11

The Reproducibility Project: A Model of Large-Scale Collaboration for Empirical Research on Reproducibility

Open Science Collaboration*

CONTENTS

* See Endnote at the end of the chapter for a listing of authors.

The goal of science is to accumulate knowledge that answers questions such as "How do things work?" and "Why do they work that way?" Scientists use a variety of methodologies to describe, predict, and explain natural phenomena. These methods are so diverse that it is difficult to define a unique scientific method, although all scientific methodologies share the assumption of reproducibility (Hempel and Oppenheim, 1948; Kuhn, 1962; Popper, 1934/1992; Salmon, 1989).

In the abstract, reproducibility refers to the fact that scientific findings are not singular events or historical facts. In concrete terms, reproducibility—and the related terms repeatability and replicability—refers to whether research findings recur. "Research findings" can be understood narrowly or broadly. Most narrowly, reproducibility is the repetition of a simulation or data analysis of existing data by reexecuting a program (Belding, 2000). More broadly, reproducibility refers to *direct replication*, an attempt to replicate the original observation using the same methods of a previous investigation but collecting unique observations. Direct replication provides information about the reliability of the original results across samples, settings, measures, occasions, or instrumentation. Most broadly, reproducibility refers to *conceptual replication*, an attempt to validate the *interpretation* of the original observation by manipulating or measuring the same conceptual variables using different techniques. Conceptual replication provides evidence about the validity of a hypothesized theoretical relationship. As such, direct replication provides evidence that a finding can be obtained, and conceptual replication provides evidence about what it means (Schmidt, 2009).

These features of reproducibility are nested. The likelihood of direct replication is constrained by whether the original analysis or simulation can be repeated. Likewise, the likelihood that a finding is valid is constrained by whether it is reliable (Campbell et al., 1963). All of these components of reproducibility are vitally important for accumulating knowledge in science, with each directly answering its own specific questions about the predictive value of the observation. The focus of the present chapter is on direct replication.

An important contribution of direct replication is to identify false-positives. False-positives are observed effects that were inferred to have occurred because of features of the research design but actually occurred by chance. Scientific knowledge is often gained by drawing inferences about a population based on data collected from a sample of individuals to make inferences about the population as a whole. Since this represents an example of induction, the knowledge gained in this way is always uncertain. The best a researcher can do is estimate the likelihood that the research findings are not a product of ordinary random sampling variability and provide a probabilistic measure of the confidence they have in the result. Independently reproducing the results reduces the probability that the original finding occurred by chance alone and, therefore, increases confidence in the inference. In contrast, false-positive findings are unlikely to be replicated.

Given the benefits of direct replication to knowledge building, one might expect that evidence of such reproducibility would be published frequently. Surprisingly, this is not the case. Publishing replications of research procedures is rare (Amir and Sharon, 1990; Makel et al., 2012; Morrell and Lucas, 2012; Open Science Collaboration, 2012). One recent review of psychological science estimated that only 0.15% of published studies were attempts to directly replicate a previous finding (Makel et al., 2012). As a consequence, there is a proliferation of scientific findings, but little systematic effort to verify their validity, possibly leading to a proliferation of irreproducible results (Begley and Ellis, 2012; Prinz et al., 2011). Despite the low occurrence of published replication studies, there is evidence that scientists believe in the value of replication and support its inclusion as part of the public record. For example, a survey of almost 1300 psychologists found support for reserving at least 20% of journal space to direct replications (Fuchs et al., 2012).

In this chapter, we first briefly review why replications are highly valued but rarely published. Then we describe a collaborative effort—the *Reproducibility Project*—to estimate the rate and predictors of reproducibility in psychological science. We emphasize that, while a goal of direct replication is to identify false-positive results, it does not do so unambiguously. Direct replication always includes differences in sample, setting, or materials that could be theoretically consequential boundary conditions for obtaining the original result. Finally, we detail how we are conducting this project as a large-scale, distributed, open collaboration. A description of the procedures and challenges may assist and inspire other teams to conduct similar projects in other areas of science.

11.1 Current Incentive Structures Discourage Replication

The ultimate purpose of science is the accumulation of knowledge. The most exciting science takes place on the periphery of knowledge, where researchers suggest novel ideas, consider new possibilities, and delve into the unknown. As a consequence, innovation is a highly prized scientific contribution, and the generation of new theories, new methods, and new evidence is highly rewarded. Direct replication, in contrast, does not attempt to break new ground; it instead assesses whether previous innovations are accurate. As a result, there are currently few incentives for conducting and publishing direct replications of previously published research (Nosek et al., 2012).

Current journal publication practices discourage replications (Collins, 1985; Mahoney, 1985; Schmidt, 2009). Journal editors hope to maximize the impact of their journals and are inclined to encourage contributions that are associated with the greatest prestige. As a consequence,

all journals encourage innovative research, and few actively solicit replications, whether successful or unsuccessful (Neuliep and Crandall, 1990). An obvious response to these publication practices is to create journals devoted to publishing replications or null results. Of multiple attempts to start such a journal over the last 30 years, success is fleeting. Several versions exist today (e.g., http://www.jasnh.com/; http://www.jnr-eeb.org/; http://www.journalofnullresults.com/), but challenges remain: journals that publish what no other journal will publish ensures their low status (Nosek et al., 2012). It is not in a scientist's interest to publish in low-status journals.

Because prestigious journals do not provide incentives to publish replications, researchers do not have a strong incentive to conduct them (Hartshorne and Schachner, 2012a; Koole and Lakens, 2012). Scientists make reasonable assessments of how they should spend their time. Publication is the central means of career advancement for scientists. Given the choice between replication and pursuing novelty, career researchers can easily conclude that their time should be spent pursuing novel research. This may be especially true for researchers that do not yet have academic tenure.

