Discussion: What Is a Replication?

Samuel C. Fletcher*, Galin Jones*, & Alexander Rothman*

Abstract: Machery (2020) has recently proposed a "resampling" account of experimental replication to dissolve a debate in psychology about the relative merits of direct and conceptual replication. We argue that (i) on matters of replication's function and typology, the resampling account is not substantially different from the functional account of replication extant in the literature; (ii) on what generalizations can be drawn from replications, the resampling account is too restrictive and relies on a misunderstanding of the relation between random sampling and generalizability; and (iii) Machery's reading of the debate on the relative importance of direct and conceptual replication elides a deeper debate about values and the distribution of research resources in science.

1 Introduction: What Is a Replication?

Edouard Machery (2020) has recently proposed in this journal an account of what it means for one experiment to be a replication of another:

> 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. (556)

The "experimental components" include the experimental units studied, the treatments to (or independent variables for) those units manipulated in the study, the measurements of (or dependent variables for) those units, and the settings of the study, which include any relevant contextual information (550–551). On his account,

> 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. (555)

To treat a component as a random factor is, on Machery's account, to model it statistically as a random sample from a population, whether that population consists of units, treatments, measurements, and so on (552).

Besides contrasting his "Resampling" account with the "Functional" account of replication (Schmidt 2009; 2017), Machery explores how accepting his account bears on an ongoing debate within psychology about the relative value of direct and conceptual replication. Roughly speaking, a direct replication is an experiment that narrowly repeats the experimental procedure of another along all the dimensions that matter scientifically. A conceptual replication, in contrast, is an experiment that tests or explores the same hypotheses or research questions

*University of Minnesota, Twin Cities

using different methods. (We shall say more about these types in section 2.) In the aforementioned debate, some psychologists seem to suggest that one type of replication is more valuable or informative than the other. Machery in opposition argues that the debate is ill-founded, for on his account conceptual replication is not a sort of replication at all (545, 547, 563).

We agree with much about Machery's "Resampling" account of replication and applaud his goal to address the debate about the relative value of direct and conceptual replication. We also believe, however, that typical readers of *Philosophy of Science* may not be aware of the (recent) theoretical context in which Machery's essay attempts to intervene, context which could well affect how one should evaluate the details of his arguments and goals. So, in the present discussion, we'd like to add some of this context, drawing on our experience across statistics, methodology in psychology, and the philosophy of experiment.

In what follows, we argue in particular for three main points. First, in section 2, we review the functional account of replication in a bit more depth to show that, despite Machery's presentation, the resampling account is very similar to it. Both accounts in fact yield the same replication typologies. The main difference between them, Machery alleges, is that the functional account does not permit treatments, measurement, and settings to be random factors. Depending on how one understands the functional account, either this difference is illusory or it can be interpreted as adding compatible further detail to the functional account. We conclude that, at least as far as matters of function, sampling, and classification are concerned, any remaining dispute between the functional and resampling accounts is a merely verbal matter of emphasis or foregrounding.

Second, there is a further, more substantive difference between the two accounts that Machery does not highlight. It concerns the populations from which one can draw conclusions with experiments, which Machery insists are only those whose members are treated as random factors in the statistical model for the experimental data. The functional account does not impose this requirement. We show in section 3 that while the requirement has been a part of statistical lore, it is not and has never been supported by statistical theory. Machery is correct that the inferences we can draw from an experiment critically depend on how the features of an experiment are selected, but treating a variable as a random factor is neither necessary nor sufficient to justify inferences about a population of which it is a part.

Third, when this difference is set aside, the functional account fares equally well in achieving Machery's goal to dissolve the debate about the relative value of direct and conceptual replication, if construed as a debate about which is a more valuable *type of replication*. But as we outline in section 4, there is a deeper issue that we believe Machery's depiction has elided, namely, a debate about which *balance* of activities is better for the progress of science. That debate involves consideration of the values of different types of research and cannot be dissolved by a purely methodological investigation.

