

# What Is a Replication?

Edouard Machery\*

---

This article develops a new, general account of replication (the Resampling Account of replication). I argue that a replication is an experiment that resamples the experimental components of an original experiment that are treated as random factors and that the function of replications is, narrowly, to assess the reliability of the replicated experiments. On this basis, I argue that the common notion of conceptual replication is confused and that the ongoing controversy about the relative value of direct and conceptual replications should be dissolved.

---

**1. Introduction.** Over the last 10 years, we have learned that a surprisingly large proportion of alleged findings in psychology fail to replicate—up to 60% according to some way of identifying replication failures (Open Science Collaboration 2015). This “repligate” (Machery and Doris 2017) has spread to other areas of contemporary science. By some measures of replication success, more than a third of experiments in experimental economics fail to replicate (Camerer et al. 2016) as do most reported findings about cancer treatment (Begley and Ellis 2012). Repligate has moved beyond the confines of academic research, and it is now widely discussed in public media.

The surprising frequency of replication failures has led to a heated controversy about the best form of replication. While most agree that both forms of replication are valuable, many argue for the superiority of *direct* (i.e., exact) replications over *conceptual* replications, yet some argue for the opposite.<sup>1</sup>

Received April 2019; revised July 2019.

\*To contact the author, please write to: Department of History and Philosophy of Science, University of Pittsburgh, 1117 CL, Pittsburgh, PA 15260; e-mail: machery@pitt.edu.

1. While discussing the failed replication of one of her papers, Schnall (2014) asserts that some view direct replications as “more valid”: “Our entire literature is built on those conceptual replications, but those are not the ones that people are now discussing. . . . They’re called direct replications. The idea there is that you take an experiment in exactly the same way and repeat it with that precise method. . . . *That’s what some people consider more valid in a way*” (my emphasis). She holds the opposite view.

Philosophy of Science, 87 (October 2020) pp. 545–567. 0031-8248/2020/8704-0001\$10.00  
Copyright 2020 by the Philosophy of Science Association. All rights reserved.

Direct and conceptual replications are defined in various ways, but it is possible to capture what many, although admittedly not all, mean by these terms. As a first approximation, a replication is direct if and only if it aims to be identical to an original experiment save for its sample of participants. Thus, Hüffmeier, Mazei, and Schultze (2016, 82) define exact replications as follows: “studies that aspire to be comparable to the original study in all aspects.” Similarly, Schmidt (2017, 237) describes direct replication as follows: “the repetition of an experimental procedure.” By contrast, roughly, a replication is conceptual if and only if it attempts to establish the same theoretical conclusion as an original experiment with different experimental manipulations or measures. Thus, Schmidt describes conceptual replication as follows: “the repetition of a test of a hypothesis or of a result of an earlier research work with different methods” (237).

Pashler and Harris (2012) feel the need to address the belief in the superiority of conceptual replications, which they ascribe to “senior psychologists” and which they put as follows: “Researchers frequently attempt (and publish) conceptual replications, *which are more effective than direct replications* for assessing the reality and importance of findings” (533; my emphasis). To rebut this belief, they give the following argument: “A failure to confirm a result based on a serious direct replication attempt is interesting gossip, and the fact is likely to circulate at least among a narrow group of interested parties. . . . If a conceptual replication attempt fails, what happens next? Rarely, it seems to us, would the investigators themselves believe they have learned much of anything. We conjecture that the typical response of an investigator in this (not uncommon) situation is to think something like ‘I should have tried an experiment closer to the original procedure—my mistake’” (533). By contrast, Stroebe and Strack (2014, 64) endorse the superiority of conceptual replications as follows: “Because failures of exact replications do not tell us why findings cannot be replicated, they are ultimately not very informative. The believers will keep on believing, pointing at the successful replications and derogating the unsuccessful ones, whereas the non-believers will maintain their belief system drawing on the failed replications for support of their rejection of the original hypothesis.” Strikingly, both articles appeal to similar considerations to defend opposite conclusions. On the one hand, the failure of a conceptual replication is said to be uninformative because it could result from relevant differences between the original experiment and its conceptual replication; on the other hand, the failure of a direct replication is said to be uninformative because it does not tell us why the replication failed.

The controversy about the best form of replication is on going, and the available arguments have failed to sway scientists’ opinion in one direction. The stakes are high, however. While replication is meant to allow scientists to correct the empirical record (but see Romero 2016), this controversy may

stand in the way of prompt and consensual self-correction. To defend themselves, psychologists often appeal to successful conceptual replications when direct replications of their work have failed (e.g., Bargh 2012; Baumeister, Tice, and Vohs 2018), and people who have run failed direct replications are prone to dismiss successful conceptual replications (e.g., Chambers 2017).

To assess the respective merits of direct and conceptual replications, a general account of what a replication is and what it is for (its function or functions) seems needed. (Similarly, to know which font is better, it is useful to know its purposes: readability on screen, advertisement, etc.) Surprisingly, there is little discussion of what a replication is, in general, and of its function (but see Schmidt 2009, 2017; Asendorpf et al. 2013). Much of the related literature proposes typologies of replications (e.g., Hüffmeier et al. 2016) or argues for the superiority of one kind of replication (e.g., Cesario 2014; Simons 2014; Stroebe and Strack 2014; Lynch et al. 2015; Crandall and Sherman 2016).

This article develops a general account of replication (the Resampling Account<sup>2</sup>) that is applicable in many disciplines and that helps resolve the controversy about the value of direct and conceptual replications. I argue that a replication is an experiment that resamples the experimental components of an experiment that are treated as random factors and that the function of replications is, narrowly, to assess the reliability of the replicated experiments.<sup>3</sup> (Much of this article is dedicated to explaining this account.) On the basis of the Resampling Account, I argue that the usual notion of a conceptual replication is confused, and I end up rejecting the very distinction between direct and conceptual replication, as it is usually drawn. I conclude that the debate about the relative value of direct and conceptual replications should be dissolved rather than resolved.

To develop a general account of replication will require a fair amount of stage setting, which will take place in section 2 of this article. On this basis, section 3 will present the general account of replication. Section 4 investigates the implications of this general account for the controversy about direct and conceptual replications. Section 5 responds to two objections.