Complicating matters is the presence of additional forces rewarding positive over negative results. A common belief is that it is easier to obtain a negative result erroneously than it is to obtain a positive result erroneously. This is true when using statistical techniques and sample sizes designed to detect differences (Nickerson, 2000) and when designs are underpowered (Cohen, 1962; Lipsey and Wilson, 1993; Sedlmeier and Gigerenzer, 1989). Although both of these features are common, researchers can design studies so that they will be informative no matter the outcome (Greenwald, 1975). There are many reasons why a null result may be observed erroneously such as imprecise measurement, poor experimental design, or other forms of random error (Greenwald, 1975; Nickerson, 2000). There are also many reasons why a positive result may be observed erroneously such as introducing artifacts into the research design (Rosenthal and Rosnow, 1960), experimenter bias, demand characteristics, systematic apparatus malfunction, or other forms of systematic error (Greenwald, 1975). Further, false-positives can be inflated through selective reporting and adventurous data analytic strategies (Simmons et al., 2011). There is presently little basis other than power of research designs to systematically prefer positive results compared to negative results. Decisions about whether to take a positive or negative result seriously are based on evaluation of the research design, not the research outcome.

Layered on top of legitimate epistemological considerations are cultural forces that favor significant (Fanelli, 2010, 2012; Greenwald, 1975; Sterling, 1959) and consistent (Giner-Sorolla, 2012) results over inconsistent or ambiguous results. These incentives encourage researchers to obtain and publish positive, significant results and to suppress or ignore inconsistencies that disrupt the aesthetic appeal of the findings. As examples, researchers

might decide to stop data collection if preliminary analyses suggest that the findings will be unlikely to reach conventional significance, examine multiple variables or conditions and report only the subset that "worked," accept those studies that confirm the hypothesis as effective designs, and dismiss those that do not confirm the hypothesis as pilots or methodologically flawed because they fail to support the hypothesis (LeBel and Peters, 2011). These practices, and others, can inflate the likelihood that the results are false-positives (Giner-Sorolla, 2012; Ioannidis, 2005; John et al., 2012; Nosek et al., 2012; Schimmack, 2012; Simmons et al., 2011). This is not to say that researchers engage in these practices with deliberate intent to deceive or manufacture false effects. Rather, these are natural consequences of motivated reasoning (Kunda, 1990). When a particular outcome is better for the self, then decision making can be influenced by factors that maximize the likelihood of that outcome. Researchers may tend to carry out novel scientific studies with a confirmatory bias such that they—without conscious intent—guide themselves to find support for their hypotheses (Bauer, 1992; Nickerson, 1998).

11.2 Publishing Incentives Combined with a Lack of Replication Incentives May Reduce Reproducibility

The strong incentives to publish novel, positive, and clean results may lead to problems for knowledge accumulation. For one, the presence of these incentives leads to a larger proportion of false-positives, which produces a misleading literature and makes it more likely that future research will be based on claims that are actually false. Any individual result is ambiguous; but because the truth value of a claim is based on the aggregate of individual observations, ignoring particular results undermines the accuracy of a field's collective knowledge. This occurs both by inflating the true size of the effect and by concealing potential limitations to the effect's generalizability. Knowing the rate of false-positives in the published literature would clarify the magnitude of the problem and indicate whether significant intervention is needed. However, there is very little empirical evidence on the rate of false-positives. Simulations, surveys, and reasoned arguments provide some evidence that the false-positive rate could be very high (Greenwald, 1975; Hartshorne and Schachner, 2012a; Ioannidis, 2005). For example, asking psychologists about the proportion of research findings that would be reproduced from their journals in a direct replication yielded an estimate of 53% (Fuchs et al., 2012). The two known empirical estimates of nonrandom samples of studies in biomedicine provide disturbing reproducibility estimates of 25% or less (Begley and Ellis, 2012; Prinz et al., 2011). There are few other existing attempts to estimate the rate of false-positives in any field of science.

The theme of this chapter is reproducibility, and the focus of this section is on the primary concern of irreproducibility: that the original results are false. Note, however, that the reproducibility rate is not necessarily equivalent to the false-positive rate. The *maximum* reproducibility rate is 1 minus the rate of false-positives tolerated by a field. The ubiquitous alpha level of 0.05 implies a false-positive tolerance of 5%, meaning a reproducibility rate of 95%. However, in practice, there are many reasons why a true effect may fail to replicate. A low-powered replication, one with an insufficient number of data points to observe a difference between conditions, can fail for mathematical rather than empirical reasons.

The reproducibility rate can be lowered further for other reasons. Imprecise reporting practices can inadvertently omit crucial details necessary to make research designs reproducible. Description of the methodology—a core feature of scientific practice—may become more illustrative than substantive. This could be exacerbated by editorial trends encouraging short-report formats (Ledgerwood and Sherman, 2012). Even when the chance to offer additional online material about methods occurs, it may not be taken. For example, a Google Scholar search on articles published in *Psychological Science*—a short-report format journal—for the year 2011 revealed that only 16.8% of articles included the phrase "supplemental material" denoting additional material available online, even without considering whether or not that material gave a full accounting of methods. As a consequence, when replication does occur, the replicating researchers may find reproduction of the original procedure difficult because key elements of the methodology were not published. This makes it difficult both to clarify the conditions under which an effect can be observed and to accumulate knowledge.

In sum, both false-positives and weak methodological specification are challenges for reproducibility. The current system of incentives in science does not reward researchers for conducting or reporting replications. As a consequence, there is little opportunity to estimate the reproducibility rate, to filter out those initial effects that were false-positives, and to improve specification of those initial effects that are true but specified inadequately. The *Reproducibility Project* examines these issues by generating an empirical estimate of reproducibility and identifying the predictors of reproducibility.

