studies in History and Philosophy of Biological and Biomedical Sciences* 42 (4): 497–507.