ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(5) 552–555
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614543974
pps.sagepub.com
SAGE

An Introduction to Registered Replication Reports at *Perspectives on Psychological Science*

Daniel J. Simons[1], Alex O. Holcombe[2], and Barbara A. Spellman[3]
[1]University of Illinois, [2]University of Sydney, and [3]University of Virginia

Much of science, including psychology, consists of eliciting, measuring, and documenting effects. In psychology, these effects are used to test and support theories about how the mind works or why people behave the way they do. Yet, for several reasons, the first report of an effect in the published literature rarely provides enough evidence to draw firm conclusions about the actual size of the effect. First, the sample sizes typically used in psychology studies are not large enough to measure an effect with precision (Marszalek, Barber, Kohlhart, & Holmes, 2012). Second, scientific publishing favors statistically significant, novel results over inconclusive or negative ones (Fanelli, 2012). If discrepant results rarely enter the literature, then published studies will, on average, overestimate the true size of the effects they report. And third, just by chance, some published studies will be false positives (Ioannidis, 2005).

Uncertainty about the true size of important effects hampers psychological theorizing. Moreover, single studies provide little evidence for the robustness of an effect across the population to which it is assumed to generalize. This issue of *Perspectives on Psychological Science* includes the first example of a new type of journal article, one designed to provide a more definitive measure of the size and reliability of important effects: the Registered Replication Report (RRR; see Simons & Holcombe, 2014). RRRs compile a set of studies from a variety of laboratories that all followed an identical, vetted protocol designed to reproduce the original method and finding as closely as possible. By combining the resources of multiple labs, RRRs provide the ingredients for a meta-analysis that can authoritatively establish the size and reliability of an effect.

The RRR Process

Given the nature and scope of these multilab replications, the processes for planning, running, analyzing, and publishing an RRR differ from those of a more typical empirical journal article. The first step is a proposal. A researcher or a team of researchers submits a presubmission inquiry to the editors (details here: http://www.psychologicalscience.org/index.php/replication/instructions-for-proposers). In it, they make a case for the "replication value" of the original finding. Has the effect has been highly influential? Is it methodologically sound? Is the size of the effect uncertain due to controversy in the published literature or a lack of published direct replications? These original findings often have generated tremendous interest among scientists and in the media. They might also have practical implications. At this point, the RRR editor in charge (either Simons or Holcombe) consults experts in the research area to confirm that the original study does merit the large-scale effort involved in such a multilab replication project. Only those effects that meet such criteria proceed beyond the initial proposal stage.

The first RRR

The preliminary submission and review process described above is how RRRs normally will be initiated. However, for the first RRR, the editors selected the study to be replicated—a prominent finding in the literature on eyewitness accuracy.

In 1990, Schooler and Engstler-Schooler reported a phenomenon that they called "verbal overshadowing." Participants viewed a video of a simulated bank robbery and later attempted to pick the robber out of a line-up of photographs. After watching the video, but before the line-up task, participants were asked either to write a description of the robber (experimental condition) or to generate a list of U.S. states and their capitals (control

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

condition). Surprisingly, those who wrote a description of the robber were less likely to later pick him from the lineup than were those in the control condition. This finding is theoretically important because it suggests that recalling and describing a person's appearance can impair rather than improve later recognition memory performance. The finding is practically important as well, with ramifications for how police obtain descriptions of perpetrators from actual witnesses.

The verbal overshadowing effect in the original paper was large: approximately a 25% difference in accuracy between those who wrote a description of the robber and those who did the control task. However, the original study provided only an initial estimate of this effect, and the sample sizes were relatively small. A later meta-analysis (Meissner & Brigham, 2001), averaging across studies that used a range of materials and procedures, found a smaller effect than the original report.

Given the theoretical and practical importance of the verbal overshadowing effect and the uncertainty about the true size of the effect, we at *Perspectives* believed that it would be a good choice for the first RRR. Having an editor steer the first RRR was beneficial as we encountered the inevitable frustrations in designing and implementing this new process. One of us (Simons) led the effort and it proceeded in consultation with many others, especially the lead author of the original study, Jonathan Schooler.

Developing the study protocol

Regardless of how the process is initiated (by editors or researchers), the second step is to establish a protocol for the study. The proposing researchers complete a form detailing the methodological and analysis details of the original study, suggesting how those details will be implemented in the replication, and identifying any discrepancies or missing information.

The editor then asks the original author(s) to help verify that the proposed replication procedures are true to the original study and effect. The editor also asks the original author(s) to identify manipulation checks or boundary conditions necessary to measure the effect accurately. Unlike a traditional review that occurs after data have been collected, this vetting process is intended to be a constructive, mediated dialog between the proposing team and the original authors, with a common goal of developing an accurate and complete protocol. (Note that if an original author does not want to be involved, the editor, preferably with the help of the original author, will identify another qualified researcher to vet the protocol.)

In some cases, the replication protocol might improve on shortcomings of the original study's method. For example, advances in technology might permit more precise timing of stimulus presentation. Or the replication procedure might implement tighter controls for confounds inherent to the original design or add measures of moderators that were identified in subsequent research. Working together, the editor, original author, and proposing team may introduce minor changes that minimize the potential for variability in how the experiment is conducted by the multiple labs involved in an RRR. For example, for the verbal overshadowing RRR, we replaced the original videotape used to show the robbery with a digitized movie file and allowed for computer-based presentations rather than having the video shown on a television.

