Maximizing the Reproducibility of Your Research

Open Science Collaboration[1]

Open Science Collaboration (in press). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*. New York, NY: Wiley.

Authors' Note: Preparation of this chapter was supported by the Center for Open Science and by a Veni Grant (016.145.049) awarded to Hans IJzerman. Correspondence can be addressed to Brian Nosek, nosek@virginia.edu.

[1] Alexander A. Aarts, Nuenen, The Netherlands; Frank A. Bosco, Virginia Commonwealth University; Katherine S. Button, University of Bristol; Joshua Carp, Center for Open Science; Susann Fiedler, Max Planck Institut for Research on Collective Goods; James G. Field, Virginia Commonwealth University; Roger Giner-Sorolla, University of Kent; Hans IJzerman, Tilburg University; Melissa Lewis, Center for Open Science; Marcus Munafò, University of Bristol; Brian A. Nosek, University of Virginia; Jason M. Prenoveau, Loyola University Maryland; Jeffrey R. Spies, Center for Open Science

Commentators in this book and elsewhere describe evidence that modal scientific practices in design, analysis, and reporting are interfering with the credibility and veracity of the published literature (Begley & Ellis, 2012; Ioannidis, 2005; Miguel et al., 2014; Simmons, Nelson, & Simonsohn, 2011). The reproducibility of published findings is unknown (Open Science Collaboration, 2012a), but concern that is lower than desirable is widespread - even among scientists themselves (Fuchs, Jenny, & Fiedler, 2012; Pashler & Wagenmakers, 2012). Further, common practices that interfere with reproducibility are maintained by incentive structures that prioritize innovation over accuracy (Nosek, Spies, & Motyl, 2012). Getting deeper into the metascience literature reviewing scientific practices might lead to a discouraging conclusion for the individual scientist - I cannot change the system on my own, so what should I do?

This chapter provides concrete suggestions for increasing the reproducibility of one's own research. We address reproducibility across the research lifecycle: project planning, project implementation, data analysis, reporting, and programmatic research strategies. We also attend to practical considerations for surviving and thriving in the present scientific culture, while simultaneously promoting a cultural shift toward transparency and reproducibility through the collective effort of independent scientists and teams. As such, the practical suggestions to increase research credibility can be incorporated easily into the daily workflow without requiring substantial additional work in the short-term and perhaps saving substantial time in the long-term. Further, emerging requirements by journals, granting agencies and professional organizations are adding recognition and incentives for reproducible science. Doing reproducible science will increasingly be seen as the way to advance one's career, and this chapter may provide a means to get a head start.

Project Planning

Use High Powered Designs

Within the nearly universal null hypothesis significance testing (NHST) framework there

are two inferential errors that can be made: (I) falsely rejecting the null hypothesis (i.e., believing that an effect exists even though it doesn't) and (II) falsely failing to reject it when it is false (i.e., believing that no effect exists even though it does). "Power" is the probability of rejecting the null hypothesis when it is false given that an effect actually exists. Power depends on the size of the investigated effect, the alpha level, and the sample size.[2] Low statistical power undermines the purpose of scientific research; it reduces the chance of detecting a true effect but also, perhaps less intuitively, reduces the likelihood that a statistically significant result reflects a true effect (Ioannidis, 2005). The problem of low statistical power has been known for over 50 years: Cohen (1962) estimated that in psychological research the average power of studies to detect small and medium effects was 18% and 48% respectively, a situation that had not improved almost 25 years later (Sedlmeier & Gigerenzer, 1989). More recently, Button and colleagues (Button et al., 2013) showed that the median statistical power of studies in the neurosciences is between 8% and 31%.

Considering that many of the problems of low power are well-known and pernicious, it should be surprising that low power research is still the norm. Some reasons for the persistence of low powered studies include: (1) resources are limited, (2) researchers know that low power is a problem but do not appreciate its magnitude, and (3) there are insidious, perhaps unrecognized, incentives for engaging in low powered research when publication of positive results is the primary objective. That is, it is easier to obtain false positive results with small samples, particularly by using one's limited resources on many small studies rather than one large study (Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Ioannidis, 2005; Nosek et al., 2012). Given the importance of publication for academic success, these are formidable barriers.

What can you do? To start, consider the conceptual argument countering the

[2] Even outside of the dominant NHST model, the basic concept of higher power still holds in a straightforward way - increase the precision of effect estimates with larger samples and more sensitive and reliable methods.

publication incentive. If the goal is to produce accurate science, then adequate power is *essential*. When studying true effects, higher power increases the likelihood of detecting them. Further, the lure of publication is tempting, but the long-term benefits are greater if the published findings are credible. Which would you rather have: more publications with uncertain accuracy or fewer publications with more certain accuracy? Doing high-powered research will take longer, but the rewards may last longer.