The following caveat might be useful. The Resampling Account of replication is not meant to capture what scientists mean when they use the word “replication.” It is not a piece of conceptual analysis. Rather, it is a characterization

2. This appellation has the downside of suggesting a connection to the use of resampling in statistics (thanks to Geoff Cumming for this point), where there is none. I have decided to still use this appellation because it captures the core idea of the account and because my account has already been discussed under this appellation.

3. This account shares similarities with Asendorpf et al.’s (2013) Brunswickian account of replication but was developed before getting to know this article.

of what a replication is, based on a principled account of what experiments are; it is meant to replace scientists' often vague understanding of what replications are and what they are for. That is, it is a piece of conceptual engineering (on conceptual engineering, see, e.g., Machery [2017]).

**2. Phenomena, Experiments, Experimental Components, and Reliability.** Four pieces are needed to develop the Resampling Account of replication: the distinction between data and phenomena, the narrow notion of an experiment, the notion of an experimental component, and the notions of reliability and validity. I examine them in turn.

*2.1. Data and Phenomena.* While there is no consensus about how to draw the distinction between data and phenomena (Bogen and Woodward 1988; McAllister 1997; Colaço 2019), most philosophers of science would agree with the following minimal account. Phenomena are what scientific theories by themselves predict and explain. They are what the empirical consequences of theories are about.<sup>4</sup> They correspond to what scientists often call “effects.” Data are the values of particular measurements in an empirical (i.e., experimental or observational) context. A data set resulting from an experiment or a sequence of observations is unique: it is a particular. While the ontology of phenomena is not fully clear, they are undoubtedly not particulars. Phenomena and data differ in uncontroversial ways. Phenomena are predictable; by contrast, the exact values of the data points are not predictable since data points are influenced by a myriad of factors, including sampling and measurement error. Naturally, the average tendencies of data are predictable within a margin of errors, but the data points themselves are not. For all that, data and phenomena are related. In particular, the reality and nature of phenomena are typically inferred from the data, which themselves are the products of measurement.

To illustrate, the conjunction fallacy is a well-known phenomenon in psychology (Tversky and Kahneman 1982). People tend to judge that an individual is more likely to fall under a conjunction (e.g., to be a feminist bank teller) than under one of the conjuncts (e.g., to be a bank teller) when the individual is more typical of the conjunction than of the relevant conjunct. Participants' answers to the question of whether Linda is more likely to be a feminist bank teller or just a bank teller in Tversky and Kahneman's (1982) experiment are the data. These data served as evidence for the existence of the conjunction fallacy.

*2.2. What Is an Experiment?* The word “experiment” is used in various ways in and outside philosophy. Sometimes, it encompasses any scientific activity that involves some measurement. Here, however, the notion of

4. Naturally, when combined with further details such as information about initial conditions, scientific theories can also predict and explain singular events.

experiment will be used more narrowly, standing in contrast with the notion of observation.

As a first approximation, a token experiment is a sequence of events ( $e_1, \dots, e_j$ ) brought about in order to produce data relevant to inferring the reality and nature of some phenomenon. Experiments differ from observations in that the former, but not the latter, involve experimenters manipulating a system, that is, intervening on a system so that it enters a given state before measurement. For instance, participants in a pharmacology experiment are assigned to the treatment or placebo condition. As characterized here, experiments do not require a comparison between an experimental and a control condition (in contrast to the characterization of experiments typically found in textbooks in the behavioral sciences).

The sequence of events that constitutes a token experiment ( $e_1, \dots, e_j$ ) belongs to a given type ( $E_1, \dots, E_j$ ). Two token experiments ( $e_{1A}, \dots, e_{jA}$ ) and ( $e_{1B}, \dots, e_{jB}$ ) are the same experiment if and only if their constitutive sequences of events fall under the same sequence type ( $E_1, \dots, E_j$ ). Event types can be individuated more or less coarsely (as can any type), and depending on how these types are individuated two token experiments count or fail to count as the same experiment. For example, a token psychological experiment could involve collecting data from 100 participants from a given population (perhaps to obtain a given power). The corresponding type could specify the exact number of participants, in which case any study with a different number of participants would not count as the same experiment, however similar it is to the original study in other respects. Or, more coarsely, it could just specify that some participants were sampled from the relevant population, in which case a follow up study would count as the same experiment whatever its sample size is.

Event type individuation also determines the specific respects in which two events must be identical to count as instances of the same type. Because only specific respects matter, two events need not be identical in every respect to be the same, that is, to fall under the same event type. Similarly, a paperback and a hardcover version of *Ulysses* differ in various physical respects, but they are the same book because they fulfill the sameness criteria associated with being *Ulysses*.

Lynch and colleagues (2015, 333) are thus too hasty when they write: “The very concept of an ‘exact replication’ in social science is flawed. Even if one used the exact same procedures, respondents may have changed over time. Exact replication is impossible. Therefore, the only issue is how close the replication is to the original, and whether it is desirable to be ‘as close as possible’” (see also Schmidt 2009, 92). The mere existence of unavoidable differences between an original experiment and its follow-up does not entail that an experiment cannot be exactly replicated. Two token experiments are the same if they are identical in the relevant respects, even if they differ in other respects. It would be correct to say that no experiment could ever be exactly

replicated only if the differences between an original experiment and its follow-up would always be relevant to the individuation of the event types, but there is no reason to believe this to be the case. Admittedly, it can be difficult to specify what the relevant respects are, but specifying them is certainly possible; indeed, psychologists typically agree that some differences just do not matter, while others undoubtedly do.

Scientists are fully aware of the importance of reporting the events that constitute experiments. In psychology, journals require authors to describe in an article section typically called “Procedure” the events that constitute the token experiment the results of which are reported (e.g., Gigerenzer and Hoffrage 1995, 693). It is, however, rarely clear which aspects of the experiment are meant to be included in the type of experiment conducted (in contrast to the token experiment) and which are not. Furthermore, as Collins (1985) has rightly emphasized, sometimes (particularly when new instruments at the cutting edge of science are used) scientists’ knowledge of the events that constitute experiments is tacit and hardly verbalizable and thus hard, if not impossible, to report.

*2.3. Experimental Components.* An experimental component is an aspect of an experiment that can be independently modified. Psychologists often distinguish four different experimental components: experimental units, treatments (i.e., “independent variables”), measurements (i.e., “dependent variables” or “response variables”), and settings.<sup>5</sup> The experimental units are the entities to which treatments are applied and whose reaction (or behavior) is measured. Experimental units can be individuals (e.g., human beings or animals) or groups (e.g., factories or countries).

