aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

The Value of Direct Replication

Perspectives on Psychological Science
2014, Vol 9(1) 76–80

DOI: 10.1177/1745691613514755
pps.sagepub.com
SAGE

Daniel J. Simons
University of Illinois

Abstract
Reproducibility is the cornerstone of science. If an effect is reliable, any competent researcher should be able to obtain it when using the same procedures with adequate statistical power. Two of the articles in this special section question the value of direct replication by other laboratories. In this commentary, I discuss the problematic implications of some of their assumptions and argue that direct replication by multiple laboratories is the only way to verify the reliability of an effect.

Keywords
direct replication, conceptual replication, reliability, generalizability

Reproducibility is the cornerstone of science. The idea that direct replication undergirds science has a simple premise: If an effect is real and robust, any competent researcher should be able to obtain it when using the same procedures with adequate statistical power. Many effects in psychology replicate reliably across laboratories and even work as classroom demonstrations, but the reproducibility of some prominent effects in psychology has come under fire (see Pashler & Wagenmakers, 2012, for a summary). As a way to address this problem, *Perspectives on Psychological Science* recently launched a new type of article, the Registered Replication Report, devoted to verifying important psychology findings. These reports compile multiple replications of a single effect, conducted by labs throughout the world who all agree to follow a preregistered and vetted protocol. The end result is not a judgment of whether a single replication attempt succeeded or failed—it is a robust estimate of the size and reliability of the original finding.

Two of the articles in this special section (Cesario, 2014, this issue; Stroebe & Strack, 2014, this issue) question the value of direct replication by other laboratories.[1] Cesario recognizes the importance of direct replication in principle, but advocates replication by the originating lab as the best way test the reliability of an effect. Until theories can specify the contingencies that govern variation in effects, he argues that replication failures by other laboratories should be viewed as not just ambiguous, but uninformative. Stroebe and Strack reject direct replication in favor of conceptual replication, arguing that "the true purpose of replications is a (repeated) test of a theoretical hypothesis rather than an assessment of the reliability of a particular experimental procedure" (p. 61)

In this commentary, I challenge these claims and their underlying assumptions by making three related points:

1. Direct replication by other laboratories is the best (and possibly the only) believable evidence for the reliability of an effect;
2. The idea that only the originating lab can meaningfully replicate an effect limits the scope of our findings to the point of being uninteresting and unfalsifiable; and
3. Situational influences and moderators should be verified rather than assumed.

Trust but Verify

All findings represent a combination of some underlying effect (the signal) and sampling error (the noise). The noise can be further decomposed into systematic error (moderators and differences in samples) and unsystematic error (measurement error). Direct replication by multiple laboratories is the only way to isolate the signal from the noise and average across different types of error. The rejection or deferral of direct replication by other labs assumes that unspecified moderators or quirks of sampling are reliably biased against finding the original effect in a new setting.

Many effects in psychology are readily obtained across laboratories and samples (e.g., Crump, McDonnell, &

Corresponding Author:
Daniel J. Simons, Department of Psychology, University of Illinois, 603 E. Daniel Street, Champaign, IL 61820
E-mail: dsimons@illinois.edu

Gureckis, 2013; Germine et al., 2012; Roediger, 2012). If a mechanism is real, it should be possible to measure its impact across a wide range of contexts even if systematic errors weaken or strengthen it. Few studies provide compelling, reliable, replicated evidence for situational influences or individual difference moderators that completely swamp what are otherwise robust effects. Fragile or easily manipulated effects, those buffeted about by moderators, are exactly the ones most in need replication by other laboratories. Otherwise, there is no way to know whether the original measurements were just the result of a fortunate convergence of unknown factors.

The mantra, "trust but verify," is a hallmark of the scientific method. Cesario is right to advocate self-replication before publication—especially for preliminary or unexpected findings; we should trust our colleagues to publish only those findings that they can reproduce themselves (or those for which effect size estimates are reasonably precise due to the use of large samples). But self-replication is not adequate verification. Those same unidentified moderators that purportedly make effects fragile could also explain why the originating lab found them in the first place, leading to an internal replication that does not isolate and verify the reliability of the underlying effect. Only direct replication by other laboratories can do that, precisely because the noise should vary across laboratories.

Limitations of Scope

The idea that we should assume that undiscovered moderators are responsible for failed replications until proven otherwise is a claim about generalization. Every study in psychology is intended to generalize beyond the tested sample. If a paper reports inferential statistics (analysis of variance, *t* test), the authors have implicitly assumed generality to similar samples from the same target population (subjects are treated as a random effect). That assumption might later be proven wrong: perhaps the effect was specific to the sample tested that day, perhaps it only worked because testing took place in cubicles, perhaps it applies only to elite college students, or perhaps it relied on experimenter expectations. Those possibilities are what make the claim a scientific one; the assumption that an effect generalizes across situations can be disproven with evidence that it does not.