Recruiting a larger sample is an obvious benefit, when feasible. There are also design strategies to increase power without more participants. For some studies, it is feasible to apply within-subject and repeated-measurement designs. These approaches are more powerful than between-subject and single-measurement designs. Repeated measures designs allow participants to be their own controls reducing data variance. Also, experimental manipulations are powerful as they minimize confounding influences. Further, reliable outcome measures reduce measurement error. For example, all else being equal, a study investigating hiring practices will have greater power if participants make decisions about many candidates compared to an elaborate scenario with a single dichotomous decision about one candidate. Finally, standardizing procedures and maximizing the fidelity of manipulation and measurement during data collection will increase power.

A complementary approach for doing high-powered research is collaboration. When a single research group cannot achieve the sample size required to provide sufficient statistical power, multiple groups can administer the same study materials and then combine data. For example, the “Many Labs” replication project administered the same study across 36 samples, totalling more than 6000 participants, producing both extremely high-powered tests of the effects and sufficient data to test for variability across sample and setting (Klein et al., 2014). Likewise, large-scale collaborative consortia in fields such as human genetic epidemiology have transformed the reliability of findings in these fields (Austin, Hair, & Fullerton, 2012). Even just combining efforts across 3 or 4 labs can increase power dramatically, while minimizing the labor

and resource impact on any one contributor. Moreover, concerns about project leadership opportunities for publishing can be minimized with quid pro quo agreements - you run my study, I'll run yours.

Create an Analysis Plan

Researchers have many decisions to make when conducting a study and analyzing the data. Which data points should be excluded? Which conditions and outcome variables are critical to assess? Should covariates be included? What variables might moderate the key relationship? For example, Carp (2012a) found that among 241 studies using functional magnetic resonance imaging (fMRI) there were 223 unique combinations of data cleaning and analysis procedures (e.g., correction for head motion, spatial smoothing, temporal filtering). The inordinate flexibility in analysis options provides researchers with substantial degrees-of-freedom to keep analyzing the data until a desired result is obtained; Carp (2012b) reports that when using the over 30,000 possible combinations of analysis methods on a single neuroimaging experiment, 90.3% of brain voxels differed significantly between conditions in at least one analysis. This flexibility could massively inflate false positives (Simmons et al., 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

The best defense against inflation of false positives is to reduce the degrees of freedom available to the researcher by writing down, prior to analyzing the data, how the data will be analyzed. This is the essence of confirmatory data analysis (Wagenmakers et al., 2012). The key effect of committing to an analysis plan in advance is to preserve the meaning of the *p*-values resulting from the analysis. The *p*-value is supposed to indicate the likelihood that these data would have occurred if there was no effect to detect. This interpretation is contingent on how many tests on the data were run and reported. Once the data have been observed, the universe of possible tests may be reduced to those that appear to be differences, and tests that do not reveal significant effects may be ignored. Without an *a priori* analysis plan, the extent to which the likelihood of a false positive has occurred is entirely unknown.

Writing down an analysis plan in advance stimulates a more thorough consideration of potential moderators and controls as well as a deeper involvement with the previous research and the formulated theories. By committing to a prespecified analysis plan one can avoid common cognitive biases (Kunda, 1990; Nosek et al., 2012). This approach also allows researchers to be open about and rewarded for their exploratory research (Wagenmakers et al, 2012), and highlights the value of conducting pilot research in order to clarify the qualities and commitments for a confirmatory design.

Project Implementation

Determine Data Collection Start and Stop Rules

It is not uncommon for researchers to peek at their data and when just shy of the "magical" *alpha* = .05 threshold for significance to add participants to achieve significance (John, Loewenstein, Prelec, 2012). This is problematic because it inflates the false-positive rate (Simmons et al., 2011). Likewise, particularly with difficult to collect samples (e.g., infant studies, clinical samples), there can be ambiguity during pilot testing about when a study is ready to begin. A few particularly "good" participants might be promoted to the actual data collection if the status of piloting versus actual data collection is not clear. Defining explicit data collection start and stop rules is effective self-protection against false positive inflation. These could be defined as a target number of participants per condition, a target period of time for data collection, as a function of an *a priori* power analysis, or by any other strategy that removes flexibility for deciding when data collection begins and ends (Meehl, 1990). Some journals, such as *Psychological Science*, now require disclosure of these rules.

Register Study and Materials

Many studies are conducted and never reported. This "file-drawer effect" is a major challenge for the credibility of published results (Rosenthal, 1979). Considering only the likelihood of reporting a null result versus a positive result and ignoring the flexibility in analysis strategies, Greenwald (1975) estimated the false-positive rate to be greater than 30%. However,

it is difficult to imagine that every study conducted will earn a full write-up and published report. A more modest solution is to register every study at the onset of data collection in a public registry. Registration involves, at minimum, documentation of the study design, planned sample, and research objectives. Registration ensures that all conducted studies are discoverable, and facilitates the investigation of factors that may differentiate the universe of studies conducted from the universe of studies published.