2 The Resampling Account vs. the Functional Account

The resampling account of replication foregrounds specification of the populations which replications share with the studies they replicate. The functional account of replication foregrounds instead the purposes in scientific inquiry towards which investigators perform

replications (Schmidt 2009, 2017). Fletcher (2021, sec. 2) identifies that these purposes or roles are unified in excluding various sorts of underdetermination in the conclusions of an experiment, ones regarding the theory or hypothesis tested or regarding auxiliary hypotheses concerning the experiment:

1. They are not due to mistakes in the data analysis.
2. They are not due to sampling error.
3. They do not depend on contextual factors, according to the theory or hypothesis tested.
4. They do not arise from fraud or questionable research practices.
5. They generalize, according to the theory or hypothesis tested, to a larger or different population than that sampled in the original.
6. Their aspects pertaining to the theoretical hypothesis of interest hold even when that hypothesis is operationalized or tested in completely different ways.

A replication directed towards one or more of these roles repeats an original experiment by keeping fixed the latter's experimental procedure, inasmuch as is possible and according to a background's theory's specification of what differences are and are not relevant, except for one (or more) of the following four classes of variables:

A. the procedures for constituting the independent variables, the ones whose explanatory, predictive, or controlling features are under test;
B. the study's context, i.e., possible moderators such as the properties and history of the research units and the people running the study, their relevant historical or cultural context, and the physical setting of the study and its material realization;
C. the procedures for the selection and allocation of the research units; and
D. the procedures for constituting the dependent variables, the ones to be explained, predicted, or controlled.

For instance, to test function 2, all classes of variables would be held constant except for the particular research unit selected. That would constitute a resampling from the relevant populations. (See Schmidt (2009, 93–95; 2017, 239–241) for more on which variables should vary for a replication to perform which function.) Function 1 is called methodological replication, since it can be performed using the same methods (or at least what the original experiment reported) without a new data set; functions 2–5 are species of direct replication; and function 6 corresponds to conceptual replication.

How does the functional account compare with Machery's resampling account? Both agree that one can separate one function for (direct) replication from conceptual replication/extension. However, the functional account provides a more fine-grained typology of the sources of experimental reliability, in the sense that Machery describes above. While Machery correctly identifies a useful notion of experimental reliability that appeals to resampling with "everything else being kept constant" (555) and acknowledges the subtlety of this qualifier (558), he does not distinguish between the different ways in or dimensions along which reliability could fail, as given by functions 2–5. Nevertheless, distinguishing between these different ways seems to be compatible with the rest of the resampling account.

Both accounts also agree that (direct) replication involves resampling from relevant populations. However, the resampling account provides a more fine-grained typology of components that can be resampled, dividing them into subjects, treatments, measurements, and settings. In light of this, Machery surprisingly concludes, "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" (560). Moreover,

> The usual typology of replications [such as that found in Schmidt (2009)] is unprincipled because a single type of replication [conceptual] corresponds to two distinct experimental components [treatment and measurement], while another type of replication [direct] corresponds to experimental units; no replication corresponds to setting. (561–562)

Should Machery therefore also accept, for analogous reasons, that the functional account reveals that his resampling account's notions of replication and extension are "confused," yielding an "unprincipled" typology? For, as we adumbrated just above, the resampling account's notions of replication and extension do not distinguish between many of the distinct functions that these activities fulfill. We propose rather that both inferences should be rejected as invalid. No notion is "confused" or "unprincipled" simply because it is coarse-grained.

We anticipate two objections to this proposal. First, what of the claim that the functional account's typology has no item which varies or samples from the possible settings of an experiment? If true, it would require us to qualify our claim that the functional account simply has a coarse-grained specification of which populations are sampled in a replication. Now, for Machery, "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" (551). This is however included in the study's *context*, as specified in the above variables for the functional account, which in fact Schmidt proposed should be varied in replications for functions 3, 4, and 6 (2009, 94–95). So this objection is mistaken.

