Improving Psychological Science through Transparency and Openness: An Overview

Andrew H. Hales[1], Eric D. Wesselmann[2], Joseph Hilgard[2]

[1]University of Virginia

[2]Illinois State University

Author Note

Correspondence concerning this article should be addressed to Andrew Hales, Frank Batten School of Public Policy and Leadership, University of Virginia, Charlottesville, VA, 22902; ahales@virginia.edu.

We thank Thomas Critchfield for valuable comments on a draft of this article.

Word Count: 9,642

Abstract

The ability to independently verify and replicate observations made by other researchers is a hallmark of science. In this article, we provide an overview of recent discussions concerning replicability and best practices in mainstream psychology with an emphasis on the practical benefits to both researchers and the field as a whole. We first review challenges individual researchers face in producing research that is both publishable and reliable. We then suggest methods for producing more accurate research claims, such as transparently disclosing how results were obtained and analyzed, preregistering analysis plans, and publicly posting original data and materials. We also discuss ongoing changes at the institutional level to incentivize stronger research. These include officially recognizing open science practices at the journal level, disconnecting the publication decision from the results of a study, training students to conduct replications, and publishing replications. We conclude that these open science practices afford exciting low-cost opportunities to improve the quality of psychological science.

Keywords: Replication, Reproducibility, Preregistration, Meta-Analysis

Improving Psychological Science through Transparency and Openness: An Overview

Behavior is complicated, and understanding its antecedents and consequences has challenged the best minds since at least the time of ancient philosophers. Unlike those philosophers, rather than rely on armchair theorizing, today's psychologists apply the scientific method; that is, they assume that they can better understand psychological phenomena through *observation*. Further, as scientists, psychologists value the principle that no claim has to be taken on authority – any observation can be independently verified by a curious skeptic.

In recent years, many psychologists have shifted from valuing *the principle* of independent observation to enacting *the practice* of independent observation. Psychological science is experiencing a crisis of confidence (Spellman, 2015). Concerns about the state of the science were prompted by the realization that long-common research practices could lead to false-positive results (Simmons, Nelson, & Simonsohn, 2011) and that a troubling number of findings are difficult to replicate (Open Science Collaboration, 2015). Happily, the last few years have seen a surge in interest in the topic of replicability and recommendations for best-practices to address this crisis of confidence. The result: actionable suggestions that are generally easy to implement and likely to improve the state of the science.

In the current article, we discuss these recommendations, focusing on how to implement them in practice and the ways in which they can increase the confidence of both psychologists and the general public in psychological science. These recommended changes often come with minimal cost and notable benefit for the researchers who use them. Before considering these recommendations, we will briefly review the challenges and conflicting contingencies faced by individual researchers that could lead to irreplicable findings, and consider them in the context of

behavioral science. We then consider 1) behavioral changes that can (and are) being implemented by *individual researchers* to improve replicability, followed by 2) the behavioral changes that can (and are) being implemented by *journals/institutions/organizations* to improve replicability. It is important to note that behavioral changes at these two levels are not independent; as top-down guidelines are issued from journals and organizations, contingencies facing individuals will change, and as replicable practices become more common, contingencies facing institutions will change.

Much of our analysis is based on the logic and consequences of traditional null hypothesis statistical testing. We recognize that many readers of the present journal approach their research questions in a way that does not incorporate formal hypothesis testing (e.g., Baron & Perone, 1998; Perone, 2018). However, even among those who are predisposed toward other approaches, hypothesis testing is growing in prevalence, if for no reason other than making results persuasive to those outside of behavior analysis (e.g., Davison, 1999; Fox, 2018). The recommendations that follow here will be most relevant to those who already incorporate statistical modeling and hypothesis testing into their work. However, we hope it will also provide insights and ideas for those who take alternative approaches, including qualitative research.

The Challenge Facing Researchers

When a researcher wishes to publish research demonstrating a causal relationship between two variables, it is their responsibility to demonstrate that changes in the dependent variable are due to changes in the independent variable. This requires careful attention to potential confounding variables, as these would provide alternative, and often less interesting,

accounts for the results. However, there is a possible account that is even less interesting than confounds - that the results were a fluke due to chance (or, a *Type I error*; Abelson, 1995)[1].

Two samples drawn from the same population will rarely be identical. So, how different must they be at the end of a study to rule out mere chance as explaining the difference? Traditional publishing standards in many fields require that researchers submit their results to *null hypothesis statistical testing*, by first entertaining the assumption that the groups are really drawn from the same population, and then calculating the probability of a difference as large as, or larger than, the observed difference under such an assumption. If the probability is less than 5%, the researcher rejects chance as explaining the difference. From there they may consider more interesting accounts (such as confounds or, preferably, a tidy effect of the independent variable). Accordingly, in 5% of cases that there is no effect, a researcher will falsely conclude that there is an effect.

Contrary to common misconception, a p value of $< .05$ does not mean that the null hypothesis has < 5% chance of being true. Readers may be familiar with criticisms levied by Branch (2014), who correctly argued that a p value does not represent "an indication of the probability of replication nor of the probability that the results are due to chance" (p. 257). Branch also correctly noted that p values are commonly misunderstood, which leads to "malignant side-effects," such as focusing narrowly on the dichotomy of *whether or not* there is an effect, instead of the potentially more interesting question of the *size* of the effect. But, in our

[1] *Chance*, here refers to fluctuating factors that are uncontrolled/unaccounted for in an experiment, as well as the sampling procedures used to obtain one particular subset of a population, as opposed to another. It does not signify that behavior itself is *ultimately* random or probabilistic. As Poincaré observed, "Every phenomenon, however trifling it be, has a cause, and a mind infinitely powerful and infinitely well-informed concerning the laws of nature could have foreseen it… Chance is only the measure of our ignorance." (1914/1952, p. 65). To say that a difference is "due to chance" is to say that, for unspecifiable reasons not involving systematic treatment, one's sample is atypical of the population.

view, it is an overstatement to conclude that *p* values have *no bearing* on the reliability of a finding - at least when they are used and interpreted correctly (which we note, is not a trivial qualification). A small *p* value, compared to a large *p* value, is indeed associated with greater probability of the null hypothesis being false, *all else being the same*. In a recent large-scale replication of 21 experiments, researchers were able to predict with some success which results would replicate; one cue that was used was the *p* value of the original study (Camerer et al., 2018). Our perspective more closely resembles that of Fox (2018); in most circumstances, behavior analysis is studying clear effects with considerable within-subject sample sizes, and so *p* values will be small, and graphical data compelling. However, in many cases, the *p* value can be a helpful safeguard against subjectivity, particularly with regard to less-well-powered tests like between-subject differences.