A treatment is an exogenous cause that changes the state of some aspect of the experimental units (i.e., the value of some variable characterizing them). Scientists intend to determine whether and how this change influences some other aspect of the experimental units (i.e., some other variable). When an experiment has several conditions (e.g., drug vs. placebo), the treatment can be in one of several states; psychologists often say that it has several “levels.” In a typical psychology experiment, participants assigned to different levels of the treatment are presented with different stimuli. An experiment can involve several treatments. Two treatments are crossed when measurement happens for each combination of the levels of these two treatments. They are nested when measurement happens for some combinations of the levels of these two treatments.<sup>6</sup>

5. Experimental components are constitutive of an experiment type’s defining event types.

6. More generally, all experimental components can be crossed or nested. For instance, in a between-subjects experiment experimental units are nested under treatment since different participants are exposed to different levels of the treatment.

Measurement is a causal interaction with the experimental units aimed at determining what state a particular aspect of the experimental units is in. Measurement typically involves creating an observable variable (e.g., the height of the column of mercury in a mercury thermometer) that is causally influenced by (and thus provides information about) the variable (e.g., temperature) scientists hope to influence by means of the treatment (e.g., ingesting a dose of paracetamol). The setting is a vague and umbrella construct, which includes the identity of the experimenter and of the lab conducting the experiment, whether the experiment is done online or in a lab, and so on.

Psychology journals require authors to describe the experimental components that constitute their experiments, with a special focus on experimental units, treatments, and measurements. (Less emphasis is put on settings.) Gigerenzer and Hoffrage (1995) describe the experimental units—the participants—as follows: “Sixty students, 21 men and 39 women from 10 disciplines (predominantly psychology) from the University of Salzburg, Austria, were paid for their participation. The median age was 21 years. None of the participants was familiar with Bayes’ theorem” (692). Their description of the sample does not describe precisely the population sampled from. It is not clear which aspect of the sample is meant to characterize this population (e.g., whether it is made of students, of people of a certain age, of people ignoring Bayes’ theorem).

Gigerenzer and Hoffrage’s study crosses two treatments (called “format” and “menu”), each with two levels (frequency vs. probability format and short vs. standard menu), resulting in four conditions. A third treatment, which we could call “vignette,” is then nested under these two treatments. Depending on which of the four conditions a participant is in, she will be exposed to different versions of 15 vignettes. Table 1 of their article presents the generic form of the four conditions as well as the relevant version of one of the 15 vignettes for each of these conditions together with the dependent variable (Gigerenzer and Hoffrage 1995, 688; the first condition is reproduced here as table 1). Each

TABLE 1. ONE OF THE FOUR COMBINATIONS OF INFORMATION FORMATS AND MENUS FOR THE MAMMOGRAPHY PROBLEM

Format and Menu	Description of Problem
Standard probability format	The probability of breast cancer is 1% for women at age forty who participate in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? ____%



participant is presented with two versions (out of the four possible) of 15 vignettes (or “problems” as Gigerenzer and Hoffrage call them). Finally, the setting is briefly described as follows: “Participants were studied individually or in small groups of 2 or 3 (in two cases, 5). We informed participants that they would need approximately 1 hr for each session but that they could have more time if necessary. On the average, students worked 73 min in the first session (range = 25–180 min) and 53 min in the second (range = 30–120 min)” (693).

Experimental components are either fixed or random factors. If they are random factors, their levels are randomly sampled from a population (i.e., universe). They could have been different. For instance, other participants could have been sampled (if experimental unit is a random factor) or other stimuli could have been used (if treatment is a random factor). The experimenter intends to generalize statistically from the sample (i.e., the observed levels) to the population as a whole and, thus, to unobserved but possible levels, instead of limiting her conclusions to the observed levels. Because the levels of random factors are sampled from a population, they can represent this population more or less accurately, and sampling error must be taken into account.<sup>7</sup> By contrast, if the levels of experimental components are fixed factors, the levels used in an experiment exhaust the relevant population. That is, the experiment examines all the possible values the experimental component could take, and the levels of an experimental component could not have been different. The experimenter does not aim to generalize statistically to unobserved levels; rather, she limits her conclusion to the observed levels. In a randomized controlled trial testing a new drug, treatment is often conceived as a fixed factor. The new drug is not conceived as one of the many drugs participants could have received, and the experimenter does not intend to generalize to other drugs. There is no sampling error related to the choice of treatment to be concerned with.

Whether an experimental component should be treated as a random or a fixed factor depends on the particular scientific context. Experimental units are typically explicitly treated as the levels of a random factor since they are typically meant to stand for a population. The participants in a psychology experiment are often assumed to be randomly sampled from a particular population, although the identity of the relevant population is rarely made explicit (as is the case in Gigerenzer and Hoffrage 1995), and the sample is

7. How to take into account that the sampling error introduced by treating treatment as a random factor has been discussed extensively (e.g., Kenny 1985; Richter and Seay 1987; Judd, Westfall, and Kenny 2012).



almost never genuinely random. Treatment is sometimes properly conceived as a fixed factor, as is illustrated in the drug testing example above. In other cases it should be viewed as a random factor. When participants are exposed to some particular stimuli (words in psycholinguistics experiments, vignettes in judgment and decision-making experiments, faces in face perception experiments, etc.) that are meant to stand for a broader class of stimuli (e.g., all the words participants could have been presented with in a psycholinguistics experiment), treatment should be conceived as a random factor, although psychologists rarely explicitly do so in such conditions (except in psycholinguistics). The points just made about treatment extend to measurement. In psychology, the particular measurement used in an experiment is rarely explicitly treated as the value of a random sample from a population of measurements.

The distinction between fixed and random factors determines which statistical generalization is allowed by an experiment. It is only when an experimental component is a random factor that one can generalize statistically from the observed levels of this experimental component to the unobserved levels.

