RESEARCH METHODS
IN THE SOCIAL SCIENCES

AN A-Z OF KEY CONCEPTS

Falsification: How does it relate to reproducibility?

Brian D. Earp
Yale University

This is the author's copy of published handbook chapter. Please cite as follows:

Earp, B. D. (2020). Falsification: How does it relate to reproducibility? In J.-F. Morin, C. Olsson, & E. O. Atikcan (Eds.), *Research Methods in the Social Sciences: An A-Z of Key Concepts* (pp. 119-123). Oxford: Oxford University Press..

Introduction

A common way to distinguish science from pseudoscience is that the former puts forward **HYPOTHESES** that are at least in principle *falsifiable* (Popper 1959). This means that they are capable of being refuted—or falsified—by counter-evidence. To illustrate, astrology is often characterized as a pseudoscience since its predictions are so vague that any number of occurrences could be interpreted as confirmatory. Astronomy, by contrast, is usually seen as scientific in part because it makes extremely precise and hence 'risky' predictions: the predictions could turn out to be false, which would require astronomers to update their theories in light of the evidence.

In reality, matters are rarely so clear. One can always question the validity of the seemingly contrary evidence rather than change one's theory, for example, and this is often justified. Accordingly, contemporary philosophers of science tend to look down on falsifiability as overly simplistic. Nevertheless, among many practicing scientists, the notion is still regarded as a useful—if imperfect—heuristic for judging the strength of a **HYPOTHESIS** in terms of its ability to generate new insights when combined with careful observation (Earp and Trafimow 2015).

Consider the **HYPOTHESIS** that all swans are white (H1). H1 is falsifiable because a single valid observation of a non-white swan would show H1 to be wrong. This, in turn, would teach us something new about the world (not all swans are white). Now consider the hypothesis that all

swans are either white or some other color (H2). H2, while true, is not falsifiable because it could never be shown wrong no matter how many swans were observed. It renders swan observation uninformative. The lesson for researchers is to try to formulate hypotheses that could turn out to be wrong, and to test those hypotheses in such a way that, if they are wrong, this could be legitimately inferred from the data.

As can be seen with H1, some falsifiable hypotheses can be tested by making simple observations. Others must be tested by experiment, for example, the **HYPOTHESIS** that a given treatment reduces the symptoms of a disease. Based on the principle of falsifiability, designing a convincing experiment requires asking: "What would count against my hypothesis?" If an experiment is carefully designed to produce such contrary evidence, and yet reliably fails to produce it, the researcher will have good reason to place provisional confidence in the soundness of the hypothesis: it has not yet been refuted. The more such experiments that fail to falsify the hypothesis, the more robust the evidence amassed in its favor (Lebel et al. 2017).

Falsification and self-correction

Falsification also relates to *self-correction* in science. Often, erroneous findings make their way into the literature. If subsequent researchers conduct the same experiment as the original and yet fail to yield the same finding, they are often described as having 'falsified' (that is, shown to be incorrect) the original result. In this way, mistakes, false alarms, and other non-reproducible output is thought to be identifiable and thus able to be corrected.

But this account is far too simple. First, actual attempts at falsification are relatively uncommon. At best, scientists (typically) conduct what are sometimes called 'conceptual' **REPLICATIONS**: essentially, *variations* on previously published experiments designed to extend existing findings or to generalize across new conditions or methodologies. Such quasi-replication, however, is not sufficient to confirm—or falsify—original results. Put simply, if a conceptual replication 'fails' it could be due to whatever change was made in the materials or methodology, as opposed to any problems with the original experiment or finding (Doyen et al. 2014). Negative results from such experiments are therefore inconclusive. For self-correction of science through falsification, then, what is needed are 'direct' replications.

Direct replication

'Direct' **REPLICATIONS** are those that (attempt to) follow the original study as closely as possible. Some differences are unavoidable. Ideally, however, the replication is sufficiently similar (SS) to the original along all theoretically relevant dimensions—from design, to method, materials, **SAMPLING TECHNIQUE** and so on—that the results from the two experiments can be meaningfully compared.

Sometimes, it is hard to determine if the SS condition has been met. Especially in younger fields such as psychology, it is often the case that certain *auxiliary assumptions* have not yet been exhaustively worked out. An auxiliary assumption is simply a logical or practical assumption needed to link an 'abstract' theory or **HYPOTHESIS** to something directly observable, such as a response-time score or a check-mark on a scale. One or more such assumptions may be (inadvertently) violated in a **REPLICATION** study, even if it is designed to be 'direct.'