Some behavior analysts have avoided the use of statistical models, fit statistics, and *p* values, preferring instead to study clear effects with many repeated within-subject trials, which are then replicated in another organism. The main output of these studies is often a visually compelling graphical display of behavior - that is, a figure that is so compelling that it passes the interocular trauma test, meaning that the effect hits you right between the eyes (Edwards, Lindman & Savage, 1963). Indeed, strong graphical approaches (particularly of complete original observations, not just group averages) is a strength of behavior analysis - one that ought to be exported to other areas of psychology (e.g., McCabe, Kim, & King, 2018).

There remains, however, an element of subjectivity in this approach that leaves it insufficient on its own (e.g., Fisch, 2001). What does it mean to say that an effect has replicated in another organism? In the absence of formal models, we risk letting subjectivity guide our

decisions. The use of multi-level modeling, or similar techniques, in behavior analysis seems promising (e.g., Kirkpatrick, Marshall, Steele, & Peterson, 2018; Huitema & McKean, 2000); it provides a way to describe data featuring repeated trials nested within several organisms. Using these models, one is able to describe the parameter values and the uncertainty around them.

In principle, formal hypothesis testing limits the rate at which researchers report effects where there is no effect. In practice, however, things get messy. When researchers sit down to analyze their data, they face many choices. Do they exclude outliers? If so, which ones? Do they use statistics to adjust for characteristics of the individual? How exactly do they calculate the dependent variable? Do they need fifty observations, or two hundred?

These choices are referred to as *researcher degrees of freedom* (Simmons et al., 2011), *p-hacking, questionable research practices,* (John, Loewenstein, & Prelec, 2012) or navigating *a garden of forking paths* (Gelman & Loken, 2014). Because journals tend to treat $p = .05$ as a hard cutoff, researchers are motivated to obtain results that are statistically significant, and they have an incentive to probe different combinations of decisions and report only those that yield positive results. Even though most researchers are well-intentioned, the motivation to achieve a significant result can bias them towards justifying a particular analytic choice after the fact (Nosek, Spies, & Motyl, 2012). These researcher degrees of freedom can substantially inflate Type I error rates.

The problem, in general, is not necessarily that a particular set of analyses that are reported are inherently inappropriate; these analyses may well have been the most appropriate analyses had they been selected before seeing the results. The problem is that for null hypothesis statistical testing to work without producing excess false positives, these decisions need to be

made *independent of their actual effect on the p value*. Without this independence, researcher degrees of freedom make *p* values uninterpretable regardless of whether one is studying a null effect or a nonzero effect. Even if the underlying effect happens to be real, it is an error to make this conclusion based on methods that increase the chance of false positives (Hales, 2016).

The magnitude of this problem was memorably illustrated by Simmons and colleagues (2011). Simmons and colleagues randomly assigned participants to listen to either *When I'm Sixty-Four* by The Beatles or a control song. After the song, participants indicated their age. The authors reported the impossible conclusion that listening to The Beatles made participants significantly younger. Simmons and colleagues were able to obtain this result by exploiting researcher degrees of freedom: dropping conditions, selecting among outcomes, trimming outliers, strategically using covariates, and using the results to decide when to end data collection. Their point: Anything can be made statistically significant through undisclosed flexibility in data analysis and reporting.

A further complication is that scientific journals favor publication of statistically significant results. Non-significant results are likely to be discarded as uninteresting. As a result, the published literature consists primarily of significant results (indeed, more than 80% of psychology articles report support for their hypothesis; see Fanelli, 2012), while nonsignificant results may be left in a file drawer somewhere. This *file-drawer problem* (Rosenthal, 1984) leads to *publication bias*, the overestimation of the size and robustness of phenomena due to censoring of nonsignificant results. This preference for significant results may not only censor non-significant results from report, but it may also encourage researchers to turn non-significant

results into significant results through p-hacking. The result is a literature that appears to show strong evidence for a particular effect, when in fact the true evidence is much weaker.

Clearly this is not an ideal way for a scientific community to produce and share research findings. What can be done?

Individual Behaviors

Individual researchers can change their behavior in ways that increase the replicability of their findings and the trust other researchers give their research. These behaviors, described below, give readers more information, constrain researcher degrees of freedom, and improve the ratio of true findings to false positives.

Transparent Disclosure

Part of the problem with statistically-guided hypothesis testing is that the actions that qualify as researcher degrees of freedom are hidden from public view; hence their impact on research conclusions remains obscure. Several new tools give greater insight into the research process, allowing readers to better evaluate the strength of evidence.

Disclosure of all manipulations and measures. One simple and direct action that researchers can take to increase confidence in replicable research findings is to disclose whether or not they exploited researcher degrees of freedom. Simmons and colleagues (2012) introduced a straightforward 21-word solution: authors may state at the beginning of the method section "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study." Simmons and colleagues point out that such a disclosure statement is analogous to food-labeling practices. Marketers of food products that are sugar-free

are allowed to proudly display so on their packaging. Researchers should be encouraged to do the same.

Open data and materials. Another opportunity for increasing the transparency and reproducibility of studies is for researchers to provide open access to their study materials and data files. The Open Science Framework (https://osf.io) allows researchers to store materials, data, and analysis scripts for private, internal use or for public access. This service is useful as an online hub for cross-laboratory collaborations (see https://osf.io/env5w/ for an example). Many journals also host online supplemental materials sections which are convenient for sharing additional research materials. Figshare (https://figshare.com/) provides an online platform for data management and sharing, and GitHub (https://github.com/) provides version control and cloud storage of data and analysis code.[2] Sharing research materials has the benefit of enabling high-fidelity replications (Funder et al., 2014) and can increase the efficiency of scientific discovery (Perrino et al., 2013). Posting data also has the benefit of making cases of research fraud easier to detect (Simonsohn, 2013).