Importantly for the Resampling Account, experimental unit is not the only experimental component that can be legitimately treated as a random factor. Treatment and measurement can too, and in some circumstances they should (e.g., Wells and Windschitl 1999; Judd et al. 2012). It is admittedly unusual to do so explicitly, as noted above. One could perhaps argue that it is not an accident that treatment and measurement are rarely explicitly viewed as random factors. First, the treatment and measurement used in an experiment are often intentionally developed through painstaking processes of piloting and validation; at the very least they are intentionally chosen. How then could they be properly thought of as randomly sampled from a population of treatments and measurements? Treating treatments and measurements as randomly sampled is an idealization, exactly as it is an idealization to treat an experiment's experimental units as randomly sampled when they are a convenience sample. A psychologist could require the students in one of her classes to complete an experiment for credit but still treat her participants as a random sample. One might resist this comparison as follows. Even when not randomly sampled, participants in an experiment are not chosen because of their properties; they are chosen because of their availability, and any other sample would have been equally good. By contrast, treatments and measurement are often chosen because of their distinctive properties, and it is not the case than any other treatment or measurement would have been equally good. This response fails, however. When treatment and measurement are viewed as random factors, their distinctive properties (which result from piloting, measure development, etc.) are taken to be shared by

the other members of the population from which they are by idealization assumed to be randomly sampled. Similarly, if a psychologist asks her male students to complete a study for credit because she studies how men react to some stimuli, being male is taken to be a property of all the members of the population she assumes to be sampling from.

Second, it is often unclear what the populations of treatments and measurements could be, and it might thus seem to make little sense to treat treatments and measurements as being sampled from a population. However, it is also often unclear which population experimental units are sampled from, and in this respect the difference between treatments and measurements on the one hand and experimental units on the other is only a matter of degree. Just as scientists should specify the population of experimental units, it is incumbent on them to specify the population of treatments and measurements when those are random factors. In response, one may insist that at least we have ways of specifying the population of experimental units (e.g., all human beings or all adult Westerners) but not the alleged populations of treatments and measurements. The difficulty of specifying such populations is often exaggerated. For instance, for the standard probability format condition in Gigerenzer and Hoffrage (1995), one can, and arguably should (sec. 5.1), specify the population of vignettes by describing a recipe for producing more vignettes:

1. A vignette should refer to a given condition that can be detected by a test;
2. it should specify the base rate of this condition in a population by means of a percentage;
3. it should specify the test's hit rate by means of a percentage;
4. it should specify the test's false positive rate by means of a percentage;
5. it should indicate that a member of the population has received a positive test.

One may wonder whether it is realistic to require specifying populations of stimuli. However, to the extent that the psychologist is making a claim that goes beyond the stimuli she is using, she must specify the population of stimuli she intends to generalize to, even if she could learn that her hypothesis was mistaken.

*2.4. Reliability.* 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 or if it were invalid.

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 (e.g., repeatedly sampling new participants from the original population of participants—say Americans—or repeatedly sampling new stimuli from the original population of stimuli), everything else being kept constant, the same experimental outcome would be found with high frequency. A token experiment is unreliable just to the extent that its experimental outcome is an outlier. Most experimental outcomes resulting from repeated resampling would be different. Unreliability results from several sources. It can result from nondirectional error: a token experiment is reliable only if its experimental outcome is not due to nondirectional error. Nondirectional error is error (distance between the estimate of a quantity and its true value) that is as likely to result in overestimation as in underestimation of the quantity to be estimated. It results from measurement error, sampling error, or imprecise manipulation. The less precise measurement is (think, e.g., about an imprecise thermometer) or the less precise the manipulation is (e.g., the more the vignettes in psychology are open to interpretation), the less frequently the same result will be found. Unreliability can also result from other sources such as honest mistakes, frauds, or questionable research practices. When a scientist makes up her data, if one sampled new values for the experimental units, one would likely get a different experimental outcome; the experiment is unreliable.

This characterization of reliability appeals to the notion of sameness of outcome and thus requires a way of individuating experimental outcomes. There is no interest-independent way of individuating experimental outcomes. Some outcome that counts as the same result as another outcome given some interests may well count as a different result given other interests. For instance, in some circumstances finding again that a manipulation has an effect in a given direction (it either increases or decreases the average dependent measure) counts as obtaining the same experimental outcome; in other circumstances, one would find the same experimental outcome only if the effect sizes in the original experiment and the replication were similar (what counts as sufficiently similar itself is interest dependent). Naturally, once interests are fixed, there is a matter of fact as to whether an experimental outcome counts as the same as another one.

A token experiment is valid just in case it actually supports the conclusion it claims to establish. It is internally valid just in case it actually supports the causal claim that the treatment caused the measured difference between the conditions. Plausible but uncontrolled confounds undermine the internal validity of an experiment. A token experiment is externally valid just in case it actually supports a conclusion about a situation outside the lab that is of interest to scientists and motivated the research in the first place.

### 3. What Is a Replication?

*3.1. The Resampling Account of Replication.* We can now characterize the notion of replication:

Experiment A replicates experiment B if and only if A consists of a sequence of events of the same type as B while resampling some of its experimental components in order to assess the reliability of the original experiment.

In the remainder of this section, I comment on the main aspects of this characterization and highlight some of its most significant consequences.<sup>8</sup>

Experiment A can only replicate experiment B if A and B's constitutive events are of the same type. We have already noted that events can be typified more or less coarsely (an issue known in epistemology as the generality problem and in philosophy of science as the reference class problem). As a result, whether two experiments are of the same type is always a matter of interpretation. It depends on how the relevant types are individuated, a matter often left implicit. In Open Science Collaboration (2015), differences were intentionally introduced between some of the original experiments and their replications. For instance, participants in Shnabel and Nadler (2008) were Israeli, while its replication took place in the United States. Because what is meaningful for Israeli participants might not be meaningful for American participants, superficial features of the vignettes used as stimuli were modified, while the significant structural features remained the same. Nosek (2016) takes this difference to be irrelevant for the identity of the stimuli: they are the same type of stimuli because they share the same structural features. Gilbert et al. (2016) take this difference to be significant and contest that the alleged replication is a genuine replication of the original study (sec. 5 discusses this issue further).

According to my proposal, replications should be conceived as involving the resampling of the experimental components when these are random factors. Each of the experimental components so treated can be the target of resampling. Most obviously, replications nearly always, but not necessarily, involve sampling another group of experimental units from the original population. But the three other experimental components can also be the targets of resampling. If treatment is viewed as a random factor, then the stimuli used in experiment A can be replaced with other stimuli in experiment B (see sec. 2.2 for a discussion of whether it makes sense to treat experimental components other than experimental units as random factors). Consider, for instance, the use of vignettes as stimuli in the judgment and

8. I do not use "replication" as a success term, and I distinguish between a replication and a successful replication.

decision-making literature. If treatment is a random factor, then they are a random sample from a population of vignettes, and any finding obtained with particular vignettes is meant to be replicable on the basis of other samples from the same population of vignettes.<sup>9</sup> To determine whether an experiment type is reliable with respect to its treatment, we resample the particular treatment from the relevant population of treatments. The same is true of measurement (think, e.g., of various scales meant to measure the same construct) or of settings. By contrast, when an experimental component is a fixed factor, changing its value is not a replication; it is a different experiment.