Public registration of studies is required by law in the United States for clinical trials (De Angelis et al., 2004) and is a pre-condition for publication in many major medical journals. The 2013 Declaration of Helsinki, a possible bellwether of ethical trends, recommends that this requirement be extended to all studies involving human participants (http://www.wma.net/en/30publications/10policies/b3/). This movement towards more transparency of all research can improve accessibility of findings that did not reach the published literature in order to evaluate potential biases in publishing and aggregate all evidence for a phenomenon.

A common concern about public registration is that one's ideas may be stolen by others before the research is completed and published. Registration actually certifies the originator of ideas with a time and date stamp. But, for the cautious researcher, some modern registries allow researchers to register studies privately and then reveal the registration later (e.g., https://osf.io/ described later).

Data Analysis

Perform Confirmatory Analyses First

For confirmatory analyses to retain their interpretability, they must be conducted and reported in full. Consider, for example, pre-registering 20 unique tests and reporting the single test that achieved a *p*-value below .05. Selective reporting renders a confirmatory analysis plan irrelevant. Likewise, a confirmatory analysis plan does not eliminate interpretability challenges of multiple comparisons. So, disclosing all 20 registered tests does not make the one significant

result less vulnerable to being a false positive. The key for registering an analysis plan is that it constrains the initial analyses conducted and makes clear that any potential Type 1 error inflation is limited to those confirmatory analyses.

In the ideal confirmatory analysis, the analysis script is created in advance and executed upon completion of data collection. In some cases, this ideal will be difficult to achieve. For example, there may be honest mistakes in the pre-registration phase or unforeseen properties of the data – such as non-normal data distributions or lack of variation in a key variable – that make deviations from the original analysis plans necessary. Having an analysis plan - with whatever degree of specificity is possible - makes it easy to clarify deviations from a strictly confirmatory analysis; explanations of those deviations makes it easier to judge their defensibility.

Conduct Exploratory Analysis for Discovery, not for Hypothesis Testing

Exploratory analysis is a valuable part of data analysis (Tukey, 1977). Much of progress in science is through accidental discovery of questions and hypotheses that one did not think to have in advance (Jaeger & Halliday, 1998). The emphasis on confirmatory designs does not discourage exploratory practice. Rather, it makes explicit the difference between outcomes resulting from confirmatory and exploratory approaches.

In exploratory analysis, inductive reasoning is used to form tentative *a posteriori* hypotheses that explain the observations (Stebbins, 2001). Popper (1959) proposed that a hypothesis that is derived from a given set of observations cannot be falsified by those same observations. As Popper noted “a hypothesis can only be empirically *tested*—and only *after* it has been advanced” (p.7). Making explicit the distinction between confirmatory and exploratory analysis helps clarify the confidence in the observed effects, and emphasizes the fact that effects from exploratory analysis require additional investigation.

Test Discoveries with Confirmatory Designs

With discovery in hand, the temptation for publication is understandable. Replication

offers only the dreaded possibility of “losing” the effect (Nosek et al., 2012). There may be no palliative care available other than to point out that while many exploratory results are opportunities to develop hypotheses to be tested, they are not the hypothesis tests themselves. The long-term view is that it is better to learn quickly that the effect is irreproducible than to expend yours and others’ resources on extensions that falsely assume its veracity. Following a discovery with a high-powered confirmatory test is the single best way to enhance the credibility and reproducibility of research findings.

This point is not universal. There are instances for which the effects in exploratory analysis are estimated with such precision that it is highly unlikely that they are chance findings. However, the circumstances required for this are uncommon in most psychological applications. The most common cases are studies with many thousands of participants. Even in these cases, it is possible to leverage chance and exaggerate results. Further, data collection circumstances may not be amenable to conducting a confirmatory test after an exploratory discovery. For example, with extremely hard-to-access samples, the effort required to conduct a confirmatory test may exceed available resources.

It is simply a fact that clarifying the credibility of findings can occur more quickly for some research applications compared with others. For example, research on the development of infant cognition often requires laborious laboratory data collections with hard to reach samples. Adult personality investigations, on the other hand, can often be administered via the Internet to hundreds of people simultaneously. The former will necessarily accumulate information and knowledge more slowly than the latter.

These constraints do not exempt research areas from the tentativeness of exploratory results and the need for confirmatory investigations. Rather, because of practical constraints, some research applications may need to tolerate publication of more tentative results and slower progress in verification.

Keep Records of Analyses

Some challenges to reproducibility are more a function of deficient record-keeping than analysis and reporting decisions. Analysis programs, like SPSS, provide easy-to-use point-and-click interfaces for conducting analyses. The unfortunate result is that it can be very easy to forget the particulars of an analysis if only the output persists. A simple solution for increasing reproducibility is to retain scripts for exactly the analyses that were conducted and reported. Coupled with the data, re-executing the scripts would reproduce the entire analysis output. This is straightforward with script-based analysis programs like R, STATA and SAS, but is also easy with SPSS by simply generating and saving the scripts for the conducted analyses. Taking this simple step also offers practical benefits to researchers beyond improved reproducibility. When analysis procedures are carried out without the use of scripts, adding new data points, revising analyses, and answering methodological questions from reviewers can be both time-consuming and error-prone. Using scripts makes these tasks incomparably simpler and more accurate.