Of course, it is not always practical to share study materials or results publicly. For example, some data are proprietary, or sometimes sharing data may compromise the privacy of research subjects. Indeed for this reason many recent journal reforms stop short of making this a requirement (Eich, 2014; Grahe, 2018; Vazire, 2016). A primer by Meyer (2018) provides

[2] Version control allows users to maintain a history of their files and to fall back to any previous version. This is useful in the case that one deletes material from a manuscript that should be later added back in, changes something in analytic code that accidentally breaks everything, or accidentally saves changes to the original raw data file. Version control eliminates the need for a cluttered folder full of files named “manuscript.docx”, “manuscript_final.docx”, and “manuscript_final2.docx”. It is generally not considered an open-science development, but it is a useful tool nonetheless.

guidance for how to maximize the accessibility of one's original data within the limits of ethical and legal considerations.

Constraints on generality. Researchers can preemptively outline the conditions under which they anticipate that a given finding can be replicated. Consider a study that seeks to directly replicate a previously published effect. Under current practices of null hypothesis significance testing, if the replication attempt yields non-significant results, the results are inherently ambiguous (Greenwald, 1975). Why was the new study not significant? One reason is that the original effect is simply not reliable; perhaps it was an inadvertent result of researcher degrees of freedom or a straightforward a Type I error. However, another reason is that the effect is reliable, but only occurs under conditions that the first study satisfied, but the second study did not. The replication may have drawn from a different population, or used different procedures. How can we tell which it is?

One proposed practice helps to preemptively disambiguate these situations: authors of original findings can disclose hypothesized *Constraints on Generality* (Simons, Shoda, & Lindsay, 2017). According to this approach, research reports should openly and candidly state the limits under which a given phenomenon is expected to be observed. Will the effect occur in different species? In members of different cultures? Should future researchers expect similar results to occur if measured remotely through the internet instead of in-person? Of course, researchers cannot say with complete confidence what the essential boundary conditions are for an effect; often this will be informed speculation (so these sections can be thought of as "likely" constraints on generality). The assumption is that the team that performed the research is in the best position to notice and share any potentially important contextual factors necessary to

replicate an effect (e.g., particular training apparatus, type of stimuli, laboratory conditions, etc.). By pre-specifying these conditions, researchers can help avoid situations where others fruitlessly attempt to replicate effects under conditions where they were never hypothesized to occur, only to discover this after the fact. At least one journal now currently advises authors to include a Constraints on Generality statement in their discussion section (Kitayama, 2017). Even when not required, authors would benefit by preemptively volunteering this information to other researchers, to pave the way for more successful replications down the road. Importantly, a researcher may believe that his or her effect is essentially universal. Simply saying so allows future replicators to proceed with more confidence that they will observe the same results. Additionally, it should be noted that whether these factors actually constrain the effect of an experimental variable is an empirical question. The point of these sections is not to convince an audience that a causal relationship depends on particular factors (which would most likely require additional experiments), but instead to alert the audience to conditions to keep in mind when performing a replication.

A further benefit of acknowledging constraints on generality may be to raise consciousness about the representativeness (or lack thereof) of samples typically used in psychological research. In a highly influential paper, Henrich and colleagues drew attention to the field's heavy reliance on samples that are *WEIRD*. That is, psychologists often use convenience samples made up university students who are disproportionately from societies that are Western, Educated, Industrial, Rich, and Democratic (Henrich, Heine, & Norenzayan, 2010). Greater attention to constraints on generality may motivate researchers to use more

representative samples, and, when this is not possible, draw more reserved conclusions about the generality of their findings.

Preregistration

One proposed method for improving the replicability of published research is *preregistration*, in which researchers commit to an analysis plan and post an unalterable, time stamped outline of the procedures that will be used to analyze the data (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Nosek, Ebersole, DeHaven, & Mellor, 2017). This practice constrains researcher degrees of freedom, thereby preventing inflation of Type I error. A typical preregistration explicitly states all manipulations, outcomes, and hypotheses, a stopping rule for data collection, criteria for eliminating outliers, and specific analyses that will be performed. Researchers can freely and easily preregister studies at aspredicted.org or on the Open Science Framework. Both resources include options for an embargo period, during which the information remains completely private, thereby addressing concerns that preregistration might put one at risk of having their research ideas appropriated by others.

Researchers may have a sense that conducting a preregistration is a time-consuming and onerous task. In our experience, in addition to being free and technically easy, preregistrations can be conducted in a reasonable amount of time. One need not prepare a manuscript-level introduction section. It is sufficient to simply specify the research method, hypotheses, and the specific measures and analyses that will be used to test those hypotheses.

In our experience it is ideal to preregister a study after performing at least some pilot testing, if possible. This enables a researcher to finalize their procedures, and also gives a clearer picture of potentially unexpected analytic issues that may arise. Researchers can then make these

analytic decisions *a priori*, ensuring that they do not affect the conclusions that are drawn from the results.

Benefits of Preregistration. Preregistration provides numerous benefits to researchers (Wagenmakers & Dutilh, 2016). Its primary benefit is that it distinguishes between confirmatory and exploratory hypothesis tests. This maintains correct Type I error rates for confirmatory tests, making those *p* values interpretable. In the absence of preregistration, researchers may fall victim to inadvertent cognitive biases causing them to believe that certain hypotheses were generated earlier in the research process than they actually were (Nosek, et al., 2012). Because the research process often spans many years, a preserved record of one's hypotheses and analysis plan helps to structure analysis and reporting and preserves accountability.

This signal of accountability is an asset to researchers in defending their hypotheses and analytic decisions to peer reviewers. At present, the use of one-tailed hypothesis tests is often regarded with suspicion: did the researcher truly never intend to interpret the opposite effect, or is the one-tailed test a last-ditch attempt to turn $p = .09$ into support for one's hypothesis? Similar concerns may arise regarding other analytical choices, such as which covariates are included in a statistical model (Wang, Sparks, Gonzales, Hess, & Ledgerwood, 2017) or how outcomes are aggregated. With a preregistration, authors can demonstrate to critics that these analytical choices were made in advance and were not motivated by the *p* value they yielded.