Second, Machery notes that "Schmidt limits sampling error to the selection of participants ([2009,] 93), failing to acknowledge that all the experimental components can be random factors and failing to treat all the experimental components similarly" (560). He is right that Schmidt only explicitly mentions participants when it comes to sampling error. But we do not interpret that as a claim that other components *cannot* incur sampling error; it only reflects the rarity of random sampling from populations of treatments, measurements, and settings.[1] In any case, an advocate of the functional account could accept Machery's observation as a friendly amendment without changing any other aspect of the account at all.

Likewise, the resampling account can adopt without any other changes the functional account's finer distinctions about the purposes and goals of replication (and perhaps replication-adjacent) activities. When each takes the other's finer distinctions into account, the classifications of replication (and perhaps replication-adjacent) activities they describe,

[1] There is some textual evidence that Schmidt acknowledges that these other components can vary. He writes that "The most important class [of variables] is termed *primary information focus*. This construct [theoretical concept] describes the instructions, materials, and events that create a certain stimulus complex for the participant. It is designed by the researcher with respect to the hypothesis that is investigated and will be varied accordingly" (Schmidt 2009, 93). However, Machery (2020) deserves credit for making explicit which such components can be varied or sampled.

delineated by populations and functions, is the same. They are merely notational variants of one another, the one foregrounding one dimension of that classification over another. Any further dispute about such foreground seems to us to be merely verbal (Jenkins 2014).

3 Statistical Generalization: Fixed and Random Factors

Another difference between the resampling and functional accounts of replication pertains not to their replication typologies, but to their accounts of a replication's inductive use—how replications justify statistical generalizations from the samples they measure to populations of interest. Machery distinguishes between "fixed" and "random" experimental components—subjects, treatments, measurements, and settings—and asserts that every component is either fixed or random.

> If they are random factors, their levels are randomly sampled from a population … [and because of this,] they can represent this population more or less accurately, and sampling error must be taken into account. By contrast, if the levels of experimental components are fixed factors, the levels used in an experiment exhaust the relevant population. … The experimenter does not aim [in this case] to generalize statistically to unobserved levels; rather, she limits her conclusion to the observed levels. (Machery 2020, 552)

He draws a tight connection between this distinction and generalizability: "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" (553)—cf. Cronbach, Rajaratnam, and Gleser (1963, 147, 160). Since the resampling account requires that a replication of an original experiment resamples the factors in the original which were random (556), it constrains substantially which experiments can be replications.

While much of Machery's description parallels that of the theory of statistical design of experiments (Oehlert 2000, chs. 1, 11), it diverges in the claims that random factors both *must* have levels randomly sampled from a population and are *required* for valid generalization. By contrast, the functional account makes no such tight connections. Indeed, the functional account comports more closely with the theory of statistical experimental design in affirming that treating an experimental component as a random factor (i) does not entail random sampling and is (ii) neither sufficient (iii) nor necessary for statistical generalizability. Rather, that theory requires merely that a sample be representative of the feature of the population about which one wishes to generalize. In what follows, we argue for these three points and illustrate them with simple examples.

In the statistical design of experiments, random factors can arise in two ways. One way involves random sampling of the levels of a factor. However, this is somewhat uncommon in practice, in part because it requires both access to the population and a process for generating a random sample therefrom. The other way is the formal use of random factors to induce certain properties in the statistical model so that it might better fit the data by accounting for features of the experiment. An example of this involves modeling correlations that would be expected from repeated measures on the same experimental unit.

For instance, animal breeding experiments often use the sires at hand, rather than random samples. Even so, the sires can be represented as either fixed or random factors in the statistical model. In line with Machery's description, it is common to consider them fixed if the goal is to make inferences about this particular set of sires, such as which sires lead to larger offspring or resistance to disease. But it is equally legitimate for this purpose to treat them as random to account for breeding the sires multiple times to various dams, for example. One can also treat them as random if the goal is to generalize to a larger population. However, the validity of any such generalization depends critically on whether the sires used are representative of that population. This is so even if the sires *were* randomly sampled. Regardless, if they are not representative, treating the sires as a random factor cannot compensate for this limitation. This shows that treating an experimental component as a random factor is not sufficient for generalizability.