Share and Review Data and Analysis Scripts Among Collaborators

Science is often done in teams. In most cases, team members have some specialization such as one member developing the research materials and another conducting analysis. In most cases, all members of a collaboration should have access to all of the research materials and data from a study. At minimum, shared access ensures that any member of the team could find the materials or data if other team members were not available. Moreover, sharing materials and data increases the likelihood of identifying and correcting errors in design and analysis prior to publication. For example, in software development, code review - the systematic evaluation of source code - is common practice to fix errors and improve the quality of the code for its purpose and reusability (Kemerer & Paulk, 2009; Kolawa & Huizinga, 2007). Such practices are easy to incorporate into scientific applications, particularly of analysis scripts, in order to increase confidence and accuracy in the reported analyses.

Finally, sharing data and analysis scripts with collaborators increases the likelihood that both will be documented so that they are understandable. For the data analyst, it is tempting to

forgo the time required to create a codebook and clear documentation of one's analyses because - at the moment of analysis the variable names and meaning of the analysis are readily available in memory. However, six months later when the editor requires additional analysis, it can be hard to recall what VAR0001 and VAR0002 mean. Careful documentation of analyses and methods, along with data codebooks increase reproducibility by making it easier for someone else, including your future self, to understand and interpret the data and analysis scripts (Nosek, 2014). Sharing with collaborators is a means of motivating this solid practice that otherwise might feel dispensable in the short-term, but becomes a substantial time saver in the long-term.

Archive Materials and Data

Wicherts, Borsboom, Kats and Molenaar (2006) tried to obtain original datasets from 249 studies in order to reproduce the reported results. They found that the major barrier to reproducibility was not errors in the datasets; it was not being able to access the dataset at all. Just 26% of the datasets were available for reanalysis. In a more recent case, Vines and colleagues (2013) found that just 23% of 516 requested datasets were available, and the availability of datasets declined by 7% per year over the 20-year period they studied. Further, Vines and colleagues observed that the working rate of email addresses of corresponding authors fell by 7% per year over the same span. In sum, reproducibility and reanalysis of data is most threatened by the gradual loss of information through the regular laboratory events of broken machines, rotating staff, and mismanagement of files.

The potential damage for one's own research and data management are substantial. Researchers routinely return to study designs or datasets as their research programs mature. If those materials and data are not well maintained, there is substantial loss of time and resources trying to recover prior work. Considering the substantial resources invested in obtaining the data and conducting the research, these studies reveal a staggering degree of waste of important scientific resources. It does not need to be this way. There are now hundreds of

repositories available for archiving and maintaining research materials and data. If researchers adopt the strategy of sharing research materials and data among collaborators, then it is a simple step to archive those materials for purposes of preservation and later recovery.

Reporting

Disclose Details of Methods and Analysis

The APA Manual, the style guide for report writing, suggests that methods sections need to report sufficient detail so that a reader could reasonably replicate the study (APA Manual 6th Edition, 2010, p. 29). And, for reporting analyses, authors should "mention all relevant results, including those that run counter to expectation" (APA Manual 6th Edition, 2010, p. 32) and "include sufficient information to help the reader fully understand the analyses conducted", minimally including "the per-cell sample size, the observed cell means (or frequencies of cases in each category for a categorical variable), and the cell standard deviations, or the pooled within-cell variance" (APA Manual 6th Edition, 2010, p. 33).

Even a cursory review of published articles reveals that these norms are rarely met in modal research practices. And, yet, complete methodology and analysis description is vital for reproducibility. In the ideal report, a reader should be able to identify the conditions necessary to conduct a fair replication of the original research design, and have sufficient description of the analyses to reproduce them on the same or a new dataset. Without full description, the replication will inevitably contain many unintended differences from the original design or analysis that could interfere with reproducibility.

There are occasions in which some critical elements of a research design will not fit into a written report - either because of length restrictions or because the design elements cannot be described in words effectively. For both, there are readily available alternatives. Supplementary materials, which most journals now support online during review and after publication, allow more comprehensive descriptions of methodology. Photo or video simulations of research designs can clarify key elements that are not easy to describe. What should be

included in methods descriptions will vary substantially across research applications. An example of guidelines for effective reporting of methods and results was developed by the Research Committee at the Tilburg University Social Psychology Department (2013; http://www.academia.edu/2233260/Manual_for_Data_Sharing_-_Tilburg_University).

While preparing comprehensive descriptions of research methods may add to the time required to publish a paper, it also has the potential to increase the impact of the research. Independent scientists interested in replicating or extending the published findings may be more likely to do so if the original report describes the methods thoroughly. And detailed methodological reporting increases the chances that subsequent replication attempts will faithfully adhere to the original methods, increasing the odds that the findings are replicated and the original authors' reputations enhanced.