Suppose that such a study yields a null finding, whereas the original finding was 'statistically significant' (see **STATISTICAL SIGNIFICANCE**). This would not entail that the original result (or the **HYPOTHESIS** the predicted it) had been falsified. Rather, follow-up studies would be needed to narrow in on where in the chain of auxiliary assumptions a break-down or violation may have occurred. In this way, if conducted systematically, **REPLICATIONS** can be useful for identifying and making explicit unknown or unacknowledged auxiliary assumptions, which can help with hypothesis-specification and theory advancement (Cesario 2014).

Such interpretational ambiguities complicate the project of falsification. In fact, **REPLICATION** results can never definitively prove or disprove (falsify) a reported effect: alternative explanations are always in theory possible. They can, however, be *informative* about its likely existence, magnitude, and theoretical and practical implications. In particular, they can give all-things-considered good reasons to adjust one's confidence in the reliability of the effect upward or downward, as a function both of the quality and number of replication attempts (Earp and Trafimow 2015).

The importance of auxiliary assumptions: an example

Consider a reported failure to replicate a famous study in which participants exposed to an elderly priming stimulus subsequently walked more slowly down a hallway compared to a control group, as measured by a stopwatch (Bargh, Chen and Burrows 1996). In the **REPLICATION** study, an infrared sensor was used instead of a stopwatch, and the slow-walking effect seemed to disappear (Doyen et al. 2012). Since a change was made, one could argue that the replication attempt was only 'conceptual' and therefore not even potentially falsifying. But that would be incorrect. The change was based on a sound auxiliary assumption: namely, that a time-measure that is less susceptible to human error would be superior to one that is more susceptible to such error. In other words, the change allowed for a *better* test of the original **HYPOTHESIS**.

Assuming that there were no other relevant differences between the original and **REPLICATION**, it would be reasonable to interpret the observed null finding as counting at least somewhat against the validity of the original result. The more the null finding, itself, is replicated in follow-up experiments—so long as they are competently executed and do not violate any further auxiliary assumptions—the more the total body of evidence will support the conclusion that the original result should not be relied upon (Earp and Trafimow 2015).

But now consider another change that was made. The priming materials were translated into French. Apparently, the replicating team assumed that the language used for a priming study is irrelevant to the outcome. But *this* auxiliary assumption may be mistaken. Based on a corpus analysis, a different team of researchers showed that the association between priming words and the elderly stereotype was roughly six times stronger for the English words used in the original study than for their translated French equivalents in the **REPLICATION** study. The lesson here is that potential violations of even seemingly minor auxiliary assumptions can make a big difference for the conclusions it is safe to draw from apparent acts of falsification (see Trafimow and Earp 2016).

Conclusion

According to Lakatos (1970), "given sufficient imagination … any theory can be permanently saved from 'refutation' by some suitable adjustment in the background knowledge in which it is embedded" (p. 184). Similarly, any finding can be indefinitely rescued from falsification if the original scientist is willing to claim that the replicators all made a mistake or violated an important auxiliary assumption (as indeed they may have done). But it is not enough to simply make such claims. Scientists must make explicit all relevant auxiliary assumptions, and clarify the range of conditions under which their purported finding should obtain. If a sufficient number of **REPLICATION** attempts obey those assumptions and are carried out under those conditions—as judged by an impartial expert—yet still fail to show the originally reported result, it then becomes reasonable for the community of scientists to abandon their confidence in the finding, thus treating it as effectively falsified.

References

Bargh, John, Chen, Mark, and Lara Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology* 71(2): 230.

Cesario, Joseph. 2014. Priming, replication, and the hardest science. *Perspectives on Psychological Science* 9(1): 40-48.

Doyen, Stéphane. Klein, Olivier, Pichon, Cora-Lise, and Axel Cleeremans. 2012 Behavioral priming: it's all in the mind, but whose mind? *PLOS ONE*, 7(1): e29081.

Doyen, Stéphane, Klein, Olivier, Simons, Daniel, and Axel Cleeremans. 2014. On the other side of the mirror: Priming in cognitive and social psychology. *Social Cognition* 32 (Supplement) : 12-32.

Earp, Brian and David Trafimow. 2015. Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology* 6(621): 1-11.

Lakatos, Imre. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the growth of knowledge.* London: Cambridge University Press.

LeBel, Etienne, Vanpaemel, Wolf, McCarthy, Randy, Earp, Brian and Elson, Malte. 2017. *A unified framework to quantify the trustworthiness of empirical research*. Psy ArXiv.

Popper, Karl. 1959. *The logic of scientific discovery*. London: Hutchinson & Co.

Trafimow, David and Earp, Brian. 2016. Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein. *Theory & Psychology* 26(4): 540-548.