As Cesario correctly notes, researchers have been overzealous in generalizing from their results, and they should limit the scope of their claims (see also Giner-Sorolla, 2012; Henrich, Heine, & Norenzayan, 2010). But they should not assume complete specificity either. Imagine reviewing a manuscript that concludes "Our finding applies only to the students in our subject pool who were tested in our lab cubicles on a rainy afternoon in October by our RA." Presumably, a reviewer would reject it because it is inherently unfalsifiable (and uninteresting to anyone who does not care about a case study of those subjects). Other laboratories could not measure the same effect because they could not sample from that population. The originating lab could not verify that conclusion either; even direct replications by the originating laboratory assume some generality across situations and samples.

In contrast, if researchers overreach and incorrectly assume generalization to all of humanity, then at least their claim is falsifiable. That makes it a scientific claim, albeit an unjustified one (Henrich et al., 2010). Ideally, researchers should specify the expected generality of their finding: What population are they trying to sample from when they treat subjects as a random effect in their statistics, and what population are they generalizing to? With that information, other laboratories could sample from the same population. The effect might well prove more or less general than they think, but that is an empirical question requiring direct replication by other laboratories.

The Danger of Assumed Moderation

Although Cesario rightly notes the problem of positing moderators after the fact, his solution is worse. When researchers posit a moderator explanation for discrepant results, they make a testable claim: The effect is reliable, but differences in that moderator explain why one laboratory found the effect and the other did not. They then can conduct a confirmatory study to manipulate that moderator and demonstrate that they can reproduce the effect and make it vanish. In fact, they have a responsibility to do so in order to justify their claim that the effect is robust and varies as a result of moderation. Without such confirmatory evidence, a failed replication does provide some evidence against the reliability of the original effect.

Rather than treating moderation as an empirically tractable problem, Cesario and also Stroebe and Strack make it an *a priori* assumption: Any failed replication by another laboratory could result from moderation, so it can be dismissed as uninformative about the underlying mechanism. Consider the logical ramifications of that assumption. Any two studies will differ in some respects, even when they are conducted in the same laboratory—no study can be an exact replication. Under this default assumption of moderation, a direct replication provides no evidence for or against the reliability of an effect unless the replicating lab verifies that all potential moderators were equated. That is something not even the original laboratory could do. Researchers could treat a

failed direct replication just like a failed conceptual replication: It provides no challenge to the original finding (Pashler & Harris, 2012).

This *a priori* assumption places the burden of proof on the replicating laboratory to do something that is empirically impossible because the number of possible moderators is infinite: perhaps the effect depends on the phases of the moon, perhaps it only works at a particular longitude and latitude, perhaps it requires subjects who ate a lot of corn as children. Cesario holds that theories eventually will specify all of the relevant moderators, and direct replication by other laboratories will be useful after they do. But no theory will ever be that complete. Stroebe and Strack effectively dismiss direct replication because it is not exact replication: A replication of the same method with a different sample or in a different time is not dispositive because it might test different theoretical constructs. Critically, this claim undermines not just replication failures, but also replication successes. Any successful replication might just reflect the operation of unspecified moderators rather than verification of an underlying mechanism.

If we accept the idea that we should defer direct replication by other laboratories until theories are adequately complete, then psychology is not a scientific exploration of the mechanisms that affect our behavior. We cannot accumulate evidence for the reliability of any effect. Instead, all findings, both positive and negative, can be attributed to moderators unless proven otherwise. And we can never prove otherwise.

If we reject this default assumption of moderation, as we must, what role should we grant to moderators? Almost all published papers in psychology describe positive results (Fanelli, 2010; Smart, 1964), and few provide confirmatory tests showing that an effect can be completely eliminated or reversed as a function of some moderator (Cesario cites a handful). Those few that do also require verification via direct replication using large samples to show that the moderator effects themselves are reliable. If the main effects have yet to be replicated, should we trust moderator effects that have not been either?

In the absence of published, confirmatory, replicated evidence for the influence of a moderator on an effect, what should we make of claims that such moderators matter? Either (a) those making claims of moderators are remarkably adept at guessing what factors matter in the absence of experimental evidence, or (b) they have unpublished evidence for the effect of a moderator.

If researchers are just skilled guessers, then claims of moderation are only speculative. Without experimental evidence for moderation, what are the odds that the original researchers just happened to stumble upon a sample that had the right levels of self-monitoring, used research assistants who happened to have the right social skills for that effect, tested at the right time of day and the right month of the year, and used the appropriate laboratory arrangement? Given that these moderators presumably have different consequences for different effects, how is it that researchers successfully chose the right settings when they report only one test of an effect? And, what are the odds that researchers who conducted a failed direct replication necessarily flubbed the parameter settings, leading to a negative finding?