Follow Checklists for Good Reporting Practices

The APA manual provides specific guidance for style and general guidance for content of reporting. Following revelations of substantial opportunity for (Simmons et al., 2011) and exploitation of (John et al., 2012) flexibility in data analysis and reporting, new norms are emerging for standard disclosure checklists of research process. Following Simmons et al. (2011) and LeBel and colleagues (2013), *Psychological Science* has established four items that must be disclosed in all its articles (Eich, 2013): (1) how samples sizes were determined, (2) how many observations, if any, were excluded, (3) all experimental conditions that were tested, including failed manipulations, and (4) all items and measurements that were administered. These are easy to implement for any report, regardless of journal, and they disclose important factors where researchers may take advantage of, or avoid, leveraging chance in producing research findings.

More generally, checklists can be an effective way of making sure desired behaviors are performed (Gawande, 2009). There are a variety of checklists emerging for particular research practices and reporting standards. For example: (1) CONSORT is a checklist and reporting

standard for clinical trials (Moher et al., 2010); (2) the ARRIVE checklist has a similar purpose for animal research (Kilkenny, Browne, Cuthill, Emerson, & Altman, 2010); (3) Kashy and colleagues (2009) provided recommendations for methods and results reporting for authors of articles in *Personality and Social Psychology Bulletin* that have wider applicability; (4) Poldrack and colleagues (2008) offered a reporting standards checklist for fMRI analysis pipelines; (5) Klein and colleagues (2012) suggested standard reporting of participant and experimenter characteristics for behavioral research; (6) Brandt and colleagues (2014) offer 36 questions to address for conducting effective replications; (7) members of a laboratory and course at the University of Virginia generated three brief checklists for managing research workflow, implementing a study, and reporting the results to facilitate transparency in research practices (Open Science Collaboration, 2012b); and (8) the headings of this chapter can serve as a checklist for reproducibility practices as presented in Table 1.

Share Materials and Data with the Scientific Community

When Wicherts et al. (2006) received just 26% of requested datasets of published articles, they speculated that the low response rate was primarily a function of the time and effort it takes for researchers to find, prepare, and share their data and code books after publication. It is also possible that some were reluctant to share because the present culture perceives such requests as non-normative and perhaps done in effort to discredit one's research. Explicit, widespread embrace of openness as a value for science may help neutralize this concern. More directly to the point, when materials and data are archived from the start of the research process, it will be much easier for researchers to adhere to data-sharing requests.

Some archiving solutions make it trivially easy to move a private repository into public or controlled access. Researchers who shared their materials and data with collaborators in a web-based archive can select which of those materials and data to release to the public. This may be particularly helpful for addressing the file-drawer effect. For those studies that researchers do not intend to write up and publish, their presence in a registry and public access

to the materials and data ensures their discoverability for meta-analysis and assists researchers investigating similar questions in informing their research designs.

Sharing research materials and data is not without concern. First, researchers may be concerned about the amount of work that will be required from them once method and data sharing becomes the standard. However, if researchers incorporate the expectation of sharing materials and data with collaborators, and potentially more publicly, into their daily workflow, sharing becomes surprisingly easy and encourages good documentation practices that assists the researcher's own access to the materials and data in the future. This may even save time and effort in the long run.

Second, some data collections require extraordinary effort to collect and are the basis for multiple publications. In such cases, researchers may worry about the cost:benefit ratio of effort expended to obtain the data against the possibility of others' using the data before they have had sufficient time to develop their own published research from it. There are multiple ways to address this issue including: (a) releasing the data in steps exposing only the variables necessary to reproduce published findings; (b) establishing an embargo period during which the original authors pursue analysis and publication, but then open the data to others following that; or (c) embracing the emerging evidence that open data leads to greater scientific output and impact (Piwowar & Vision, 2013). Further, there are now journals such as the *Journal of Open Psychology Data* (http://openpsychologydata.metajnl.com/) and organizational efforts like Datacite (http://www.datacite.org/) that aim to make datasets themselves citable and a basis for earning reputation and citation impact.

Finally, the most potent concern is protecting participant privacy with human participant research. At all times, the individual researcher bears fundamental responsibility to meet this ethical standard. Data sharing cannot compromise participants' rights and well-being. For many research applications, making the data anonymous is relatively easy to do by removing specific variables that are not essential for reproducing published analyses. For other research

applications, a permissions process may be needed to obtain datasets with sensitive information.

In summary, reproducibility will be maximized if the default practice for materials and data is to share them openly. Restrictions on open data are then the exceptions to the default practice. There are many defensible reasons for closing access, particularly to data. Those reasons should be made explicit in each use case.

Report Results to Facilitate Meta-Analysis

A single study rarely settles a scientific question. Any single finding could be upwardly or downwardly biased (i.e., larger or smaller than the true effect, respectively) due to random or systematic sources of variance. Meta-analysis addresses this concern by allowing researchers to model such variance and thereby provides summary estimates worthy of increased confidence. However, if the sources used as input to meta-analyses are biased, the resulting meta-analytic estimates will also be biased. Biased meta-analytic findings are especially problematic because they are more likely than primary studies to reach scientific and practitioner audiences. Therefore, they affect future research agendas and evidence-based practice (Kepes & McDaniel, 2013).