On the Resampling Account, sampling from a different population (of participants, of stimuli, etc.) is not replicating an experiment; rather, one extends a previous experiment. I propose to contrast *replications*, which involve resampling from a given population, and *extensions*, which involve either sampling from a different population (for the experimental components treated as random factors) or changing the level of an experimental component treated as a fixed factor (see also Bonett 2012). Researchers may want to know whether the outcome of an experiment run with neurotypical individuals (e.g., Machery 2008) holds for people with Autism Spectrum Disorder. Their experiment, which repeats the original one except that participants are now sampled from the population of people with Autism Spectrum Disorder (e.g., Machery and Zalla 2015), should not be considered a replication according to the Resampling Account, and it typically would not be. We would not suspect that the original result is a false positive if it is not found with the new population. The difference between this second experiment and a genuine replication is that the latter, but not the former, samples from the same population as the original experiment. (I discuss how one identifies population in sec. 5.1.) Or consider a case study of a patient suffering from some lesion in cognitive neuropsychology (e.g., HM in Scoville and Milner 1957; Squire 2009). In this case, the experimental unit is a fixed factor: its value is the patient under consideration (e.g., HM). Doing another case study with a patient suffering from a similar lesion (e.g., DC in Scoville and Milner 1957) is not replicating the first study.

The Resampling Account seems to have two counterintuitive consequences. First, on this account a follow-up experiment that would use the same participants, manipulations, and measures (e.g., to assess measurement error) would not count as replicating the original experiment. In response, it is enough to notice that the setting of the experiment would change since the experiment would not be done at the same time. Resampling would thus be taking

9. Because of random error, among other sources of noise, one cannot expect any finding to replicate successfully with all the samples from the populations of vignettes. By chance, some false negatives are to be expected among replications, even if the original finding is genuine.

place. Second, if all the experimental components are resampled, a follow-up experiment could replicate an original experiment despite looking very different (because it involves different manipulations and measures). In response, one should note that this consequence holds on any account of replication. What are called “conceptual replications” can be very different from the original experiments. In addition, while the follow-up experiment would be superficially different, it would be the same experiment at a deeper level. The manipulations of the original experiment and its replication, for instance, would belong to the same population of treatments and would be the same in this respect. A critic may insist that failure of this replication would say nothing about the reliability of the original experiment, but this objection should be resisted. Even if the sample of, say, stimuli in the follow-up experiment looks very different from the original sample, the replication examines whether the original result is due to the original stimuli’s peculiarities, hence, whether the original experiment was reliable.

The Resampling Account of replication might not seem compatible with the fact that populations can change. Psychological phenomena that held, say, in the 1970s might not hold anymore in the 2010s (for related discussion, see Lovett and Munger [2019]). If the phenomena change, then a replication would not regularly give the same result despite being *prima facie* reliable. However, on second thought, such “time sensitivity” does not challenge the Resampling Account since if the population changes, psychologists are not sampling from the same population anymore. What looks like a replication is in fact an extension.

The Resampling Account is not narrowly tailored to psychology but extends to other disciplines, including pharmacology, experimental economics, molecular biology, and other life sciences. It is meant to apply to any experiment that fulfills the following two conditions: (1) it is possible to distinguish its experimental units, treatment, measurement, and setting, and (2) at least some of its experimental components can be viewed as sampled from a population. Experiments in many scientific disciplines fulfill these conditions. For example, a drug test can involve a sample of participants assigned to one of two conditions (e.g., drug vs. placebo), and the effect of treatment is measured by some outcome variable: experimental unit is a random factor, while other experimental components such as treatment are treated as fixed factors.<sup>10</sup>

Finally, the Resampling Account converges with Hacking’s (1992) description of experiments. Hacking distinguishes the target of modification (which corresponds to the experimental units), the source of modification (treatment), and “the detector” (measurement). Although Hacking does

10. Experiments in some disciplines perhaps do not meet these conditions, but no clear example comes to mind.

not insist on sampling, the target of modification of a particular experiment, the source of modification, and the detector can be (often are) conceived as samples from some populations.

*3.2. The Function of Replications.* According to the Resampling Account, the function of replications is to check whether a token experiment  $e$  that claims to identify and characterize a candidate phenomenon is reliable, that is, whether an experimental outcome similar to the one found in  $e$  would be found frequently were one to resample from the populations corresponding to the experimental components treated as random factors (if similar results would not be found frequently, then the results of  $e$  are due to the sources of unreliability such as sampling or measurement error). Reliability can be assessed with respect to the four experimental components identified in section 2. For instance, when the treatment of  $e$  is modified, it is to test whether the experimental outcome of  $e$  is not due to one of the threats to reliability, including (but not limited to) sampling error in the type of manipulations used in this experiment (i.e., if the outcome of  $e$  is not due to atypical stimuli or manipulations).

So, on the Resampling Account, the function of replications is really to test reliability rather than validity. Further, it is not to determine the invariance range of a finding: one does not replicate in order to control for confounds or to find out whether the phenomenon still occurs when background conditions are modified. In this respect, the Resampling Account differs starkly from the vague characterizations of replication in psychology and other sciences.

*3.3. The Superiority of the Resampling Account.* There are few general accounts of replication and of its function. Before comparing the Resampling Account to one of them, it is worth highlighting the three main virtues of this characterization of replication. First, it treats all the components of experiments (experimental units, treatment, measurement, and setting) similarly. Treatments or measurements can be sampled exactly as experimental units, and when one uses new stimuli or measurements (sampled from their respective populations) one does exactly the same thing as when examining new experimental units (e.g., new participants). There is no need to introduce a fundamental distinction between those replications that modify experimental units and those that modify treatments and measurements (more on this in sec. 3.4). Second, the resulting account of replication is satisfactorily delimited. On this account, not every experiment that is in some respect or other similar to an original experiment counts as a replication. Third, and relatedly, it allows us to distinguish extensions from replications in a principled manner. In practice, scientists usually distinguish these two forms of experiments, as shown by the fact that psychologists redoing a psychological study in a



new cultural context do not characterize their work as a replication. Any acceptable account of replication must be able to draw this distinction.