11.3 Reproducibility Project

The Reproducibility Project began in November 2011 with the goal of empirically estimating the reproducibility of psychological science. The concept was simple: Take a sample of findings from the published literature in psychology and see how many of them could be replicated. The implementation, however, is more difficult than the conception. Replicating a large

number of findings to produce an estimate of reproducibility is a mammoth undertaking, requiring much time and diverse skills. Given the incentive structures for publishing, only a person who does not mind stifling their own career success would take on such an effort on their own even if they valued the goal. Our solution was to minimize the costs for any one researcher by making it a massively collaborative project.

The Reproducibility Project is an open collaboration to which anyone can contribute according to their skills and available resources. Project tasks are distributed among the research team, minimizing the demand on each individual contributor but still allowing for a large-scale research design. As of this writing (March 2013), 118 researchers have joined the project, a complete research protocol has been established, and more than 50 replication studies are underway or completed. The project, though incomplete, has already provided important lessons about conducting such large-scale, distributed projects. The remainder of this chapter describes the design of the project, what can be learned from the results, and the lessons for conducting a large-scale collaboration that could be translated to similar efforts in other disciplines.

11.3.1 Project Design

To estimate the rate and predictors of reproducibility in the psychological sciences, we selected a quasi-random sample of studies from three prominent psychological journals (*Journal of Personality and Social Psychology*; *Journal of Experimental Psychology: Learning, Memory, and Cognition*; and *Psychological Science*) from the 2008 publication year—a year far enough in the past that there is evidence for variation in impact of the studies and variability in independent replication attempts and not so far in the past that original materials would not be available. Studies were selected for replication as follows: Beginning with the first issue of 2008, the first 30 articles that appeared in each journal made up the initial sample. As project members started attempting to replicate studies, additional articles were added to the eligible pool in groups of 10. This strategy minimized selection biases by having only a small group of articles available for selection at any one time while maintaining a sufficient number of articles so that interested replication teams could find tasks that match their resources and expertise.

Each article in the sampling frame was reviewed with a standard coding procedure*. The coding procedure documented (1) the essential descriptors of the article such as authors, topic, and main idea; (2) the key finding from one of the studies and key statistics associated with that finding such

* Linked resources are also available via the Reproducibility Project's page on the Open Science Framework website: http://openscienceframework.org/project/VMRGu/wiki/home.

as sample size and effect size; (3) features of the design requiring specialized samples, procedures, or instrumentation; and (4) any other unusual or notable features of the study. This coding provided the basis for researchers to rapidly review and identify a study that they could potentially replicate. Also, coding all articles from the sampling frame will allow systematic comparison of the articles replicated with those that were available but not replicated.

Most articles contain more than one study. Since the Reproducibility Project is concerned with the state of replicability in general, a single key finding was sampled from a single study. By default, the last study reported in a given article was the target of replication. If a replication of that study was not feasible, then the second to the last study was considered. If no studies were feasible to replicate, then the article was excluded from the replication sample. A study was considered feasible for replication if its primary result could be evaluated with a single inference test and if a replication team on the project had sufficient access to the study's population of interest, materials, procedure, and expertise. Although every effort is made to make the sample representative, study designs that are difficult to reproduce for practical reasons are less likely to be included. In psychology, for example, studies with children and clinical samples tend to be more resource intensive than others. Likewise, it is infeasible to replicate some study designs with large samples, many measurements over time, a focus on one-time historical events, or expensive instrumentation. It is not obvious whether studies with significant resource challenges would have more or less reproducible findings as compared to those that have fewer resource challenges.

11.3.2 Maximizing Replication Quality

A central concern for the Reproducibility Project was the quality of replication attempts. Sloppy, nonidentical, or underpowered replications would be unlikely to replicate the original finding, even if that original finding was true. While these are potential predictors of reproducibility, they are not particularly interesting ones. As a consequence, the study protocol involved many features to maximize quality of the replications. As a first step, each replication attempt was conducted with a sufficient number of observations so that replications of true findings would be likely. For each eligible study, a power analysis was performed on the effect of interest from the original study. The power analysis determined the samples necessary for 80%, 90%, and 95% power to detect a statistically significant effect the same size as the prior result using the same analytic procedures. Replication teams planned their sample size aiming for the highest feasible power. All studies were designed to achieve at least 80% power, and about three-fourths of the studies conducted to date have an anticipated power of 90% or higher.

In another step to maximize replication quality, replication teams contacted the original authors of each study to request copies of project materials and clarify any important procedures that did not appear in the original report (http://bit.ly/rpemailauthors). As of this writing, authors of every original article have shared their materials to assist in the replication efforts, with one exception. In the exceptional case, the original authors declined to share all materials that they had created and declined to disclose the source of materials that they did not own so that the replication team could seek permission for their use. Even so, a replication attempt of that study is underway with the replication team using its own judgment on how to best implement the study.

Next, for all studies, the replication team developed a research methodology that reproduced the original design as faithfully as possible. Methodologies were written following a standard template and included measurement instruments, a detailed project procedure, and a data analysis plan. Prior to finalizing the procedure, one or two Reproducibility Project contributors who were not a part of the replication team reviewed this proposed methodology. The methodology was also sent to the original authors for their review. If the original authors raised concerns about the design quality, the replication teams attempted to address them. If the design concerns could not be addressed, those concerns were documented as a priori concerns raised by the original authors. The evaluations of the original authors were documented as endorsing the methods of the replication, raising concerns based on informed judgment or speculation (which are not part of the published record as constraints on the design), raising concerns that are based on published empirical evidence of the constraints on the effect, or no response. This review process minimized design deficiencies in advance of conducting the study and also obtained explicit ratings of the design quality in advance. These steps should make it easier to detect post hoc rationalization if the replication results violate researchers' expectations.

Some studies that were originally conducted in a laboratory were amenable to replication via the Internet. Using the web is an excellent method for recruiting additional power for human research, but it could also alter the likelihood of observing the original effects. Thus, we label such studies "secondary replications." These studies remained eligible to be claimed for "primary replications"—doing the study in the laboratory following the original demonstration. As of this writing, there were more than 10 secondary web replications underway in addition to the more than 50 primary replications. This provides an opportunity to evaluate systematically whether the change in setting affects reproducibility.

