

Preregistration, Replication, and Nonexperimental Studies

September 09, 2016



In last month's column, I worried about whether encouraging us to preregister our hypotheses and analysis plan before running studies would stifle discovery. I came to the conclusion that it needn't — but that we need to guard against letting the practice run away with itself. In this column, I take up a second concern about preregistration: That it seems to apply only to certain types of studies, and thus runs the risk of marginalizing studies for which preregistration is less fitting.

Preregistration is designed to ensure that if the data we collect confirm our hypotheses, those hypotheses were the ones we intended to test before the study began — and not new hypotheses we're generating based on what we're observing. If we are seeing patterns for the first time, we need to make it clear to ourselves, and to our readers, that the study is generating new hypotheses rather than testing old ones. In a sense, preregistration entails replication (at least conceptual replication if not exact replication; Crandall & Sherman, 2015), since the preregistered hypothesis-testing study rests on the foundations constructed on the basis of earlier hypothesis-generating studies.

Preregistration and replication lend themselves well to short-term experimental studies conducted on participants who are easy to find. But it's just too costly or unwieldy to generate hypotheses on one sample and test them on another when, for example, we're conducting a large field study or testing hard-to-find participants. Do we have to give up on the hope of replication and robustness for this type of study? There are two reasons not to despair.

First, some kinds of studies, by their nature, may be more robust than others. As Jon K. Maner (2015) notes, studies conducted in the field have two advantages over lab studies. The first advantage is obvious: The findings of a field study have clear relevance to the real world. The second advantage is less obvious: It is difficult to control all, or even many, of the variables in a field study. Why is this lack of control a good thing? If a phenomenon is discoverable under these messy conditions, it is likely to be a robust one that is worthy of explanation. Jean Piaget's discoveries, which were made at home on his three infants, are a good example. Although his sample was small and therefore obviously not

representative, the conditions under which Piaget made his observations varied extensively from trial to trial. Having a large number of naturalistic observations on a small number of participants can lead to robustness. In 1973, Roger Brown made his initial discoveries about language learning also by studying only three children at home talking about random topics. Piaget's observations have stood the test of time, in part because he was a brilliant observer who could zero in on invariances that mattered, and in part because his observations came from a range of situations and thus were less likely to depend on the details of any one of those situations. Happily, this means that in areas where it is difficult to repeat a study, exact replication may not be essential in ensuring a phenomenon's robustness.

The second reason not to despair is that there can be, and often is, replication built into observational studies — it just doesn't get reported as such. For example, we can develop a coding system on the basis of a subset of the data, establish the reliability of the coding system, and then apply that system to the rest of the data (e.g., Goldin-Meadow & Mylander, 1991, pp. 322–324). This procedure allows us to discover hypotheses on one part of the data and test them on another part, a type of replication that can be conducted on populations that are rare or exist in difficult-to-recreate conditions.

Discovering the right coding system (i.e., the coding system that captures what's interesting about the data) is analogous to piloting an experimental study to find the right parameters to reveal the phenomenon. Neither procedure is cheating — it's the discovery part of science. But perhaps researchers should be encouraged to report these steps, along with the details of the coding system in an observational study (which is typically the heart of this type of study), in supplementary materials. Doing so could save others a great deal of time and, more importantly, could provide a preliminary sense of the boundary conditions under which the phenomenon does, and does not, hold.

There is currently an effort to raise the status of replication in experimental studies and devote some of our precious journal space to making sure a phenomenon is robust across labs (e.g., Nosek & Lakens, 2014). These efforts seem reasonable to me as long as they do not become exercises in fault-finding but are seen as what they are — ways to test the robustness and generality of a phenomenon. Barring intentional fraud, every finding is an accurate description of the sample on which it was run. The question — an important one — is whether the findings extend beyond the sample and its particular experimental conditions. If we're going to take replication seriously in experimental studies, then I suggest we do the same for studies that use other methods. For example, when using observational methods, researchers can be encouraged not only to report iterative tests of a coding system on a single sample but also to recognize these tests as the replications that they are.

What we don't want to do is require that the procedures used to ensure robustness and generalizability in experimental studies (e.g., preregistration, multiple-group replications of a single study) be applied to all types of psychological studies, and then devalue or marginalize the studies for which the preregistration procedures don't fit. Rather, we need to think creatively about how to achieve robustness for the wide range of methods that comprise the richness of psychological studies. α

References

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Crandall, C. S., & Sherman, J. W. (2015). On the scientific superiority of conceptual replications for

scientific progress. *Journal of Experimental Social Psychology*, 66, 93–99.
doi:10.1016/j.jesp.2015.10.002

Goldin-Meadow, S., & Mylander, C. (1991). Levels of structure in a communication system developed without a language model. In K. R. Gibson & A. C. Peterson (Eds.), *Brain maturation and cognitive development: Comparative and cross-cultural perspectives* (pp. 315–344). New York, NY: Aldine de Gruyter.

Maner, J. K. (2015). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*, 66, 100–106. doi:10.1016/j.jesp.2015.09.018

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45, 137–141.

Piaget, J. (1952). *The origins of intelligence in children* (translated by M. Cook). New York, NY: International Universities Press, Inc.