Individual researchers can facilitate effective aggregation of research evidence by (a) making their own research evidence - published and unpublished - available for discovery by meta-analysts, and (b) structuring the results reports so that the required findings are easy to find and aggregate. The first is addressed by following the archiving and sharing steps described previously. The second is facilitated by ensuring that effect sizes for effects of interest and all variable pairs are available in the report or supplements. For example, authors can report a correlation matrix, which serves as an effect size repository for a variety of variable types (Dalton et al., 2012).

Programmatic Strategies

Replicate-and-Extend

The number of articles in psychology explicitly dedicated to independent, direct replications of research appears to be 1% or less of published articles (Makel, Plucker & Hegarty, 2012). It would be easy to conclude from this that psychologists do not care about replicating research, and that journals reject replication studies routinely because they do not make a novel enough contribution. However, even when researchers are skeptical of the value of publishing replications, they may agree that replication-and-extension is a profitable way to meet journals' standards for innovation while simultaneously increasing confidence in existing findings.

A great deal of replication could be carried out in the context of replicate-and-extend paradigms (Nosek et al., 2012; Roediger, 2012). Researchers may repeat a procedure from an initial study within the same paper, adding conditions or measures, but also preserving the original design. For example, a Study 2 might include two conditions that replicate Study 1 (disgust prime and control), but also add a third condition (anger prime), and a second outcome measure. Thus, Study 2 offers a direct replication of the Study 1 finding with an extension comparing those original conditions to an anger prime condition. This provides greater certainty about the reproducibility of the original result than a Study 2 that tests the same hypothesis after changing all the operationalizations.

Participate in Crowdsourced Research Projects

The prior section alluded to the fact that some challenges for reproducibility are a function of the existing culture strongly prioritizing innovation over verification (Nosek et al., 2012). It is not worth researchers' time to conduct replications or confirmatory tests if they are not rewarded for doing so. Similarly, some problems are not theoretically exciting, but would be practically useful for developing standards or best practices for reproducible methodologies. For example, the scrambled sentence paradigm is used frequently to make particular thoughts accessible that may influence subsequent judgment (e.g., Bargh, Chen, & Burrows, 1996). Despite being a frequently used paradigm, there is no direct evidence for which procedural features optimize the

paradigm's effectiveness, and there is great variation in operationalizations across studies. Optimizing the design would be very useful for maximizing power and reproducibility, but conducting the required studies would be time consuming with uncertain reward. Finally, some problems are acknowledged to be important, but are too large to tackle singly. It is difficult for individual researchers to prioritize doing any of these when confronted with the competitive nature of getting a job, keeping a job, and succeeding as an academic scientist.

One solution for managing these incentive problems is crowdsourcing. Many researchers can each contribute a small amount of work to a larger effort. The accumulated contribution is large, and little risk is taken on by any one contributor. For example, the Reproducibility Project: Psychology is investigating the predictors of reproducibility of psychological science by replicating a large sample of published findings. Almost 200 researchers are working together with many small teams each conducting a replication following a standardized protocol (Open Science Collaboration, 2012a, in press).

Another approach is to incorporate replications into teaching. This can address the incentives problem and provide pedagogical value simultaneously (Frank & Saxe, 2012; Grahe et al., 2012). The CREP project (https://osf.io/wfc6u/) identifies published research for which replication could be feasibly incorporated into undergraduate methods courses. Also, the Archival Project (http://archivalproject.org/) integrates crowdsourcing and pedagogical value with a crowdsourced effort to code articles to identify the rates of replications and characteristics of methods and results in the published literature.

Request Disclosure as a Peer Reviewer

Individual researchers can contribute to promoting a culture of reproducibility by adapting their own research practices, and also by asking others to do so in the context of their role as peer reviewers. Peer reviewers have influence on the articles they review and, in the aggregate, on editors and standard journal practices. The Center for Open Science (http://cos.io/) maintains a standard request that peer reviewers can include in their reviews of

empirical research to promote a culture of transparency:

> "I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open Science [see http://osf.io/hadz3]. I include it in every review."

Including this as a standard request in all reviews can (a) show the broad interest in making the disclosure standard practice, and (b) emphasize it as a cultural norm and not an accusatory stance toward any individual. A culture of transparency works best if all members of the culture are expected to abide by it.

Implementing These Practices: An Illustration with the Open Science Framework

There are a variety of idiosyncratic ways to implement the practices discussed in this chapter. Here we offer an illustration using an open source web application that is maintained by the Center for Open Science called the Open Science Framework (OSF; http://osf.io/). All of the practices summarized can be supported by the OSF.