Upon finalization, the replication methodology was registered and added to an online repository. At this point, data collection could start. After data collection, the replication teams conducted confirmatory analyses following the registered methodology. The results and interpretation were documented and submitted to a team member (who was not part of the

replication team) for review. In most cases, an additional attempt was made to contact the authors of the original study in order to share the results of the replication attempt and to consult with them as to whether any part of the data collection or data analysis process may have deviated from that of the original study. Finally, the results of the replication attempt were written into a final manuscript, which was logged in the central project repository. As additional replication attempts are completed, the repository is updated and a more complete picture of the reproducibility of the sample emerges (http://openscienceframework.org/project/EZcUj/).

The project is ongoing. In principle, there need not be an end date. Just as ordinary science accumulates evidence about the truth value of claims continuously, the Reproducibility Project could accumulate evidence about the reproducibility, and ultimately truth value, of its particular sample of claims continuously. Also, new resources provide opportunities to improve and enlarge the sample of replication studies. For example, in February 2013, the project received a grant of more than $200,000 to support replication projects. The project team formed a committee and grant application process to encourage more researchers to join the project and strengthen the study. Eventually, the collaborative team will establish a closing date for replication projects to be included in an initial aggregate report. That aggregate report will provide an estimate of the reproducibility rate of psychological science and examine predictors of reproducibility such as the publishing journal, the precision of the original estimate, and the existence of other replications in the published literature.

11.4 What Can and Cannot Be Learned from the Reproducibility Project

The Reproducibility Project will produce an estimate of the reproducibility rate of psychological science. In fact, it will produce multiple estimates, as there are multiple ways to conceive of evaluating replication (Open Science Collaboration, 2012). For example, a standard frequentist solution is to test whether the effect reaches statistical significance with the same ordinal pattern of means as the original study. An alternative approach is to evaluate whether the meta-analytic combination of the original observation and replication produces a significant effect. A third possibility is to test whether the replication effect is significantly different from the original effect size estimate. Each of these will reveal distinct reproducibility rates, and each offers a distinct interpretation. Notably, none of the possible interpretations will answer the question that is ultimately of interest: At what rate are the conclusions of published research true?

11.4.1 Of the Studies Investigated, Which of Their Conclusions Are True?

The relationship between the validity of a study's results and the validity of the conclusions derived from those results is, at best, indirect. Replication only addresses the validity of the results. If the original authors used flawed inferential statistics, then replicating the result may say nothing of the accuracy of the conclusion (e.g., Jaeger, 2008). Similarly, if the study used a confounded manipulation, and that confound explains the reported results rather than the original interpretation, then the interpretation is incorrect regardless of whether the result is reproducible. More generally, replication cannot help with misinterpretation Piaget's (1952, 1954) demonstrations of object permanence and other developmental phenomena are among the most replicable findings in psychology. Simultaneously, many of his interpretations of these results appear to have been incorrect (e.g., Baillargeon et al., 1985).

Reinterpretation of old results is the ordinary process of scientific progress. That progress is facilitated by having valid results to reinterpret. Piaget's conclusions may have been overthrown, but his empirical results still provide the foundation for much of developmental psychology. The experimental paradigms he designed were so fruitful, in part, because the results they generate are so easily replicated. In this sense, reproducibility is essential for theoretical generativity. The Reproducibility Project offers the same contribution as other replications toward increasing confidence in the truth of conclusions. Findings that replicate in the Reproducibility Project are ones that are more likely to replicate in the future. The aggregate results will provide greater confidence in the validity of the findings, whether or not the conclusions are correct.

11.4.2 Of All Published Studies, What Is the Rate of True Findings?

It is of great importance to know the rate of valid findings in a given field. Even under the best of circumstances, at least some findings will be false due to random chance or simple human error. While there is a concern that science may be far from the ideal (e.g., Ioannidis et al., 2001), there are little systematic data in any field and hardly any in psychology. There are at least two barriers to obtaining empirical data on the rate of true findings. The first is that accumulating such data across a large sample of findings requires a range of expertise and a supply of labor that is difficult to assemble. In that respect, one of the contributions of the Reproducibility Project is to show how this can be accomplished. The second is that, as discussed earlier, failure to replicate a result is not synonymous with the result being a false-positive.

The Reproducibility Project attempts to minimize the other factors that are knowable and undesirable (e.g., low power and poor replication design) and to estimate the influence of others. There are three possible interpretations of a failure to replicate the results of an original study:

Interpretation 1: The original effect was false. The original result could have occurred by chance (e.g., setting alpha = 0.05 anticipates a 5% false-positive rate), by fraud, or unintentionally by exploiting flexible research practices in design, analysis, or reporting (Greenwald, 1975; John et al., 2012; Simmons et al., 2011).

Interpretation 2: The replication was not sufficiently powered to detect the true effect (i.e., the replication is false). Just as positive results occur by chance when there is no result to detect (alpha = 0.05), negative results occur by chance when there is a result to detect (beta or power). Most studies are very underpowered (Lipsey and Wilson, 1993; Sedlmeier and Gigerenzer, 1989; see Cohen, 1962, 1992). Adequate power is a necessary feature of fair replication attempts. The Reproducibility Project sets 80% as the baseline standard power for replication attempts (Cohen, 1988) and encourages higher levels of power whenever possible. The actual power of our replications can be used as a predictor of reproducibility in the analytic models and as a way to estimate the false-negative rate among replications. For example, an average power of 85% across replications would lead us to expect a false-negative rate of 15% on chance alone.