Being able to defend these decisions through preregistration can make data collection faster and more cost-effective. Researchers can preregister the ways they wish to spend their alpha level (two-tailed, one-tailed, or even one-and-a-half tailed; see Ramsey, 1990), maximizing statistical power to test a hypothesized effect. Researchers can even use *planned sequential*

analysis to check the *p* value after a certain number of subjects to decide whether to collect more data or to stop and publish (Lakens, 2014). In the absence of preregistration, these techniques may be criticized as increasing the Type I error rate, but with a preregistration, they are defensible and appropriate. These techniques may benefit researchers by yielding more statistical power per research dollar: preregistering a one-tailed test increases statistical power, and planned sequential analysis may provide an opportunity to stop data collection early.

Preregistration also makes analysis and writing easier by specifying analytic decisions beforehand. Considering the specific analysis in advance sometimes reveals fatal flaws in study design; taking time to preregister can often reveal these flaws before data collection, when there is still time to fix the design, rather than in data analysis or peer review, when it is much too late. Finally, making decisions can be a frustrating experience (e.g., Iyengar & Lepper, 2000). We have found that making these decisions ahead of time in preregistration can make the data analysis process more enjoyable and less confusing on the back end.

We must emphasize that preregistration *does not* prevent publication of exploratory findings (Nosek et al., 2017). Exploratory findings are often interesting and important, and researchers justifiably want to share them with readers. Indeed, in many cases, withholding the results of exploratory analyses would produce an incomplete or misleading portrayal of the results. Instead, the purpose of preregistration is to separate exploratory results from confirmatory results. Doing so will ultimately increase confidence in those findings that were predicted *a priori*. One appealing strategy to balancing exploratory and confirmatory research is to *explore small, confirm big* (Sakuluk, 2016). That is, one might explore novel hypotheses with

an initial modestly-sized sample, then confirm those findings with a preregistered, high-powered confirmatory study.

Power Analysis

For years, psychologists have been conducting studies that are underpowered (Maxwell, 2004). That is, studies have had a relatively poor chance of obtaining significant results given the magnitude of the effect studied and the number of observations collected. Conducting underpowered studies is inefficient and unwise; what is the value of a study that only has a 30% chance of detecting a true effect? Despite this poor power, most published studies nonetheless find significant results. This is demonstrably improbable and suggests that research findings have been distorted through publication bias or researcher degrees of freedom (see, e.g., Francis, 2012, Schimmack, 2012).

So how can one increase power? Observers in mainstream psychology echo points made independently by behavior analysts such as Perone (1998). The first steps are to use reliable measurements and within-subjects research designs when possible, as these minimize sources of noise (Greenwald, 1976; Kanyongo et al., 2007). Second, stronger manipulations tend to produce larger effect sizes and greater statistical power, so they should be favored when possible (Meyvis & van Osselaer, 2018). Finally, it is necessary to collect a sufficient number of observations. A common approach is to conduct an a priori *power analysis* by first estimating the likely effect size (e.g., based on similar research or heuristics, such as $d = .50$ is "medium"; Cohen, 1988), and then calculating the number of subjects necessary to achieve a desired probability of a significant result (often 80%, Cohen, 1988; though this is arbitrary, and researchers may not want to settle for a 1 in 5 chance of say, their dissertation, not finding a true effect). These

calculations can easily be performed using the free software G*Power (Faul, Erdfelder, Buchner, & Lang, 2009). Power analysis is also important, and also calculable, for small-*N* designs (see Kyonka, 2018). Resources are also available for more complicated research scenarios, such as multi-level modeling, when observations are clustered within individuals, pairs, or groups (e.g., Hennes & Lane, 2017).

Researchers often struggle with power analysis because, to estimate the number of observations required, researchers must specify not only the degree of desired power, but the true effect size. This is challenging because researchers do not know the true effect size; if they did, they would not need to do the study in the first place. However, one can often get a reasonable estimate by consulting the literature. Meta-analyses typically report the effect sizes of individual studies and the average effect size across studies. These can serve as estimates of the likely effect size; however, one should be aware that these may overestimate the likely effect size due to publication bias (see Vasishth & Gelman, 2017).

An alternative approach is to conduct a *sensitivity analysis* (also available in G*power), by starting with the number of subjects one is able to run based on resources, and then calculating the effect size which they have a reasonable probability of detecting. Using this approach, for example, a researcher intending to run 40 total participants in a two-condition between-subjects experiment would learn that, unless the effect they are studying is massive (almost a full standard deviation difference), they will have less than 80% power. Similarly, a researcher studying a smaller number of subjects could calculate, based on a fixed number of observations, how large an effect would have to be for it to be detectable.

Effect Sizes

Consider a researcher who conducts three studies seeking to detect a certain effect, yielding *p* values of .02, .14, and .07, with all of the mean differences in the same direction. According to traditional null hypothesis statistical testing, one study provides evidence for the hypothesized effect, and the other two do not; the count would be one success, and two misses (Giner-Sorolla, 2012).

However, more nuanced statistical thinking recognizes that although any given study may not be independently significant, it may still offer useful information about the population parameter which we are using samples to estimate (Cumming, 2014). According to this approach, rather than treating *p* values as all-important determinants of whether an effect *exists*, we can use standard errors to calculate *confidence intervals* around estimates of a mean, effect size, or other statistical parameter. This is a reasonable approach, given that any effect almost certainly exists; the difference between two population means is never *literally* zero "when carried to enough decimal places" (Jones & Tukey, 2000, p. 411). Smaller samples will produce wide and ambiguous confidence intervals; larger samples will produce narrow and precise confidence intervals. Researchers are free to interpret the effect sizes accordingly.

Internal Meta-Analysis. Consideration of the effect size allows researchers to take a *meta-analytic* perspective (e.g., Faith, Allison, & Gorman, 1996). Researchers can conduct an *internal* meta-analysis of a set of studies within a given paper to summarize the results (McShane & Böckenholt, 2017). Non-significant findings are not necessarily embarrassing evidence that one's effect does not exist (Giner-Sorolla, 2012) – they are to be expected as a consequence of sampling error, especially when power is modest.