If, instead, researchers have unpublished, confirmatory evidence for the importance of a moderator that shows how the effect hinges on the correct settings, then they can report that evidence in their Method section so that others can control for it. Lacking such documentation, other researchers must assume that following the published method will reproduce the published effect (assuming adequate power, of course: Simmons, Nelson, & Simonsohn, 2013; Tversky & Kahneman, 1971)—that is the purpose of a Method section, after all. If researchers have empirical evidence for moderation and relegate it to the file drawer, then, ironically, they have buried the very evidence they need in order to document moderator effects, and they have engaged in a questionable research practice (John, Loewenstein, & Prelec, 2012).

Conclusion

Psychology has come under fire for producing unreliable effects (Pashler & Wagenmakers, 2012), and we must confront the questionable research practices that contribute to them (John et al., 2012; Simmons et al., 2011). Stroebe and Strack, however, deny the existence of a problem ("no solid data exist on the prevalence of such research practices," p. 60), ignoring both circumstantial evidence as well as the fact that researchers themselves report using such practices (John et al., 2012). Stroebe and Strack even dispute that such practices would be a problem ("the discipline still needs to reach an agreement about the conditions under which they are inacceptable," p. 60). Although they correctly note that a single failure to replicate should not be treated as definitive evidence against the existence of an effect, they dismiss the value of direct replication by adopting the premise that the purpose of replication is to provide new tests of a theory (i.e., conceptual replication) rather than to determine the reliability of an effect.

Unlike Stroebe and Strack, Cesario acknowledges the legitimacy of the problems and recommends a number of worthwhile changes to improve the reliability of published research: direct replication by the originating laboratory, increased sample sizes, an emphasis on effect

sizes, and more limited claims of generality (see Brandt et al., in press, and Asendorpf et al., 2013, for similar suggestions). He also recognizes the limits of conceptual replication as a way to measure reliability. Unfortunately, his proposal to defer direct replication by other laboratories until theories are suitably complete rejects a foundational principle of science: Direct replication by other scientists is the only way to verify the reliability of an effect.

Accumulated evidence for reliable effects is the lasting legacy of science—theories come and go, changing to account for new evidence and making new predictions. Numerous theories can account for any data and make predictions along the way. The value of direct replication comes from making sure that theories are accounting for the signal rather than the noise. A single failed replication does not prove that an original result was noise and it will not disprove a theory, but it does add information about the reliability of the original effect. A theory based on unreliable evidence will inevitably be a flawed description of reality, and if we discount direct replication by other researchers, then all findings will stand unchallenged and unchallengeable.

Only with direct replication by multiple laboratories will our theories make useful, testable, and generalizable predictions. We should not defer replication by other labs until we have suitably rich theories, and we cannot rely on tests of different effects (conceptual replication) as a way to assess reliability. We cannot have suitably precise theories without direct replication by other laboratories. Direct replication, ideally by multiple laboratories, is the only way to measure the reliability and generality of an effect across situations. Direct replication is the only way to make sure our theories are accounting for signal and not noise.

Acknowledgments

Thanks to Brent Roberts, Chris Fraley, Rolf Zwaan, Christopher Chabris, and Hal Pashler for their comments on earlier drafts of this commentary. Their feedback was essential, but I take responsibility for the views expressed in this commentary. Thanks also to Joe Cesario for an earlier email correspondence about these issues.

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

Note

1. The other article in this special section (Klatzky & Creswell, 2014, this issue) focuses less on the value of direct replication. Instead, it extrapolates from a model in sensory integration to develop a theoretical framework to account for the fragility of some priming results. Given that several recent articles, including those in a special issue of *Perspectives*, have addressed other aspects of the replicability crisis, I have focused this commentary on claims about direct replication.

References

Asendorpf, J. B., Conner, M., de Fruyt, F., de Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. doi:10.1002/per.1919

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . Van 't Veer, A. (in press). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*. Retrieved from http://dx.doi.org/10.1016/j.jesp.2013.10.005

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*, 40–48.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. doi:10.1371/journal.pone.0057410

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*(4), e10068. doi:10.1371/journal.pone.0010068

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*, 847–857.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*, 562–571. doi:10.1177/1745691612457576

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral & Brain Sciences*, *33*, 61–83. Retrieved from http://dx.doi.org/10.1017/S0140525X0999152X

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

Klatzky, R. L., & Creswell, D. (2014). An intersensory interaction account of priming effects—and their absence. *Perspectives on Psychological Science*, *9*, 49–58.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536. doi:10.1177/1745691612463401

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *Observer*, *25*(2), *9*, 27–29.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013, January). *Life after P-Hacking (2013)*. Paper presented at the Society for Personality and Social Psychology annual meeting, New Orleans, LA. Retrieved from http://ssrn.com/abstract=2205186

Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychologist*, *5a*, 225–232. doi:10.1037/h0083036

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.