Interpretation 3: The replication methodology differed from the original methodology on unconsidered features that were critical for obtaining the true effect. There is no such thing as an exact replication. A replication necessarily differs somehow, or else it would not be a replication. For example, in behavioral research, even if the same participants are used, their state and experience differ. Likewise, even if the same location, procedures, and apparatus are used, the history and social context have changed. There are infinite dimensions of sample, setting, procedure, materials, and instrumentation that could be conditions for obtaining an effect. Keeping with the principle of Occam's razor, these variables are assumed irrelevant until proven otherwise. Indeed, if an effect is interpreted as existing only for the original circumstances, with no explanatory value outside of that lone occasion, its usefulness for future research and application is severely limited. Consequently, authors almost never exhaustively report procedural details when writing about effects.

Part of standard research practice is to understand the conditions necessary to elicit an effect. Does it depend on the color of the walls? The hardness of the pencils used? The characteristics of the sample? The context of measurement? How the materials are administered? There is an infinite number of possible conditions, and a smaller number of plausible conditions, that could be necessary for obtaining an effect.

A replication attempt will necessarily differ in many ways from the original demonstration. The key question is whether a failure to replicate could

plausibly be attributed to any of these differences. The answer may rest upon what aspect of the original effect each difference violates:

1. *Published constraints on the effect*: Does the original interpretation of the effect suggest necessary conditions that are not part of the replication attempt? If the original interpretation is that the effect will only occur for women, and the replication attempt includes men, then it is not a fair replication. The existing interpretation (and perhaps empirical evidence) already imposes that constraint. Replication is not expected. Replication teams avoid violating these constraints as much as possible in the Reproducibility Project. Offering original authors an opportunity to review the design provides another opportunity to identify and address these constraints. When the constraints cannot be addressed completely, they are documented as potential predictors of reproducibility.
2. *Constraints on the effect, identified a priori*: An infinitely precise description requires infinite journal space, and thus every method section is necessarily an abridged summary. Thus, there may be design choices that are known (to the original experimenters, if to no one else) to be crucial to obtaining the reported results, but not described in print. By contacting the original authors prior to conducting the replication attempt, the Reproducibility Project minimizes this flaw in the published record.
3. *Constraints on the effect, identified post hoc*: Constraints identified beforehand are distinct from the reasoning or speculation that occurs after a failed replication attempt. There are many differences between any replication and its original, and subsequent investigation may determine that one of these differences, in fact, was crucial to obtaining the original results. That is, the original effect is not reproducible as originally interpreted but is reproducible with the newly discovered constraints. The Reproducibility Project only initiates this process: For studies that do not replicate, interested researchers may search for potential reasons why. This might include additional studies that manipulate the factors identified as possible causes of the replication failure. Such research will produce a better understanding of the phenomenon.
4. *Errors in implementation or analysis for the original study, replication study, or both*: Errors happen. What researchers think and report that they did might not be what they actually did. Discrepancies in results can occur because of mistakes. There is no obvious difference between "original" or "replication" studies in the likelihood of errors occurring. The Reproducibility Project cannot control errors in original studies, but it can make every effort to minimize their occurrence in the replication studies. For example, it is conceivable

> that the Reproducibility Project will fail to replicate studies because some team members are incompetent in the design and execution of the replication projects. While this possibility cannot be ruled out entirely, procedures including carefully detailed experimental protocols minimize its impact and maximize the likelihood of identifying whether competence is playing a role. Moreover, features of the replication team (e.g., relevant experience, degrees, publishing record) can be used as predictors of reproducibility.

The key lesson from this section is that failure to replicate does not unambiguously suggest that the original effect is false. The Reproducibility Project examines all of the possibilities described earlier in its evaluation of reproducibility. Some can be addressed effectively with design. For example, all studies will have at least 80% power to detect the original effect, and the power of the test will be evaluated as a predictor for likelihood of replication. Also, differences between original and replication methods will be minimized by obtaining original materials whenever possible and by collaborating with original authors to identify and resolve all possible published or a priori identifiable design constraints. Finally, original authors and other members of the collaborative team review and evaluate the methodology and analysis to minimize the likelihood of errors in the replications, and the designs, materials, and data are made available publicly in order to improve the likelihood of identifying errors. Notwithstanding the ambiguity surrounding the interpretation of a replication failure, the key value of replication remains: as data accumulate, the precision of the effect estimate increases.

11.4.3 What Practices Lead to More Replicable Findings?

Perhaps the most promising possible contribution of the Reproducibility Project will be to provide empirical evidence of the correlates of reproducibility or to make a more informed assessment of the reproducibility of existing results. Researchers have no shortage of hypotheses as to what research practices would lead to higher replicability rates (e.g., LeBel and Paunonen, 2011; LeBel and Peters, 2011; Nosek and Bar-Anan, 2012; Nosek et al., 2012; Vul et al., 2009). Without systematic data, there is no way to test these hypotheses (for discussion, see Hartshorne and Schachner, 2012a,b). Note that this is a correlational study, so it is possible that some third factor, such as the authors' conscientiousness, is the joint cause of both the adoption of a particular research practice and high replicability. However, the lack of a correlation between certain practices and higher replicability rates is—assuming sufficient statistical power and variability—more directly interpretable, suggesting that researchers should look elsewhere for methods that will meaningfully increase the validity of published findings.

11.4.4 Summary

Like any research effort, the most important factor for success of the Reproducibility Project is the quality and execution of its design. The quality of the design, execution of replications, and ultimate interpretations of the findings will define the extent to which the Reproducibility Project can provide information about the reproducibility of psychological science. As with all research, that responsibility rests with the team conducting the research. The last section of this chapter summarizes the strategies we are pursuing to conduct an open, large-scale, collaborative project with the highest-quality standards that we can achieve (Open Science Collaboration, 2012).

11.5 Coordinating the Reproducibility Project

The success of the Reproducibility Project hinges on effective collaboration among a large number of contributors. In business and science, large-scale efforts are often necessary to provide important contributions. Sending an astronaut to the moon, creating a feature film, and sequencing the human genome are testaments to the power of collaboration and social coordination. However, most large-scale projects are highly resourced with money, staff, and administration in order to assure success. Further, most large-scale efforts are backed by leadership that has direct control over the contributors through employment or other strong incentives, giving contributors compelling reasons to do their part for the project.