An internal meta-analysis of a set of studies may well show that, although some results fall short of $p < .05$, confidence intervals around the total effect size indicate that there is, on average, an effect. This approach liberates researchers to report findings that are independently non-significant, but meta-analytically informative – findings that would have otherwise been destined for the file drawer. In this way, internal meta-analysis might help reduce publication bias (Lakens & Etz, 2017). For a user-friendly primer, see Goh, Hall, & Rosenthal (2016).

Another benefit of internal meta-analysis is that it can provide tests of *heterogeneity* in study results. Study results are heterogeneous when differences in the observed effect size across studies is greater than can be explained by sampling error alone. Researchers often have the sense that, given mixed significant and non-significant results, there is some difference in the method that has lead to meaningful differences in the results. Testing for heterogeneity can help to establish whether these are indeed substantial differences in the results or whether effect sizes are within sampling error of each other and the differences attributable to chance (see Goh, Hall, & Rosenthal, 2016; Higgins & Thompson, 2002).

Organizational efforts

Other open-science behaviors can and are being enacted at the level of journals and professional organizations. We detail changes in the publishing landscape that increase the transparency of research findings.

Reporting guidelines

In recent years, many journals and professional organizations have revised policies in ways that incentivize open and transparent research practices. For example, in response to a task force assembled by *The Society for Personality and Social Psychology*, the journal *Personality*

and Social Psychology Bulletin now encourages the reporting of effect sizes and confidence intervals and requires submissions to include online access to the verbatim research materials used in the study (Funder et al., 2014). Similar measures have been taken at a number of other journals. For example, the *Association for Psychological Science's* flagship journal *Psychological Science* no longer includes methods sections in word count limits so that authors can give detailed accounts of study procedures (Eich, 2014). *Psychological Science* also encourages the use of effect-size estimation and meta-analysis (Lindsay, 2015).

Badges

In a practice that is becoming increasingly common, many journals now use badges to mark papers that have taken steps towards being open and transparent (e.g., Grahe, 2014). For example, *Psychological Science* has three badges available: Preregistration, Open Materials, and Open Data. These badges recognize papers that meet these standards without mandating these standards as a requirement for publication (since it is sometimes not possible to post data/materials or preregister hypotheses; Finkel et al., 2015). These journal policies coincide with a sharp increase in the proportion of papers published that have open data (Kidwell et al., 2016). It is easy to imagine behavior analysis journals issuing similar guidelines and acknowledgement regarding preregistration, open data, and open materials.

Registered Reports

In addition to badges, many journals have introduced a Registered Report format designed to reduce publication bias. In a Registered Report, authors submit their study for peer review *before* data collection begins. Peer reviewers recommend changes to the methods, and when all are satisfied, the journal offers an *in-principle acceptance,* promising to publish the

results so long as the methods are conducted as promised. This publication format has two advantages: First, by placing peer review up front, it solicits criticism when there is still time to fix the methods, rather than when it is much too late. Second, by making publication decisions contingent on the methods, rather than the results, it alleviates publication bias and restricts researcher degrees of freedom. Registered Reports are offered at more than 100 psychology journals (see https://cos.io/rr/), and the new journal *Comprehensive Results in Social Psychology* uses this approach exclusively. Other journals may offer special issues of registered reports (e.g., Nosek & Lakens, 2013; 2014).

While registered reports may not be suitable for every research situation, they have a number of benefits. For example, they can protect authors from being stuck in an endless cycle of running additional analyses to satisfy reviewers. Eastwick (2018) tells of his experience with two papers in peer review: one under the conventional model of peer review, and another as a Registered Report. In the conventional paper, reviewers did not like the results, and Eastwick and colleagues ran more than a thousand additional post-hoc analyses to satisfy the reviewers. In the Registered Report, reviewers could only criticize the methods before observing the data; the study is conducted with the preregistered analysis plan that all agree provides the best test of the hypothesis. In Eastwick's experience, this is a better model of scientific review and report: "The data get to stand as they are, with no poking and prodding to try to make them say something else."

Publishing direct replications

Behavior analysis has a long tradition of valuing replication – both within and between subjects (e.g., Perone, 2018) and the *Journal of Applied Behavior Analysis* now has a section for

direct replications. Willingness to *directly* replicate effects may help account for why behavior analysis has not experienced the replication crisis as strongly as other fields (Branch, 2018; but see Francis, 2012 for when and why too many successful replications within a single paper can suggest non-transparent reporting).

Outside of behavior analysis, until recently, there was little incentive for researchers to conduct direct replications, and replications were rare (Makel, Plucker, & Hegarty, 2012). Insofar as results were replicated, they were *conceptually replicated* - the same research idea was tested using different methods. However, this makes interpretation difficult; if a conceptual replication fails to confirm the effect, is it because the effect does not exist, or is it because the changes in method caused the study to no longer test the relevant concept? The historical disregard for nonsignificant results has often left failed conceptual replications in the file drawer. Direct replications reduce these problems of interpretation by sticking closely to an original study's methods. Direct replications are important for the field to assess what is robust and reliable, and what is not (Zwann, Etz, Lucas, & Donnellan, 2017). More journals are becoming aware of the importance of direct replications and accepting more replications for publication (e.g., Kawakami, 2015).

Replications often make use of the registered report format. In these *registered replication reports*, researchers preregister large-scale direct replications to assess the size of a given effect (e.g. Simons, Holcombe, & Spellman, 2014). Such registered replication reports often draw collaborators from multiple laboratories in multiple states or countries. These large and diverse samples, combined with the pre-data-collection peer-review and preregistration emblematic of registered reports, provide very powerful tests of hypotheses.

The increasing prevalence of direct replications raises the question of how exactly one should go about conducting a direct replication (Brandt et al., 2014), and what one can conclude if the replication fails to find evidence for an effect. One issue that arises when conducting direct replications is exactly how *direct* to make them. Should researchers use the exact stimulus materials that were used in the original study? What if those materials are no longer available? What if the original stimulus materials do not make sense within the social or cultural context of the replication? For example, some subjects may require a different amount of time to achieve baseline learning, so repeating the *exact* procedures of the original study may make little sense. Further, with human subjects, sociopolitical or cultural changes may influence how individual cohorts respond to certain stimuli. For example, TV shows, music, or comedy used to induce certain moods cannot be guaranteed to have the same effects in a replication conducted years later (Fabrigar & Wegener, 2016).