We can illustrate these three virtues by examining Schmidt's (2009, 2017) functional account. Schmidt (2009, 94) characterizes a replication as follows: "B is a replication of A if A's primary information focus is reestablished in B." The primary information focus consists of the constructs that are manipulated and measured. A replication can involve repeating the primary information focus by the same "material means" (roughly same operationalizations) or by a "radically different material realization" (94). Schmidt also identifies the different functions a replication can have. He writes (93):

Replications serve several different functions. The general function of replication is . . . to verify a fact or piece of knowledge. However, this implies the following more specific functions:

1. To control for sampling error (chance result),
2. To control for artifacts (lack of internal validity),
3. To control for fraud,
4. To generalize results to a larger or to a different population,
5. To verify the underlying hypothesis of the earlier experiment.

Schmidt limits sampling error to the selection of participants (93), failing to acknowledge that all the experimental components can be random factors and failing to treat all the experimental components similarly. His notion of replication is extremely broad and does not distinguish a replication from the testing of a result's robustness by different means (Wimsatt 2007). It leads him to treat extensions as a distinct type of replication.

*3.4. Doing without Conceptual Replication.* The Resampling Account reveals that the notion of conceptual replication is confused: it fails to distinguish different ways of modifying the treatment, measurement, and setting of an original experiment. Psychologists often only view experimental units as a random factor, and they do not acknowledge that treatments, measurements, and settings too can be resampled. (Again, there are exceptions, such as psycholinguists.) As a result, they fail to distinguish between three distinct possibilities when it comes to modifying the treatment, measurement, and setting of an original experiment: (1) changing the value of a fixed factor (a first kind of extension); (2) resampling from the same population of treatments, measurements, and settings as in the original experiment (what I view as a genuine replication); and (3) sampling from a distinct population (a second kind of extension).

The usual notion of conceptual replication treats all modifications of treatment, measurement, and setting on the model of 1, leaving no room for

modifications along the lines of 2 and 3. It thus treats modifications of treatments differently from the modifications of experimental units: modifications of experimental units are rarely treated on the model of 1 in behavioral sciences (except in case studies). Rather, these modifications are treated on the model of 2 or 3: they either resample from the original population (what is usually called a “direct replication”) or sample from a different population, which is what happens when an experiment originally run on neurotypical individuals is again run on neuroatypical individuals.

Furthermore, because of the confusion built into the usual notion of conceptual replication, the function of modifying the treatment, measurement, and setting of an original experiment is misunderstood. Because a conceptual replication modifies the treatment and the measurement of an original experiment, in contrast to a direct replication, which only modifies the experimental units of the original experiment, psychologists often believe that it is meant to test the validity of the original experiment or the invariance range of the conclusion drawn on its basis. Rather, scientists are in fact testing the reliability of the original experiment with respect to the experimental components other than experimental units because they too can be random factors.

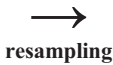
Perhaps I am uncharitably interpreting psychologists’ position about the modification of treatment, measurement, and setting. On an alternative interpretation, psychologists are aware that treatment, measurement, and setting could be random instead of fixed factors, but they do not treat them as random factors because they do not intend to generalize beyond the specific stimuli, measures, or settings of an experiment or because on their view it makes no sense to treat these as values of random factors. However, scientists often generalize beyond the original stimuli, measurements, and settings, and the concerns against treating these experimental components as random factors have already been addressed in section 2.3.

*3.5. A New Typology of Replications.* The lack of a general account of replication and the acceptance of the contrast between direct and conceptual replication has resulted in an unprincipled and unclear typology of replications (table 2). The usual typology of replications is unprincipled because a single type of replication corresponds to two distinct experimental components, while

TABLE 2. THE USUAL TYPOLOGY OF REPLICATIONS

Experimental Component	Replication
Experimental units	Exact
Treatment	Conceptual
Measurement	
Setting	

TABLE 3. THE REVISED TYPOLOGY OF REPLICATIONS

Experimental Component		Replication
Experimental units		Experimental units replication
Treatment		Treatment replication
Measurement		Measurement replication
Setting		Setting replication

another type of replication corresponds to experimental units; no replication corresponds to setting.<sup>11</sup> The usual typology is unclear because it does not specify what is required for a replication to occur. That is, the usual typology does not specify what a psychologist must do (resample, change the value of a fixed factor, etc.) to an experimental component for her experiment to count as a replication.

This typology contrasts with a principled, systematic typology of replications based on the Resampling Account (table 3). On this revised typology, a replication is a treatment replication if and only it resamples from a population of treatments. It can also resample from the population of experimental units (e.g., participants), in which case it would be an experimental units replication and a measurement replication (*mutatis mutandis* for the three other types).<sup>12</sup> A distinct type of replication corresponds to each experimental component.<sup>13</sup> Finally, the relation between the replication and the experimental component is made explicit, and it is the same for all the experimental components: resampling.

**4. The Debate about Direct and Conceptual Replications.** The debate about the best type of replication contrasts what is, according to the Resampling Account, a genuine type of replication (i.e., experimental units replication; usually called “direct replication”) and a notion that confuses extensions and replications. Rejecting the notion of conceptual replication, we must then clarify the terms of the debate and distinguish two conceptually distinct debates:

Debate 1: Experimental units versus treatment versus measurement versus setting replication

Debate 2: Replication versus extension

11. Schmidt’s (2009, 95) typology suffers from similar problems.

12. Thus, a measurement replication need not resample only from the population of stimuli.

13. Indeed, a virtue of this revised typology is to include a replication for settings, which was ignored in the usual typologies of replications.

Each debate compares comparable notions. Debate 1 asks whether it is more important to resample one of the four experimental components that have been distinguished in section 2 than the others. Debate 2 asks whether replication, as characterized by the Resampling Account, or extension is more important for science.