The Reproducibility Project differs from the modal large-scale project because it started light on resources and light on leadership. Most contributors are donating their time and drawing on whatever resources they have available to conduct replications. Project leaders cannot require action because the contributors are volunteers. How can such a project succeed? Why would any individual contributor choose to participate?

The Reproducibility Project team draws its project-design principles from open-source software communities that developed important software such as the Linux operating system and the Firefox web browser. These communities achieved remarkable success under similar conditions. In this section, we describe the strategies used for coordinating the Reproducibility Project so that other groups can draw on the project design to pursue similar scientific projects. An insightful treatment of these project principles and strategies is provided in Michael Nielsen's (2011) book *Reinventing Discovery*.

The challenges to solve are the following: (1) recruiting contributor, (2) defining tasks so that contributors know what they need to do and can do it, (3) ensuring high-quality contributions, (4) coordinating effectively so that contributions can be aggregated, and (5) getting contributors to follow

through on their commitments. The next sections describe the variety of strategies the project uses to address these challenges.

11.5.1 Clear Articulation of the Project Goals and Approach

Defining project goals is so obvious that it is easy to overlook. Prospective contributors must know what the project will accomplish (and how) to decide whether they want to contribute. The Reproducibility Project's primary goal is to estimate the reproducibility of psychological science. It aims to accomplish that goal by conducting replications of a sample of published studies from major journals in psychology. The extent to which prospective contributors find the goal and approach compelling will influence the likelihood that they volunteer their time and resources. Further, once the team is assembled, a clear statement of purpose and approach bonds the team and facilitates coordination. This goal and approach is included in every communication about the Reproducibility Project.

11.5.2 Modularity

Even though potential contributors may find the project goal compelling, they recognize that they could never conduct so many replications by themselves. The Reproducibility Project's goal of replicating dozens of studies is appealing because it has the potential to impact the field, but actually replicating those many studies is daunting. One solution is crowdsourcing (Estellés-Arolas and González-Ladrón-de-Guevara, 2012), in which work is decomposed into smaller, modular tasks that are distributed across volunteers.

Modularity is the extent to which a project can be separated into independent components and then recombined later. Also, if contributors are highly dependent on each other, then the time delay is multiplicative: delay by one affects all. The Reproducibility Project is highly modularized. Individuals or small teams conduct replications independently. Some replications are completed very rapidly, others over a longer time scale. Barriers to progress are isolated to the competing schedules and responsibilities of the small replication teams.

Besides accelerating progress, modularizing is attractive to volunteer contributors because they have complete control over the extent and nature of their participation. Modularization is useful, but it will provide limited value if there are only a few contributors. One way for crowdsourcing to overcome this problem is to have a low barrier to entry.

11.5.3 Low Barrier to Entry

Breaking up a large project into pieces reduces the amount of contribution required by any single contributor. For volunteers with busy lives,

this is vital. The Reproducibility Project encourages small contributions so that contributors can volunteer their services incrementally without incurring inordinate costs to their other professional responsibilities or allowing unfulfilled commitments to impede workflow.

Even with effective modularization, prospective contributors may have difficulty in estimating the workload required when making the initial commitment to contribute. Uncertainty itself is a formidable barrier to entry. The Reproducibility Project provides specific documentation to reduce this barrier. In particular, prospective contributors can review studies available for replication in a summary spreadsheet, consult with a team member whose role is to connect available studies to new contributors with appropriate skills and resources, and review the replication protocol that provides instruction for every stage of the process. Effective supporting material and personnel simplify the process of joining the project.

11.5.4 Leverage Available Skills

Collaborations can be particularly effective when they incorporate researchers with distinct skill sets. A problem that is very difficult for a nonexpert may be trivial for an expert. Further, there are many potential contributors that do not have resources or skills to do the central task: conducting a replication. In any large-scale project, there are additional administrative, documentation, or consulting tasks that can be defined and modularized. The Reproducibility Project has administrative contributors with specified roles and contributors who assist by documenting and coding the studies available for replication. There are also consultants for common issues such as data analysis.

11.5.5 Collaborative Tools and Documentation

As a distributed project, the Reproducibility Project coordination must embrace asynchronous schedules. Communication among the entire team occurs via an e-mail LISTSERV (https://groups.google.com/group/opensciencefframework?hl=en) that maintains a record of all communications. New ideas, procedural issues, project plans, and task assignments are discussed on the LISTSERV. Decisions resulting from team discussion are codified in project documentation that is managed with Google Docs and the Open Science Framework (OSF; http://opensciencefframework.org/).

Print documentation is extensive, as it is the primary means of providing individual contributors with knowledge of (1) what is happening in the project, (2) their role in the project, and (3) what they must do to fulfill their role. The project documentation defines the overall objective of the project, tables of subgoals and actions necessary to achieve them, protocols for conducting a replication project, and templates for communicating results. This workflow is designed to maximize the quality of the replication,

make explicit the standards and expectations of each replication, and minimize the workload for the individual contributors. With a full specification of the workflow, templates for report writing, and material support for correspondence with original study authors, the replicating teams can smoothly implement the project's standard procedures and focus their energies on the unique elements of the replication study design and data collection to conduct the highest-quality replication possible.

Unlike modular replications, administrative tasks require frequent and timely upkeep and can impact the workflow of other team members. Thus, although initially run by volunteers, dedicated administrative support was needed as the project increased in scale. Together, documentation and dedicated administrators provide continuity in the projects' objectives and methods across time and individual replication teams.

The highly defined workflow also makes it easy to track progress of one's own replication—and those of others. Each stage of the project has explicitly defined milestones, described in the project's researcher guide, and team members denote on the project tracksheet when each stage is completed. At a glance, viewers of the tracksheet can see the status of all projects. Besides its information value, tracking progress provides normative information for the research teams regarding whether they are keeping up with the progress of other teams. Without that information, individual contributors would have little basis for social comparison and also little sense of whether the project as a whole is making progress.