As a general principle, a replicator should strive to reproduce the psychological conditions of the original study to the greatest extent possible, while still being as loyal to the original procedures as is reasonable. Incomplete documentation of procedures and contextual features in the original experiment can contribute to this difficulty. Consultation with the original authors is recommended, when possible.

Case example. Depending on the specific study, replication of all relevant methods and context can be a difficult task. For example, one of us conducted a replication of the classic Schachter (1951) study (Wesselmann et al., 2014; https://osf.io/s92i4/), which examined how groups responded to members who held deviate opinions. This was a resource-intensive study with few replications since its initial publication, yet it is a considered a canonical study in social

psychology and its main findings are often taken for granted (Berkowitz, 1971; Wahrman & Pugh, 1972). One key challenge was that the original study materials only existed in dissertation form, which had to be obtained in physical copy from Schachter's doctoral-granting institution. Further, the study involved training confederates for the group discussion, yet no script was available and the original author is deceased. Thus, Wesselmann and colleagues (2014) had to create their own confederate scripts, using the guidelines provided in the original dissertation and some example dialogue in a documentary reenactment of the original study (Mayer & Norris, 1970).

The other main change involved how participants were recruited. Schachter (1951) originally recruited participants under the impression they would be participating in multiple group discussion sessions over several months. Further, Schachter did not mention providing any compensation in the original study. Wesselmann and colleagues (2014) argued for providing participants compensation in order to increase the likelihood of obtaining the necessary number of participants based on their power analysis. Even with the compensation, the researchers did not believe they would be able to recruit the necessary sample size if participants thought they were volunteering for such a prolonged study. Thus, the researchers told participants both at the beginning of the experiment and towards the end of the discussion (immediately before the dependent measures) that they were actively pursuing funding and would hopefully be able to offer participants the opportunity for future sessions. These instructions were used to increase psychological realism that the group members may have future interactions and thus the presence of a deviate group member would have important implications for these potential interactions.

The main deviance-rejection effect replicated even with these method differences. Some ancillary findings did not replicate, however, and it is unclear to what extent that may be due to method differences or other potential differences between samples (e.g., cohort effects, university setting). Regardless, these method differences were identified and justified a priori in a required section of the preregistration. As such, authors interested in replicating classic findings should not be discouraged if they cannot emulate every aspect of the original study; they simply need to provide a rationale for their decisions up front.

Conducting direct replications. Now that replications are becoming more common, and outlets are available for them to be published, the question arises of who exactly should be conducting them. The short answer is: anyone who is curious or skeptical. But many have also noticed that conducting direct replications can be a useful practice for capable and motivated students who need experience with the research process (i.e., the mechanics of achieving ethics approval and recruiting subjects) but may not have yet achieved deep familiarity with a research literature to generate their own theoretically meaningful research questions (Frank & Saxe, 2012; Kochari & Ostarek, 2018). More generally, current discussions about replicability and best practices represent an opportunity to train new graduate and undergraduate students about the importance of transparency, meta-analytic thinking, and rigorous standards (Grahe et al., 2012; for a centralized coordination of such efforts, see https://osf.io/wfc6u/wiki/home/). Some instructors have actively built open science practices into their graduate methods courses, encouraging their students to focus on basic replications for course projects (Campbell, 2016; Hawkins et al., 2018). While many of the calls for changes in research practices have been

contentious (e.g., Fiske, 2016), it is possible to imagine that younger generations of students will look back and wonder what all the fuss was about.

Interpreting direct replications. Researchers are still discussing how best to interpret the results of a replication. At times, researchers have different definitions of what it means to replicate a study. When the Reproducibility Project: Psychology (Open Science Collaboration, 2015) reported that only 36% of replication results were statistically significant, many interpreted this as indicating that only 36% of studies replicated. Commentary from Gilbert, King, Pettigrew, and Wilson (2016) argued that the estimated replication rate was biased downwards due to infidelities of methods and poor statistical power. Commentary from Patil, Peng, and Leek (2017) suggested interpreting whether the replication fell within the 95% confidence interval of the original study; by this metric, one might argue that 77% of studies replicated. However, this latter approach seems flawed in that original findings using small sample sizes can have confidence intervals so wide that they would catch any non-negative result.

Some productive suggestions recommend testing the replication effect size directly against the original study's effect size. Simonsohn (2015) recommends a "small telescopes test". In this test, the author considers the sample size of the original study and calculates the effect size that would yield 33% power. If the replication result is significantly smaller than that, one might consider the original finding to be rejected. This approach requires a lot of data; Simonsohn recommends collection of 2.5 times the number of observations made in the original study (Simonsohn, 2015). An alternative Bayesian recommendation is to collect the best sample size one can and to use Bayes Factors to measure the strength of the evidence for the null or the

alternative (Dienes, 2014 Edwards et al., 1963). This has the advantage of avoiding dichotomous decision-making and quantifies evidence in natural units of betting odds.

Funding Agencies

In addition to changes being implemented at the journal level, there are also signs that granting agencies are placing more value on replicability. For example, many agencies are now explicitly encouraging scientists to make materials and data openly available, and the National Science Foundation recently advertised call for submissions relating to *achieving new insights through replicability and reproducibility* (Cook, 2018). A list of funding agencies' open science policies is available online (https://osf.io/kgnva/wiki/home).

Conclusion

Psychologists have long been aware, in the abstract, that the ability to reproduce the results of other researchers is central to what it means to be a science. More recently we have begun to internalize what this means in concrete terms, and how best to structure our practices to make this possible. Replicability is not the only characteristic of high-quality science (Finkel, Eastwick, & Reis, 2017), but it is certainly an important one.

Researchers now have a number of tools at their disposal to promote behaviors that will produce more replicable research. Perhaps the easiest to implement of these suggestions is that authors declare all measures and procedures (i.e., the 21-word solution). This requires little time and no learning of new technology. Other behavior changes offer bigger returns, and still require (relatively) little cost from researchers. Preregistration, and posting study materials and data online for others to access requires foresight, and a change in workflow, but is free, and once learned, easy to execute the next time. We noted several benefits of these practices above, but

perhaps the most compelling reasons to make these behavior changes are selfish: First, they enhance the efficiency with which researchers uncover truth about behavior, by preventing them from investing time and resources in effects that do not exist. And second, they enhance the rhetorical power of one's methodological arguments; simply put, all else being equal, transparent research is more persuasive than traditionally-published research.