Let us now consider each debate in turn, starting with debate 1. The different types of replication are, everything else being equal, equally important to assess the reliability of experiments and to establish the reality and nature of phenomena. They complement one another rather than competing with one another. Psychologists pay more attention to experimental units replication, but this is only because they fail to acknowledge that the other experimental components can be random factors too. Admittedly, things are rarely equal: scientists sometimes have evidence that one of the experimental components is reliable. For instance, they could know that a scale provides a reliable measurement of a psychological trait (e.g., extraversion). In this case, resampling from a population of measures is less important than resampling from the other components. But absent this independent information, all experimental components matter for assessing the reliability of an experiment.

Turning to debate 2, replication and extension have different functions. Replications test the reliability of token experiments; extensions, their validity as well as the invariance range of a phenomenon. It is strange to think that there can be a meaningful comparison between these two goals. Consider a particular case: a psychologist obtains a particular result with Westerners as her explicitly stated population. Is it more important to replicate it with a different sample of Westerners so as to assess the reliability of the experiment, or is it more important to extend it to a different population of participants (e.g., people from East Asia or people from small-scale societies) so as to examine the invariance range of the finding? It is difficult to see how this question could be meaningful in the first place; there is no clear common measure to compare the importance of extension and replication.

The upshot of this discussion is that once we have a clear notion of replication, the debate about the importance of the different types of replication stops to make sense at all.

## 5. Two Objections

*5.1. Replication and Extension.* The difference between a replication and an extension hangs on whether the second experiment samples from the same population as the original experiment. Which population the original experiment sampled from is, however, often unclear and left unspecified, particularly for the experimental components other than the experimental units. Furthermore, it is a subjective matter in the sense that it depends on the intention of psychologists. The psychologist conducting

the original experiment intended to sample from a given more or less explicitly and precisely characterized population and to generalize to this population.

The first point—that we do not always know, perhaps even rarely know, what the relevant populations are—is not an objection to the Resampling Account of replication. It is rather a criticism of the actual practices in the behavioral sciences. Psychologists and others should make explicit the relevant populations experimental components are sampled from, as Simons, Shoda, and Lindsay (2017) have recently argued. Furthermore, the issue affects not only the Resampling Account of replication but also the usual notion of direct replication, since direct replications resample from an unchanged population of experimental units.

It is true that what population is sampled from and thus what experiment is conducted depends on the intention of the experimenter, but this cannot count as an objection against the Resampling Account. First, all actions are individuated by agents' intentions, and an experiment just is an action. Second, the subjective nature of the targeted population is a problem only if subjectivity is erroneously confused with privacy. Experimenters' intentions can be made explicit, even before running the experiment, as is done in preregistrations. The requirement of making explicit experimenters' intentions concerning the features of the experiments they intend to run applies as much to the populations they intend to sample from and generalize to as it does to the stopping rules for data collection. Thus, the Resampling Account provides a new argument for preregistration and extends the scope of the information to be provided in preregistrations.

*5.2. Vagueness and Controversies.* The Resampling Account does not enable psychologists to quell all possible controversies about whether an experiment replicates another one, for elements that are crucial for deciding whether an experiment replicates another one and whether replication is successful are not fixed in an objective, experimenter-independent manner. First, what population is sampled from and whether an experimental component is treated as a random or fixed factor depend on the often unknown intention of the experimenter. Second, whether experiment  $e_2$  replicates experiment  $e_1$  depends on whether the respective sequences of token events that are constitutive of  $e_1$  and  $e_2$  are of the same type, and as noted above events can be typified in more or less coarse ways. This leaves room for a psychologist to deny that  $e_2$  was genuinely a replication of  $e_1$  on the grounds that the experimental units, populations, treatments, and settings were not of the same type. Finally, whether experiment  $e_2$  successfully replicates experiment  $e_1$  depends on some individuation criterion for sameness of experimental outcomes (significant results, same  $p$ -values, same effect

sizes, etc.), something that is interest dependent too. This dependence leaves room for a psychologist to deny that  $e_2$  was a failed replication of  $e_1$ .

That intentions determine whether a second experiment replicates an original one is not a problem, as was just argued. Intentions can, and should, be made public in preregistration. Second, the Resampling Account does not solve all the controversies about whether a follow-up experiment is a genuine or a successful replication of a first experiment, but the point of this account was not to solve all controversies—just to provide a general account of what a replication is and to dissolve the controversy between proponents of direct and conceptual replications. Finally, no other account fares better when it comes to deciding whether two experiments are genuine replications (e.g., Schmidt 2009, 97–98). What the Resampling Account does, in contrast to these accounts, is identify precisely the two points from which controversies stem: (1) whether two experiments are instances of the same experiment type and (2) what the sameness of outcomes criteria are. Whether two experiments fall under the same types depends on how coarsely the constituting event types are individuated. Such individuation gives us criteria of identity. What makes two experimental outcomes the same depends on scientists' interests, and those can be made explicit.

**6. Conclusion.** A replication is an experiment that resamples from the populations targeted in an original experiment in order to assess its reliability. A distinct kind of replication corresponds to each of the four experimental components usually distinguished by psychologists: experimental units, treatments, measurements, and settings. This reveals that the traditional understanding of conceptual replication is confused, failing to distinguish different modifications of treatments, measurements, and settings. A better understanding of the notion of replication, as embodied in the Resampling Account, should thus lead us to abandon the usual distinction between direct and conceptual replication and thus any attempt to establish the epistemic superiority of one of them. Finally, the Resampling Account has some practical implications. Experimentalists should make explicit, possibly in preregistrations, whether the experimental components are fixed or random factors, and in the latter case they should describe the relevant populations.

#### REFERENCES

- Asendorpf, Jens B., et al. 2013. "Recommendations for Increasing Replicability in Psychology." *European Journal of Personality* 27:108–19.
- Bargh, John. 2012. "Priming Effects Replicate Just Fine, Thanks." *Psychology Today*, May 11. <http://www.psychologytoday.com/us/blog/the-natural-unconscious/201205/priming-effects-replicate-just-fine-thanks>.