11.5.6 Light Leadership with Strong Communication

Large-scale, distributed projects flounder without leadership. However, leadership cannot be overly directive when volunteers staff the project. Project leaders are responsible for facilitating communication and discussion and then guiding the team to decisions and action. Without someone taking responsibility for the latter, projects will stall with endless discussion and no resolution.

To maximize project investment, individual contributors should have the experience that their opinions about the project design matter and can impact the direction of the project. Simultaneously, there must be sufficient leadership to avoid having each contributor feel like they shoulder inordinate responsibility for decision making. Contributors vary in the extent to which they desire to shape different aspects of the project. Some have strong opinions about the standard format of the replication report; others would rather step on a nail than spend time on that. To balance this, the Reproducibility Project leadership promotes open discussion without requiring contribution. Simultaneously, leadership defines a timeline for decision making, takes responsibility for reviewing and integrating opinions, and makes recommendations for action steps.

11.5.7 Open Practices

The Reproducibility Project is an open project. This means that anyone can join, that expectations of contributors are defined explicitly in advance, and that the project discussion, design, materials, and data are available publicly. Openness promotes accountability among the team. Individuals have made public commitments to project activities. This transparency minimizes free-riding and other common conflicts that emerge in collaborative research. Openness also promotes accountability to the public. Replication teams are trying to reproduce research designs and results published by others. The value of the evidence accumulated by the Reproducibility Project relies on these replications being completed to a high standard. Making all project materials available provides a strong incentive for the replication teams to do an excellent job. Further, openness increases the likelihood that errors will be identified and addressed. In addition to public accessibility, the Reproducibility Project builds in error checking by requiring each replication team to contact original authors to invite critique of their study design prior to data collection and by having members review and critique each others' project reports.

11.5.8 Participation Incentives

Why participate in a large-scale project? What is in it for the individual contributor? The best designed and coordinated project will still fail if contributors have no reason to participate voluntarily. The Reproducibility Project has a variety of incentives that may each have differential impact on individual contributors. For one, many contributors have an intrinsic interest in the research questions the project has set out to answer or, more generally, view the project as an important service to the field.

Another class of incentives is experiential. Some contributors want to belong to a large-scale collaboration, try open science practices, or conduct a direct replication. For some, this may be for the pleasure of working with a group or trying something new. For others, this may be conceived as a training opportunity. Other incentives are the more traditional academic rewards. The most obvious is publication. Publication is the basis of reward, advancement, and reputation building (Collins, 1985). Contributors to the Reproducibility Project earn coauthorship on publication about the project and its findings. The relative impact for each individual contributor is most certainly reduced by the fact that there are many contributors. However, the nature of the research question, the scale of the project, and (in our humble brag opinion) quality of the endeavor mean that the project may have a high impact on psychology and science more generally. While no contributor will establish a research career using publications with the Open Science Collaboration exclusively, authorship on an important, high-profile project provides

an added bonus for the more intrinsic factors that motivate contributions to the Reproducibility Project.

11.6 Conclusion

The Reproducibility Project is the first attempt to systematically and empirically estimate the reproducibility of a subdiscipline of science. It draws on the lessons of open-source projects in software development: leveraging individuals' opinions about how things should be done while providing strong coordination to enable progress. What will be learned from the Reproducibility Project is still undetermined. But if the current progress is any indicator, the high investment of its contributors and the substantial interest and attention by observers suggest that the Reproducibility Project could provide a useful initial estimate of the reproducibility of psychological science and perhaps inspire other disciplines to pursue similar efforts.

Systematic data on replicability do not exist. The Reproducibility Project addresses this shortcoming. If large numbers of findings fail to replicate, that will strengthen the hand of the reform movements and lead to a significant reevaluation of the literature. If most findings replicate satisfactorily—as many as would be expected given our statistical power estimates—then that will suggest a different course of action. More likely, perhaps, is that the results will be somewhere in between and will help generate hypotheses about particular practices that could improve or damage reproducibility.

We close by noting that even in the best of circumstances, the results of any study—including the Reproducibility Project—should be approached with a certain amount of skepticism. While we attempt to conduct replication attempts that are as similar as possible to the original study, it is always possible that "small" differences in method may turn out to be crucial. Thus, while a failure to replicate should decrease confidence in a finding, one does not want to make too much out of a single failure (Francis, 2012). Rather, the results of the Reproducibility Project should be understood as an opportunity to learn whether current practices require attention or revision. Can we do science better? If so, how? Ultimately, we hope that we will contribute to answering these questions.

Endnote

1. Alexander A. Aarts, Nuenen, the Netherlands; Anita Alexander, University of Virginia; Peter Attridge, Georgia Gwinnett College;