It is also not necessary for generalizability. Consider the problem of predicting the cost of property damage caused by Atlantic hurricanes. Using our understanding of which properties of the hurricanes may be relevant (e.g., size and wind speed), we might use standard approaches to constructing statistical models (e.g., regression) that predict the cost of damage due to hurricanes next year, or in a past year that was not well observed. This is so even though hurricane year was not randomly sampled and was treated as a fixed factor. It is rather the representative nature of the experimental units (observed hurricanes) and our understanding of how their properties relate (e.g., size and property damage cost) that undergirds our valid generalizations.

That representing any experimental component—not just subjects—as a random factor is neither sufficient nor necessary for generalizability is important for understanding methodology in the social sciences, where researchers are especially interested in generalizing to populations of manipulations, measurements, and contexts (Shadish, Cook, and Campbell 2002). The first two, when valid, are often termed "construct validity" (Cronbach and Meehl 1955), and the third "external validity." Efforts to establish construct validity benefit from repeated experiments shaped by the theoretical assumptions that underlie the phenomenon of interest and the use of strategies such as multiple operationism (Webb et al. 1966) or heterogeneous irrelevancies (Brunswick 1956) to test those assumptions. However, neither of these approaches relies on a random sampling of manipulations or measures or treats them as random factors; instead, they proceed from theories of the phenomena and constructs of interest. Efforts to establish external validity may rely on sampling, but purposive rather than random. To determine the manner in which the causal effect of a manipulation holds across populations, settings, and time, one needs to purposely examine the effect across sources of variation within a given domain. Once again, theoretical assumptions are critical in specifying the relevant sources of variation that should be examined and provide a framework within which inferences regarding generalizability can be drawn from an experiment or set of experiments.[2] Treating a set of experimental components as a random factor does not somehow transform the properties of a set of observations into a representative sample of the population of interest.

So, there is good reason generally and in the social sciences particularly to reject Machery's claimed constraints on statistical generalization that restrict it to random factors,

[2] Purposive sampling is sometimes called "model-based" sampling for this reason. See Zhao (2020) for a recent philosophical defense of this idea.

items (treated as) randomly sampled from a population. But in doing so, one rejects the only other substantive difference between the functional and resampling accounts of replication. Nevertheless, there is still considerable value in his call for investigators to consider the manner in which *all* the features of an experiment—subjects, treatments, measures, and context—are each representative of a broader class and for investigators to utilize innovations such as pre-registrations to specify, before conducting the experiment, which class each component is meant to represent. In this regard, he is echoing prior discussions of construct validity and generalizability in the social sciences (Cronbach, Rajaratnam, and Gleser 1963; Albright and Malloy 2000; Shadish, Cook, and Campbell 2002). Investigators routinely think about how the subjects enrolled in an experiment are representative of a broader population, but are less likely to reflect on whether or how the measures, treatments, or context are so representative. Limited attention to these aspects of representativeness is one of the challenges investigators face when pursuing questions regarding replication.

4 The Deeper Conflict between Direct and Conceptual Replication

Recall that Machery developed his resampling account of replication in part to diagnose and dissolve a debate about the relative merit of direct and conceptual replication—in his terms, replication (simpliciter) and extension.[3] The dissolution comes from recognizing that "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" (Machery 2020, 563). It is no surprise that, as we have argued in section 2, the functional account also distinguishes these goals (cf. Schmidt 2017, 238). So, if the debate Machery describes is about the relative merit of these activities for the *same* goals, then the functional account comes to the same conclusion, that the debate should be dissolved.