- Baumeister, Roy F., Dianne M. Tice, and Kathleen D. Vohs. 2018. "The Strength Model of Self-Regulation: Conclusions from the Second Decade of Willpower Research." *Perspectives on Psychological Science* 13:141–45.
- Begley, C. Glenn, and Lee M. Ellis. 2012. "Drug Development: Raise Standards for Preclinical Cancer Research." *Nature* 483:531–33.
- Bogen, James, and James Woodward. 1988. "Saving the Phenomena." *Philosophical Review* 97:303–52.
- Bonett, Douglas G. 2012. "Replication-Extension Studies." *Current Directions in Psychological Science* 21:409–12.
- Camerer, Colin F., et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351:1433–36.
- Cesario, Joseph. 2014. "Priming, Replication, and the Hardest Science." *Perspectives on Psychological Science* 9:40–48.
- Chambers, Christopher D. 2017. *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton, NJ: Princeton University Press.
- Colaço, David. 2019. "An Investigation of Scientific Phenomena." PhD diss., University of Pittsburgh.
- Collins, Harry M. 1985. *Changing Order*. London: Sage.
- Crandall, Christian S., and Jeffrey W. Sherman. 2016. "On the Scientific Superiority of Conceptual Replications for Scientific Progress." *Journal of Experimental Social Psychology* 66:93–99.
- Gigerenzer, Gerd, and Ulrich Hoffrage. 1995. "How to Improve Bayesian Reasoning without Instruction: Frequency Formats." *Psychological Review* 102:684–704.
- Gilbert, Daniel T., Gary King, Stephen Pettigrew, and Timothy Wilson. 2016. "Comment on 'Estimating the Reproducibility of Psychological Science.'" *Science* 351:1037.
- Hacking, Ian. 1992. "The Self-Vindication of the Laboratory Sciences." In *Science as Practice and Culture*, ed. Andrew Pickering, 29–64. Chicago: University of Chicago Press.
- Hüffmeier, Joachim, Jens Mazei, and Thomas Schultze. 2016. "Reconceptualizing Replication as a Sequence of Different Studies: A Replication Typology." *Journal of Experimental Social Psychology* 66:81–92.
- Judd, Charles M., Jacob Westfall, and David A. Kenny. 2012. "Treating Stimuli as a Random Factor in Social Psychology: A New and Comprehensive Solution to a Pervasive but Largely Ignored Problem." *Journal of Personality and Social Psychology* 103:54–69.
- Kenny, David A. 1985. "Quantitative Methods for Social Psychology." In *Handbook of Social Psychology*, vol. 1, ed. Lindzey Gardner and Elliot Aronson, 487–508. New York: Random House.
- Lovett, Adam, and Kevin Munger. 2019. "Validity, Prediction and the Problem of Replicability." Unpublished manuscript, Center for Open Science. <https://osf.io/yzghn/>.
- Lynch, John G., Jr., Eric T. Bradlow, Joel C. Huber, and Donald R. Lehmann. 2015. "Reflections on the Replication Corner: In Praise of Conceptual Replications." *International Journal of Research in Marketing* 32:333–42.
- Machery, Edouard. 2008. "The Folk Concept of Intentional Action: Philosophical and Experimental Issues." *Mind and Language* 23:165–89.
- . 2017. *Philosophy within Its Proper Bounds*. Oxford: Oxford University Press.
- Machery, Edouard, and John M. Doris. 2017. "An Open Letter to Our Students: Doing Interdisciplinary Moral Psychology." In *Moral Psychology: A Multidisciplinary Guide*, ed. Benjamin Voyer and Tor Tarantola, 119–43. Berlin: Springer.
- Machery, Edouard, and Tiziana Zalla. 2015. "The Concept of Intentional Action in High Functioning Autism." *Oxford Studies in Experimental Philosophy* 1:152–72.
- McAllister, James W. 1997. "Phenomena and Patterns in Data Sets." *Erkenntnis* 47:217–28.
- Nosek, Brian. 2016. "Let's Not Mischaracterize Replication Studies: Authors." Retraction Watch. <http://retractionwatch.com/2016/03/07/lets-not-mischaracterize-replication-studies-authors/>.
- Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349:1422–25.
- Pashler, Harold, and Christine R. Harris. 2012. "Is the Replicability Crisis Overblown? Three Arguments Examined." *Perspectives on Psychological Science* 7:531–36.



- Richter, Martin L., and Mary B. Seay. 1987. "ANOVA Designs with Subjects and Stimuli as Random Effects: Applications to Prototype Effect on Recognition Memory." *Journal of Personality and Social Psychology* 53:470–80.
- Romero, Felipe. 2016. "Can the Behavioral Sciences Self-Correct? A Social Epistemic Study." *Studies in History and Philosophy of Science A* 60:55–69.
- Schmidt, Stefan. 2009. "Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences." *Review of General Psychology* 13:90–100.
- . 2017. "Replication." In *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research*, ed. Matthew C. Makel and Jonathan A. Plucker, 233–53. Washington, DC: American Psychological Association.
- Schnall, Simone. 2014. "Moral Intuitions, Replication, and the Scientific Study of Human Nature." *Edge*, November 18. [https://www.edge.org/conversation/simone\\_schnall-simone-schnall-moral-intuitions-replication-and-the-scientific-study-of](https://www.edge.org/conversation/simone_schnall-simone-schnall-moral-intuitions-replication-and-the-scientific-study-of).
- Scoville, William B., and Brenda Milner. 1957. "Loss of Recent Memory after Bilateral Hippocampal Lesions." *Journal of Neurology, Neurosurgery, and Psychiatry* 20:11–21.
- Shnabel, Nurit, and Arie Nadler. 2008. "A Needs-Based Model of Reconciliation: Satisfying the Differential Emotional Needs of Victim and Perpetrator as a Key to Promoting Reconciliation." *Journal of Personality and Social Psychology* 94:116–32.
- Simons, Daniel J. 2014. "The Value of Direct Replication." *Perspectives on Psychological Science* 9:76–80.
- Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay. 2017. "Constraints on Generality (COG): A Proposed Addition to All Empirical Papers." *Perspectives on Psychological Science* 12:1123–28.
- Squire, Larry R. 2009. "The Legacy of Patient HM for Neuroscience." *Neuron* 61:6–9.
- Stroebe, Wolfgang, and Fritz Strack. 2014. "The Alleged Crisis and the Illusion of Exact Replication." *Perspectives on Psychological Science* 9:59–71.
- Tversky, Amos, and Daniel Kahneman. 1982. "Evidential Impact of Base Rates." In *Judgment under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky, 153–60. Cambridge: Cambridge University Press.
- Wells, Gary L., and Paul D. Windschitl. 1999. "Stimulus Sampling and Social Psychological Experimentation." *Personality and Social Psychology Bulletin* 25:1115–25.
- Wimsatt, William C. 2007. *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.