Štìpán Bahník, Institute of Physiology, Academy of Sciences of the Czech Republic; Michael Barnett-Cowan, Western University; Elizabeth Bartmess, University of California, San Francisco; Frank A. Bosco, Marshall University; Benjamin Brown, Georgia Gwinnett College; Kristina Brown, Georgia Gwinnett College; Jesse J. Chandler, PRIME Research; Russ Clay, University of Richmond; Hayley Cleary, Virginia Commonwealth University; Michael Cohn, University of California, San Francisco; Giulio Costantini, University of Milan–Bicocca; Jan Crusius, University of Cologne; Jamie DeCoster, University of Virginia; Michelle DeGaetano, Georgia Gwinnett College; Ryan Donohue, Elmhurst College; Elizabeth Dunn, University of British Columbia; Casey Eggleston, University of Virginia; Vivien Estel, University of Erfurt; Frank J. Farach, University of Washington; Susann Fiedler, Max Planck Institute for Research on Collective Goods; James G. Field, Marshall University; Stanka Fitneva, Queens University; Joshua D. Foster, University of South Alabama; Rebecca S. Frazier, University of Virginia; Elisa Maria Galliani, University of Padova; Roger Giner-Sorolla, University of Kent; R. Justin Goss, University of Texas at San Antonio; Jesse Graham, University of Southern California; James A. Grange, Keele University; Joshua Hartshorne, M.I.T.; Timothy B. Hayes, University of Southern California; Grace Hicks, Georgia Gwinnett College; Denise Humphries, Georgia Gwinnett College; Georg Jahn, University of Greifswald; Kate Johnson, University of Southern California; Jennifer A. Joy-Gaba, Virginia Commonwealth University; Lars Goellner, University of Erfurt; Heather Barry Kappes, London School of Economics and Political Science; Calvin K. Lai, University of Virginia; Daniel Lakens, Eindhoven University of Technology; Kristin A. Lane, Bard College; Etienne P. LeBel, University of Western Ontario; Minha Lee, University of Virginia; Kristi Lemm, Western Washington University; Melissa Lewis, Reed College; Stephanie C. Lin, Stanford University; Sean Mackinnon, Dalhousie University; Heather Mainard, Georgia Gwinnett College; Nathaniel Mann, California State University, Northridge; Michael May, University of Bonn; Matt Motyl, University of Virginia; Katherine Moore, Elmhurst College; Stephanie M. Müller, University of Erfurt; Brian A. Nosek, University of Virginia; Catherine Olsson, M.I.T.; Marco Perugini, University of Milan–Bicocca; Michael Pitts, Reed College; Kate Ratliff, University of Florida; Frank Renkewitz, University of Erfurt; Abraham M. Rutchick, California State University, Northridge; Gillian Sandstrom, University of British Columbia; Dylan Selterman, University of Maryland; William Simpson, University of Virginia; Colin Tucker Smith, University of Florida; Jeffrey R. Spies, University of Virginia; Thomas Talhelm, University of Virginia; Anna van 't Veer, Tilburg University; Michelangelo Vianello, University of Padova.

References

Amir, Y. and Sharon, I. (1990). Replication research: A "must" for the scientific advancement of psychology. *Journal of Social Behavior and Personality*, 5, 51–69.

Baillargeon, R., Spelke, E. S., and Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, 20, 191–208.

Bauer, H. H. (1992). *Scientific Literacy and the Myth of the Scientific Method*. Chicago, IL: University of Illinois Press.

Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483, 531–533. doi:10.1038/483531a.

Belding, T. C. (2000). Numerical replication of computer simulations: Some pitfalls and how to avoid them. arXiv preprint nlin/0001057.

Campbell, D. T., Stanley, J. C., and Gage, N. L. (1963). *Experimental and Quasi-Experimental Designs for Research* (pp. 171–246). Boston, MA: Houghton Mifflin.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.

Collins, H. M. (1985). *Changing Order*. London, U.K. Sage.

Estellés-Arolas, E. and González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2), 189–200.

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

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90, 891–904. doi:10.1007/s1192-011-0494-7.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review*, 19, 975–991.

Fuchs, H., Jenny, M., and Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science*, 7, 634–637. doi:10.1177/1745691612459521.

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

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1–20.

Hartshorne, J. K. and Schachner, A. (2012a). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*, 6(8), 1–13. doi:10.3389/fncom.2012.0008.

Hartshorne, J. K. and Schachner, A. (2012b). Where's the data? *The Psychologist*, 25, 355.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15, 135–175.

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

Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics*, 29, 306–309.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434–446. doi: 10.1016/j.bbr.2011.03.031.

John, L., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23, 524–532. doi: 10.1177/0956797611430953.

Koole, S. L. and Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, 7, 608–614. doi:10.1177/1745691612462586.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480–498. doi:10.1037/0033-2909.108.3.480.

LeBel, E. P. and Paunonen, S. V. (2011). Sexy but often unreliable: Impact of unreliability on the replicability of experimental findings involving implicit measures. *Personality and Social Psychology Bulletin*, 37, 570–583.

LeBel, E. P. and Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, 15, 371–379. doi:10.1037/a0025172.

Ledgerwood, A. and Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, 7, 60–66. doi:10.1177/1745691611427304.

Lipsey, M. W. and Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181–1209. doi:10.1037/0003-066X.48.12.1181.

Mahoney, M. J. (1985). Open exchange and epistemic process. *American Psychologist*, 40, 29–39.

Makel, M. C., Plucker, J. A., and Hagerty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, 7, 537–542. doi:10.1177/1745691612460688.

Morrell, K. and Lucas, J. W. (2012). The replication problem and its implications for policy studies. *Critical Policy Studies*, 6, 182–200. doi:10.1080/19460171.2012.689738.

Neuliep, J. W. and Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, 5, 85–90.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 175–220.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301. doi:10.1037/1082-989X.5.2.241.

Nielson, M. (2011). Reinventing discovery: The new era of networked science. Princeton University Press.

Nosek, B. A. and Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, 23(3), 217–243. doi:1080/1047840X.2012.692215.

Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7, 615–631. doi:10.1177/1745691612459058.

Open Science Collaboration. (2012). An open, large-scale collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7, 657–660. doi:10.1177/1745691612462588.

Piaget, J. (1952). *The Origins of Intelligence in Children*. New York: International University Press.

Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic.

Popper, K. (1934/1992). *The Logic of Scientific Discovery*. New York: Routledge.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10, 712–713. doi:10.1038/nrd3439-c1.

Rosenthal, R. and Rosnow, R. L. (1960). *Artifact in Behavioral Research*. New York: Academic Press.

Salmon, W. (1989). *Four Decades of Scientific Explanation*. Minneapolis, MN: University of Minnesota Press.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17, 551–566.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13, 90–100. doi:10.1037/a0015108.

Sedlmeier, P. and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. doi:10.1177/0956797611417632.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, 54, 30–34.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives in Psychological Science*, 4, 274–90. doi:10.1111/j.1745-6924.2009.01125.x.