Certainly, some of these changes are easier to implement than others, and the learning curve will vary. It is important not to view open-science practices as an all-or-none proposition; *perfect* should not be the enemy of *better*. Even incremental behavior change in individual research practices can be expected to improve replicability. Finally, individual researchers are not alone in these behavior changes. Journals and organizations are beginning to explicitly recognize these practices and encourage more direct replications. Through greater transparency and openness, psychology can expect to become a stronger science.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Mahwah, NJ: Lawrence Erlbaum Associates.

Baron, A. & Perone, M. (1999). Experimental design and analysis in the laboratory study of human operant behavior. In K. Lattal & M. Perone (Eds.), *The handbook of research methods in human operant behavior.* (pp. 45-91). Boston, MA: Springer US.

Branch, M. (2014). Malignant side-effects of null-hypothesis significance testing. *Theory and Psychology, 24,* 256-277. https://doi.org/10.1177/0959354314525282

Branch, M. (2018). The "Reproducibility Crisis:" Might the methods used frequently in Behavior-Analysis research help? *Perspectives on Behavior Science,* doi: https://doi.org/10.1007/s40614-018-0158-5

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., 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. http://dx.doi.org/10.1016/j.jesp.2013.10.005

Camerer. C. F., Derber., A., Holzmeister, F…. & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behavior. https://doi.org/10.1038/s41562-018-0399-z

Campbell, L. (2016). *Teaching good science by conducting close replications.* https://osf.io/5yrnm/.

Cohen, J. (1962). Statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153. http://dx.doi.org/10.1037/ h0045186.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd Ed., Hillsdale, NJ: Erlbaum.

Cook, (2018, March 9). Dear colleague letter: Achieving new insights through replicability and reproducibility. Retreived from https://www.nsf.gov/pubs/2018/nsf18053/nsf18053.jsp

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29. http://dx.doi.org/10.1177/0956797613504966.

DeMets, D. L., & Lan, G. (1995). The alpha spending function approach to interim data analysis. In Thall. P. F., *Recent Advances in Clinical Trial Design and Analysis* (pp. 1-27). Boston, MA: Kluwer Academic Publishers.

Davison, M. (1999). Statistical inference in behavior analysis: Having my cake and eating it? *Behavior Analyst, 22*, 99–103. http://dx.doi.org/10.1007/ BF03391986

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 1-17. doi: 10.3389/fpsyg.2014.00781

Eastwick, P. (2018, January 9). Two Lessons from a Registered Report. Retrieved from http://pauleastwick.blogspot.com/2018/01/two-lessons-from-registered-report.html

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological review, 70*, 193-242.

Eich, E. (2014). Business not as usual. *Psychological Science, 25*, 3–6. http://dx.doi.org/10. 1177/0956797613512465.

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology, 66*, 68-80. http://dx.doi.org/10.1016/j.jesp.2015.07.009

Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), Design and analysis of single-case research (pp. 245-277). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904. https://doi.org/10.1007/s11192-011-0494-7

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi:10.3758/BRM.41.4.1149

Festinger, L., Riecken, H., & Schachter, S. (1957). *When prophecy fails.* University of Minnesota Press.

Finkel, E. J., Eastwick, P. W. & Reis, H. T. (2017). Replicability and other features of high-quality science: Toward a balanced empirical approach. *Journal of Personality and Social Psychology, 133*, 244-253, doi: 10.1037/pspi0000075

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes, 54*, 137-154. doi:10.1016/S0376-6357(01)00155-3

Fiske, S. T. (2016). A call to change science's culture of shaming. *APS Observer*. Retrieved from https://www.psychologicalscience.org/observer/a-call-to-change-sciences-culture-of-shaming

Fox, A. E. (2018). The future is upon us. *Behavior Analysis: Research and Practice,* 18, 144-150, doi: 10.1037/bar0000106

Fraley, R. C., & Vazire, S. (2014). The N-pact Factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9*. Article ID e109019.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19*, 151–156. doi.org/10.3758/s13423-012-0227-9.

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

Franklin, R. D., Allison, D. B., & Gorman, B. S. (1996). Design and analysis of single-case research. Mahwah, NJ: Lawrence Erlbaum Associates.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review, 18*, 3–12. http://dx.doi.org/10.1177/1088868313507536.

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist, 102*, 460–465.

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

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some

arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*, 535-549. http://dx.doi.org/10.1111/spc3.12267

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*, 605-607. http://dx.doi.org/10.1177/1745691612459057

Grahe, J. E. (2014). Announcing Open Science badges and reaching for the sky. *The Journal of Social Psychology, 154*, 1-3. http://dx.doi.org/10.1080/00224545.2014.853582

Grahe, J. E. (2018). Another step towards scientific transparency: Requiring research materials for publication. *The Journal of Social Psychology, 158,* 1-6. doi:10.1080/00224545.2018.1416272

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

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83,* 314-320.

Hales, A. H. (2016). Does the conclusion follow from the evidence? Recommendations for improving research. *Journal of Experimental Social Psychology, 66*, 39-46. doi: 10.1016j.jesp.2015.09.011

Hales, A. H., McIntyre, M. M., Williams, K. D., & Thomas, H. (2018, March). *Ostracized and observed: People recover more from ostracism when it is witnessed by others*. To be presented at the Society for Personality and Social Psychology, Atlanta, GA.