Running the study

The third step begins after the full protocol is finalized: The journal posts a call for other laboratories interested in contributing to the collective replication effort. The Association for Psychological Science (APS) disseminates this call via email and social media. Ongoing RRRs are posted at http://www.psychologicalscience.org/index.php/replication/ongoing-projects. Labs interested in participating must submit a form explaining how they will meet the requirements of the protocol and documenting their relevant experience. The editor reviews these proposals in an effort to make sure that the researchers are qualified to conduct the study and that the assortment of laboratories involved in the replication effort fairly represent all sides of any theoretical disagreements about the effect or its implications. To accommodate involvement by researchers in different countries, sometimes with different languages or different equipment, small variations in the implementation of the protocol are preapproved by the editor in consultation with the original author(s) as needed.

In the fourth step, individual labs that have been approved to participate document their implementation plan for the study on OpenScienceFramework.org. The editor then verifies that their plan meets all of the specified requirements, and the lab then creates a registered version of their plan. The individual labs must conduct the study by following that preregistered plan; their results are included in the RRR regardless of the outcome.

Analyzing the Data

Having all labs use the same methods and procedures, and including results from all registered labs regardless of the outcome, allows for the creation of a meta-analysis that has no file drawer problem. The data analysis concentrates on estimation of effect sizes rather than on significance testing. Neither dichotomous judgments of statistical significance nor reporting of *p* values is sufficient for understanding the size and reliability of effects. And, with the large total sample inherent to a multilab replication, even tiny effects could be statistically significant. As a result, effect sizes (and their accompanying confidence intervals) are the primary measures reported both for the individual labs and for the meta-analytic effect across all of the included studies.

The full data set from each laboratory is available for analysis, both for the RRR and for subsequent analysis by other researchers. Thus, it is possible to meta-analyze the effect directly from the data rather than from reported statistical analyses, which typically vary across published papers. Consequently, the meta-analysis in an RRR will not necessarily use the same effect size measure as was used in the original study. For example, the verbal overshadowing RRR in this issue used the percentage difference between the experimental and control condition as the basis of the meta-analysis rather than computing an effect size from a chi square or other statistical test.

Even though all of the studies in an RRR adopt the same procedures, they might not be measuring exactly the same effect. Some effects may be different across labs because of systematic differences in the subject population. For example, in the verbal overshadowing RRR, the true effect might differ for participants tested in different countries due to cultural differences (although we have no reason to expect this). For most RRRs, a random effects model will be used for the meta-analysis to allow for the possibility that individual studies are measuring different underlying effects. If there is more variability in the effect sizes than would be expected by chance, then additional analyses could explore the effects of potential moderators. If during the development of the protocol, the original authors identify a possible moderator of the effect, the RRR will include tests of that moderator. With the full availability of the data, any researcher can conduct exploratory analyses to test other moderators that were not identified in advance.

Discussing and Using the Results

RRRs can provide an unbiased and precise estimate of the size of an effect. The precision comes from combining the results of multiple independent tests using a common protocol, even though individual study results contributing to that meta-analytic estimate are necessarily imprecise (see Klein et al., 2014 for another example of this "many labs" approach). Consequently, participating laboratories are discouraged from dwelling on whether or not their individual study rejected the null hypothesis. The absence of statistical bias occurs by virtue of preregistration and the guarantee of publication, eliminating *p* hacking by individual research teams and publication bias by the journal.

The effect size estimate provided by an RRR is unlikely to match that of the original study exactly. The discussion section of an RRR considers differences in the measurement error and sample sizes across the laboratories and discusses the practical significance of an effect of the size measured by the meta-analysis. It may also discuss moderators. The full discussion sticks closely to the evidence with little speculation or new theorizing. Theory and speculative discussion are important, but as we explain below, we expect them to occur in other venues.

Following completion of the RRR manuscript, author(s) of the original study who have helped advance the RRR process are invited to submit a brief commentary. That commentary is peer reviewed, and if it makes a substantive contribution, it is published alongside the RRR.

Broader Impact

After publication of the RRR, we expect that the broad community of researchers will explore the full implications of the RRR for theory and for future experimental work. As the raw data from RRRs are available publicly (barring ethical constraints), researchers can do their own analyses and potentially discover new effects or identify possible moderators. The data can be easily incorporated into other meta-analyses. Commentaries and reanalyses that make new substantive contributions to the literature may be considered for future issues of *Perspectives* or may be published in other outlets. We hope that the RRR process of open and careful replication, estimation, and evaluation will lead to a better understanding of important effects in our field, and more generally advance the reproducibility and replicability of psychological science.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

References

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2012). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills*, *112*, 331–348.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, *15*, 603–616.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71.

Simons, D. J., & Holcombe, A. O. (2014, March). Registered replication reports: A new article type at *Perspectives on Psychological Science*. *Observer*, *27*(3). Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2014/march-14/registered-replication-reports.html