Organize a research project. The research workflow in the OSF begins with the creation of a project. The creator provides the title and description, uploads files, write documentation via the wiki, and add contributors. Users can create project components to organize the project into conceptual units. For example, a survey research project might include one component for study design and sampling procedures, another for survey instruments, a third for raw data, a fourth for data analysis, and a fifth for the published report. Each component has its own list of contributors and privacy settings. For example, the lead investigators of a project may decide to grant access to the data coding components to research assistant collaborators, but to deny those collaborators permission to modify the data analysis components.

Create an analysis plan. Once the investigator has organized her project and added her contributors, she might then add her analysis plan. The investigator might create a new

component for the analysis plan, upload analysis scripts and sample codebooks, and write a narrative summary of the plan in the component wiki.

Register study and materials. Once the investigator is ready to begin data collection, she might next register her study and materials. Materials are often used between studies and may evolve; registration at this point ensures that the exact materials used in the study are preserved. To do so, she would click a button to initiate a registration and provide some description about what is being registered. Once created, this registration becomes a frozen copy of the project as it existed at the moment it was registered. This frozen copy is linked to the project, which the researchers may continue to edit. Thus, by creating a registration, the investigator can later demonstrate that her published analysis matched her original plan--or, if any changes were necessarily, detail what was changed and why.

Keep records of analyses. As the research team collects data and conducts analysis, the tools used to generate the analysis and records of how those tools were used can be added to the data analysis component of the project. These might include analysis or data cleaning scripts written using Python, R, or SPSS, quality checking procedures, or instructions for running these scripts on new data. The OSF records all changes made to project components, so the research team can easily keep track of what changed, when it changed, and who changed it. Prior versions are retained and recoverable.

Share materials and data. At any point during the research life cycle, the team may choose to make some or all of their work open to the public. OSF users can make a project or one of its components public in a single step: clicking on the “Make Public” button on the dashboard of each project. Researchers can also independently control the privacy of each component in a project; for example, an investigator may decide to make her surveys and analysis plan public, but make her raw data private to protect the identities of her research participants.

Replicate and extend. Once the investigator’s project is complete, independent

scientists may wish to replicate and extend her work. If the original investigator made some or all of her work public, other OSF users can create an independent copy (or a "fork") of her project as a starting point for their own investigations. For example, another OSF user might fork the original researcher's data collection component to use her surveys in a new study. Similarly, another researcher planning a meta-analysis might fork the original raw data or data analysis components of several OSF projects to synthesize the results across studies. The source project/component is maintained, creating a functional citation network--the original contributors credit is forever maintained.

Closing

We started this chapter concerning how to improve reproducibility with a question: "What can I do?" We intend the suggestions made in this chapter to provide practical answers to that question. When researchers pursue open, reproducible practices they are actively contributing to enhancing the reproducibility of psychological research, and to establishing a culture of "getting it right" (Nosek et al., 2012). Though adhering to these suggestions may require some adaptation of current practices by the individual researcher, we believe that the steps are minor, and that the benefits will far outweigh the costs. Good practices may be rewarded with general recognition, badges (https://osf.io/tvyxz/), and enhanced reputation, but ultimately the reward will be having contributed to a cumulative science via reproducible findings.

Glossary

Confirmatory research: Research in which data is gathered to test a priori hypotheses.

Exploratory research: Research in which data is gathered to determine whether interesting a posteriori hypotheses might be generated from the data.

File-drawer effect: The bias introduced into the scientific literature by a tendency to publish

positive results but not to publish negative results.

Meta-analysis: The use of statistical methods to combine results of individual studies.

Power: The probability that the test will reject the null hypothesis when the alternative hypothesis is true.

Pre-registration: Registering which variables will be collected, how many participants will be tested, and how the data will be analyzed *before* any participants are tested.

References

American Psychological Association (APA). (2010). Publication Manual of the American Psychological Association (6th ed.). Washington, DC: Author.

Austin M. A., Hair M. S., Fullerton S. M. (2012). Research guidelines in the era of large-scale collaborations: An analysis of Genome-wide Association Study Consortia. *American Journal of Epidemiology, 175*, 962–969. doi: 10.1093/aje/kwr441

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554. doi: 10.1177/1745691612459060

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244. doi:10.1037/0022-3514.71.2.230

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

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R.,…Van 't Veer, A. (2014). The Replication Recipe: What Makes for a Convincing Replication? *Journal of Experimental Social Psychology, 50,* 217-224. doi: 10.1016/j.jesp.2013.10.005

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. doi: 10.1038/nrn3475

Carp, J. (2012a). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300. doi:10.1016/j.neuroimage.2012.07.004

Carp, J. (2012b). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, *6,* 149. doi: 10.3389/fnins.2012.00149

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Social Psychology, 65*, 145-153. doi: 10.1037/h0045186

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. *Personnel Psychology*, *65*, 221-249. doi: 10.1111/j.1744-6570.2012.01243.x

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R.,…Van der Weyden, M. B. (2004). Clinical trial registration: A statement from the international committee of medical journal editors. *Lancet, 364*, 911-912.