Hawkins, R. X. D., et al. (2018). Improving the replicability of psychological science through

pedagogy. *Advances in Methods and Practices in Psychological Science, 1*, 7-18.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558. doi: 10.1002/sim.1186

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences, 33*, 61-135. doi:10.1017/S0140525X0999152X

Hennes, E. P., & Lane, S. P. (2017, April). *Power to the people: Simulation methods for conducting power analysis for any model*. Workshop presented at the 89th Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Huitema, B. E. & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*, 38-58. *https://doi.org/10.1177/00131640021970358*

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*, 995-1006. http://dx.doi.org/10.1037/0022-3514.79.6.995

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

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*, 411-414. Doi: 10.1037//1082-989X.5.4.411

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparamentric statistics. *Journal of Modern Applied*

Statistical Methods, 6, 81-90. DOI: 10.22237/jmasm/1177992480

Kidwell M.C., Lazarević L.B., Baranski E., Hardwicke T.E., Piechowski S., Falkenberg L-S., et al. (2016) Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biol 14*: e1002456. doi:10.1371/journal.pbio.1002456

Kirkpatrick, K., Marshall, A. T., Steele, C. C., & Peterson, J. R. (2018). Resurrecting the individual in behavioral analysis: Using mixed effects models to address nonsystematic discounting data. *Behavior Analysis: Research and Practice*, 18, 219-238. doi: 10.1037/bar0000103

Kawakami, K. (2015). Editorial. *Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes, 108*, 58-59, doi: 10.1037/pspi0000013

Kitayama, S. (2017). Editorial. *Journal of Personality and Social Psychology: Attitudes and Social Cognition, 3,* 357-360, doi: 10.1037/pspa0000077

Kyonka, E. G. E. (2018). Tutorial: Small-N power analysis. *Perspectives on Behavior Science, doi: https://doi.org/10.1007/s40614-018-0167-4*

Kochari, A., & Ostarek, M. (2018, January 17). Introducing a replication-first rule for PhD projects (BBS commentary on Zwaan et al, "Making replication mainstream"). Retrieved from psyarxiv.com/6yv45

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*, 701-710. http://dx.doi.org/10.1002/ejsp.2023

Lakens, D., & Etz, A. J. (2017). To true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*. Advance online publication. doi: 10.1177/1948550617693058

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*, 1827-1832. http://dx.doi.org/10.1177/0956797615616374

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. http://dx.doi.org/10.1177/1745691612460688

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9*, 343-351. http://dx.doi.org/10.1177/1745691614528215

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147–163. http://dx.doi.org/10.1037/1082-989X.9.2.147.

Mayer, H. (Producer/Director) & Norris, M. (Writer). (1970). *The social animal* [Motion picture]. (Available from Indiana University Audio-Visual Center, Bloomginton, Indiana).

McCabe, C. J., Kim, D. S., King, K. M. (2018) Improving present practices in the visual display of interactions. *Advances in Methods and Practices in Psychological Science, 1*, 147-165. https://doi.org/10.1177/2515245917746792

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research, 43*, 1048-1063. doi: 10.1093/jcr/ucw085

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices*

in Psychological Science, 1, 131 - 144, doi:o0i.1or1g7/170/.21157175/2541559214757914776457656

Meyvis, T., & van Osselaer, S. M. J. (2018). Increasing the power of your study by increasing your effect size. *Journal of Consumer Research, 44*, 1157 – 1173. doi: 10.1093/jcr/ucx110

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2017, August 24). *The Preregistration Revolution*. Retrieved from osf.io/2dxu5

Nosek, B. A., & Lakens, D. (2013). Special issue of *Social Psychology* on "Replications of important results in social psychology." *Social Psychology*, *44*, 59-60.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137-141.

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. http://dx.doi.org/10.1177/1745691612459058.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 346*(6251). doi:10.1126/science.aac4716.

Perone, M. (2018). How I learned to stop worrying and love replication failures. *Perspectives on Behavior Science, doi: https://doi.org/10.1007/s40614-018-0153-x*

Perrino, T., Howe, G., Sperling, A., Beardslee, W., Sandler, I., Shern, D., ... Brown, C. (2013). Advancing science through collaborative data sharing and synthesis. *Perspectives on Psychological Science, 8*, 433–444. http://dx.doi.org/10.1177/1745691613491579.

Poincaré, H. (1914). *Science and method*. London: T. Nelson.

Ramsey, P. H. (1990). "One-and-a-half-tailed" tests of significance. *Psychological Reports, 66*, 653-654.http://dx.doi.org/10.2466/PR0.66.2.653-654

Rosenthal, R. (1984). Applied Social Research Methods Series, Vol. 6. *Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science, 9*, 293-304. http://dx.doi.org/10.1177/1745691614528214

Sakaluk, J. K. (2016). Exploring Small, Confirming Big: An alternative system to The New Statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology, 66*, 47-54. http://dx.doi.org/10.1016/j.jesp.2015.09.013

Schachter, S. (1951). Deviation, rejection and communication. *Journal of Abnormal and Social Psychology*, 46, 190-207.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566. http://dx.doi.org/10. 1037/a0029487.

Sedlmeier, P., & Gigerenzer, G. (1992). Do studies of statistical power have an effect on the power of studies? In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (pp. 389-406). http://dx.doi.org/10.1037/10109-032

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

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered

replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9*, 552-555. http://dx.doi.org/10.1177/1745691614543974

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*, 1123-1128. http://dx.doi.org/10.1177/1745691617708630

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*, 1875–1888. http://dx.doi.org/10.1177/0956797613480366.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559-569. http://dx.doi.org/10.1177/0956797614567341

Spellman, B. A. (2015). A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science, 10*, 886 - 899. doi: doi: 10.1177/1745691615609918

Vasishth, S. & Gelman, A. (2018). The statistical significance filter leads to overconfident expectations of replicability. Retrieved April 27, 2018, from https://arxiv.org/abs/1702.00556

Vazire, S. (2016). Editorial. *Social Psychological and Personality Science, 7*, 3-7. http://dx.doi.org/10.1177/1948550615603955

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

Wagenmakers, E., & Dutil, G. (2016). Seven selfish reasons for preregistration. *Observer, 29,* 13 - 14.

Wang, Y. A., Sparks, J., Gonzales, J. E., Hess, Y. D., & Ledgerwood, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and Type I error inflation. *Journal of Experimental Social Psychology, 72*, 118-124.http://dx.doi.org/10.1016/j.jesp.2017.04.011

Wesselmann, E. D., Williams, K. D., Pryor, J. B., Eichler, F. A., Gill, D. M., & Hogue, J. D. (2014). Revisiting Schachter's research on rejection, deviance, and communication (1951). *Social Psychology*, *45*, 164-169.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral and Brain Sciences*. Advance online publication. http://dx.doi.org/10.1017/S0140525X17001972