Some parties to the debate do seem to have construed the debate in these terms (e.g., Cesario 2014; Simons 2014; Stroebe and Strack 2014), but a closer reading of other texts reveals a deeper issue about the *relative resources* that psychology as a scientific discipline should invest in (direct) replication and extension (conceptual replication). The meaningful comparison between these goals is therefore in terms of what balance of emphasis should be placed on what these differing activities accomplish. Illustrating this, Crandall and Sherman (2016, sec. 3) acknowledge that direct and conceptual replications (replications and extensions) have different functions, hence are toward different goals. They write, further, that

> There is a broad consensus in favor of robust findings, for reliability in the scientific record, for high quality research with dependable reporting and replicability, and for progress in scientific knowledge. But there are sharp differences among scientists in (1) which scientific

[3] Machery (2020, 562–3) suggests that another way of reading the debate is about the relative merit of replications that resample from different components of an experiment, such as treatments and measurements. However, as he himself acknowledges (553), psychologists rarely consider levels of these factors as sampled from a population, and we are not aware of any textual evidence that the parties to the debate in question are different.

> goals should take priority over others and (2) the best way to meet those respective goals. (2016, 93)
>
> […]
>
> There is no controversy over the need for replication; virtually all scientists and philosophers of science endorse the notion that replication of one sort or another is absolutely essential. The controversy is largely over the degree to which different kinds of replications advance scientific knowledge. (2016, 94)

They go on to argue that tipping the balance of resources in favor of conceptual replications over direct replications would facilitate scientific progress.

In the present discussion, we remain agnostic about the right balance, although we are sympathetic to the idea that the balance of resources should be attuned to the values of a scientific community. It is plausible that the replication crisis, qua crisis, arose because of stark realization of the misalignment between these values and the balance of resources in psychology. Philosophers of science are increasingly aware of this issue and the challenges it poses to the reward structure of science, which may not promote these values (Romero 2017). This deeper debate about values and how science should be organized cannot be dissolved directly by stipulating a definition of what a replication is.

References

Albright, Linda, and Thomas E. Malloy. 2000. “Experimental Validity: Brunswik, Campbell, Cronbach, and Enduring Issues.” *Review of General Psychology* 4.4: 337–53.

Brunswik, Egon. 1956. *Perception and the representative design of psychological experiments*. 2nd ed. Berkeley: University of California Press.

Cesario, Joseph. 2014. “Priming, Replication, and the Hardest Science.” *Perspectives on Psychological Science* 9.1: 40–48.

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.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct validity in psychological tests." *Psychological bulletin* 52.4: 281–302.

Cronbach, Lee J., Nageswari Rajaratnam, and Goldine C. Gleser. 1963. “Theory of generalizability: A liberalization of reliability theory.” *British Journal of Statistical Psychology* 16.2: 137–163.

Fletcher, Samuel C. 2021. “The role of replication in psychological science.” *European Journal for Philosophy of Science* 11.23: 1–19.

Jenkins, C. S. I. 2014. “Merely Verbal Disputes.” *Erkenntnis* 79.S1: 11–30.

Machery, Edouard. 2020. “What is a replication?” *Philosophy of Science* 87.4: 545–567.

Oehlert, Gary W. 2000. *A First Course in Design and Analysis of Experiments*. New York: W. H. Freeman.

Romero, Felipe. 2017. “Novelty versus replicability: Virtues and vices in the reward system of science.” *Philosophy of Science* 84.5: 1031–1043.

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.

———. 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.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Simons, Daniel J. 2014. “The Value of Direct Replication.” *Perspectives on Psychological Science* 9.1:76–80.

Stroebe, Wolfgang, and Fritz Strack. 2014. “The Alleged Crisis and the Illusion of Exact Replication.” *Perspectives on Psychological Science* 9.1: 59–71.

Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest. 1956. *Unobtrusive measures: nonreactive research in the social sciences*. Chicago: Rand McNally.

Zhao, Kino. 2020. “Sample representation in the social sciences.” *Synthese* forthcoming:1-19.