Eich, E. (2013). Business not as usual. *Psychological Science*. Advance online publication. doi: 10.1177/0956797613512465

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7,* 600-604. doi: 10.1177/1745691612460686

Fuchs, H. M., Jenny, M. A., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science*, *7*, 639-–642. doi:10.1177/1745691612459521

Gawande, A. (2009). *The checklist manifesto*. New York, NY: Metropolitan Books

Grahe, J., Brandt, M. J., IJzerman, H., & Cohoon, J. (2014). Collaborative Replications and Education Project (CREP). Retrieved from Open Science Framework, http://osf.io/wfc6u.

Grahe, J. E., Reifman, A., Hermann, A. D., Walker, M., Oleson, K. C., Nario-Redmond, M., & Wiebe, R. P. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science, 7*, 605-604. doi: 10.1177/1745691612459057

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20. doi: 10.1037/h0076157

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

Jaeger, R. G., & T. R. Halliday (1998). On confirmatory versus exploratory research. *Herpetologica 54*:(Suppl.). 564–566.

John, L. K., Loewenstein, G., & Prelec (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi: 10.1177/0956797611430953

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin, 35*(9), 1131-1142. doi: 10.1177/0146167208331253

Kemerer, C. F., & Paulk, M. C. (2009). The impact of design and code reviews on software quality: An empirical study based on PSP data. *IEEE Transactions on Software Engineering, 35*, 534-550. doi: 10.1109/TSE.2009.27

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in I-O psychology? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *6*, 252-268. doi: 10.1111/iops.12045

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biology, 8*(6), e1000412. doi 10.1371/journal.pbio.1000412

Klein, O., Doyen, S., Leys, C., Magalhäes de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science, 7*(6), 572-584. doi: 10.1177/1745691612463704

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, S., Bernstein, M. J.,…Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*.

Kolawa, A., & Huizinga, D. (2007). *Automated defect prevention: Best practices in software management*. New York, NY: Wiley-IEEE Computer Society 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., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology. *Perspectives on psychological science*, *8*(4), 424-432. doi: 10.1177/1745691613491437

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

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A.,…Van der Laan, M. (2014). Promoting transparency in social science research. *Science, 343*, 30-31. doi: 10.1126/science.1245317

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtsche P. C., Devereaux, P. J.,…Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomized trials. *Journal of Clinical Epidemiology, 63*(8), e1-37. doi: 10.1016/j.jclinepi.2010.03.004

Nosek, B. A. (2014). Improving My Lab, My Science with the Open Science Framework. *APS Observer*.

Nosek, B. A., Spies, J. R., & 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 (2012a). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657-660. doi: 10.1177/1745691612462588

Open Science Collaboration (2012b). Checklists for Research Workflow. Retrieved from osf.io/mv8pj.

Open Science Collaboration (in press). The Reproducibility Project: A Model of Large-Scale Collaboration for Empirical Research on Reproducibility. In V. Stodden, F. Leish, & R. Peng (Eds.), *Implementing Reproducible Computational Research (A Volume in the R Series)*. New York, NY: Taylor & Francis.

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

Piwowar, H. A., & Vision, T. J. (2013) Data reuse and the open data citation advantage. *PeerJ* 1: e175. doi: 10.7717/peerj.175

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage, 40*(2), 409-414. doi: 10.1016/j.neuroimage.2007.11.048

Popper, K. R. (1959). *The Logic of scientific discovery*. London: Hutchinson.

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641. doi:10.1037/0033-2909.86.3.638

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309-316. doi: 10.1037/0033-2909.105.2.309

Simmons, J., Nelson, L., & 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

Smith, E. R., & Semin, G. R. (2004). Socially situated cognition: Cognition in its social context. *Advances in Experimental Social Psychology, 36,* 53-117.

Stebbins, R.A. (2001). *Exploratory research in the social sciences*. Thousand Oaks, CA: Sage. doi: 10.4135/978141298424

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T.,...Rennison, D. J. (2013). The availability of research data declines rapidly with article age. *Current Biology, 24*(1), 94-97. doi: 10.1016/j.cub.2013.11.014

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638. doi: 10.1177/1745691612463078

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726-728. doi: 10.1037/0003-066X.61.7.726

Table 1

Increasing the reproducibility of psychological research across the research lifecycle

Project Planning

(1) Use high powered designs

(2) Create an analysis plan

Project Implementation

(3) Determine data collection start and stop rules

(4) Register study and materials

Data Analysis

(5) Perform confirmatory analyses first

(6) Conduct exploratory analysis for discovery, not for hypothesis testing

(7) Test discoveries with confirmatory designs

(8) Keep records of analyses

(9) Share and review data and analysis scripts among collaborators

(10) Archive materials and data

Reporting

(11) Disclose details of methods and analysis

(12) Follow checklists for good reporting practices

(13) Share materials and data with the scientific community

(14) Report results to facilitate meta-analysis

Programmatic Strategies

(15) Replicate-and-extend

(16) Participate in crowdsourced research projects

(17) Request disclosure as a peer reviewer