ANNUAL REVIEWS

Annual Review of Psychology

Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis

Patrick E. Shrout[1] and Joseph L. Rodgers[2]

[1]Department of Psychology, New York University, New York, New York 10003; email: pat.shrout@nyu.edu

[2]Department of Psychology and Human Development, Peabody College, Vanderbilt University, Nashville, Tennessee 37205; email: joseph.l.rodgers@vanderbilt.edu

Annu. Rev. Psychol. 2018. 69:487–510

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-122216-011845

Keywords

statistics, methodology, replication

Abstract

Psychology advances knowledge by testing statistical hypotheses using empirical observations and data. The expectation is that most statistically significant findings can be replicated in new data and in new laboratories, but in practice many findings have replicated less often than expected, leading to claims of a replication crisis. We review recent methodological literature on questionable research practices, meta-analysis, and power analysis to explain the apparently high rates of failure to replicate. Psychologists can improve research practices to advance knowledge in ways that improve replicability. We recommend that researchers adopt open science conventions of preregistration and full disclosure and that replication efforts be based on multiple studies rather than on a single replication attempt. We call for more sophisticated power analyses, careful consideration of the various influences on effect sizes, and more complete disclosure of nonsignificant as well as statistically significant findings.

Contents

INTRODUCTION

In the past decade, articles in both the scientific and popular press have described a supposed crisis in science—particularly in psychological science. These articles claim that there are many findings that do not replicate in new studies and that researchers have adopted practices that will lead to false conclusions. In response to the crisis, scores of articles have been written that diagnose the problem, evaluate its impact, and suggest solutions. This burgeoning literature prompted us to write this article in the *Annual Review of Psychology*.

We review what we consider to be the most important points in the literature and make special efforts to include perspectives from clinical, cognitive, health, organizational, developmental, and social psychology, as well as neuroscience. We argue that the sense of crisis in the past decade has produced important insights and conventions, notably the emphasis on open science, whereby predictions, analysis plans, data, and supplemental material are made available to the broad scientific community. We also argue, however, that confusion about some statistical procedures and misinterpretation of important methodological issues have led some commentators to conclude that the state of our science is worse than it is. The events of the past decade may even signal natural and positive growth within psychological research and methods.[1]

Our chapter works through a series of nine questions: (*a*) Why do people say there is now a crisis? (*b*) How did scientific conventions for evaluating evidence evolve? (*c*) What are the specific problems with scientific practices in psychology? (*d*) What procedural steps have been taken to address these problems? (*e*) How can statistical theory help address the problems? (*f*) How can psychological theory help inform effect size variation? (*g*) How can replicable findings become more common in psychology? (*h*) Do new norms and procedures cause collateral damage to some scientists and disciplines? (*i*) What is the take-home message about how psychological science can speed knowledge construction?

OVERVIEW: WHY DO PEOPLE SAY THERE IS A CRISIS?

People who believe that there is a crisis point to three sets of events as evidence. The first involved scientific fraud by psychological scientists, including the highly publicized case of social psychologist Diederik Stapel (Bhattacharjee 2013), as well as less publicized cases such as those of cognitive psychologist Marc Hauser and social psychologist Lawrence Sanna (Wade 2010). The second, related set of events was the publication of articles by a series of authors (Ioannidis 2005, Kerr 1998, Simmons et al. 2011, Vul et al. 2009) criticizing questionable research practices (QRPs) that result in grossly inflated false positive error rates in the psychological literature. The third set of events involved the effort by the Open Science Collaboration to replicate 100 results that were systematically sampled from three top-tier journals in psychology: (*a*) Only 36% of the replication efforts yielded significant findings, (*b*) 32% of the original findings were no longer significant when combined with the new data, (*c*) effect sizes in the replication studies were about half the size of those in the original studies, and (*d*) failures to replicate were related to features of the original study (e.g., replication failures were more common in social than in cognitive psychological studies and in studies reporting surprising rather than intuitive findings). These events reverberated throughout the psychological and broader scientific communities, and special sessions on replication, QRPs, and open science at national meetings of major societies attracted standing-room-only crowds. In fact, concerns over research ethics, QRPs, reproducibility of reported analyses, and replication of empirical findings go far beyond the boundaries of the field of psychology (Dickersin & Rennie 2012, Ioannidis 2005), although our treatment in this article focuses on our own field.

These concerns have led to many articles by scientists and methodologists whose very attention to replication has reinforced the sense of crisis. Some authors were explicitly negative about the state of our science (Coyne 2016, Schmidt & Oh 2016), whereas others suggested new methods, norms, and explanations for the disturbing events just reviewed. Many of these articles implied that the null hypothesis statistical test (NHST) traditions of the twentieth century were susceptible to implicit and even explicit attempts to deceive reviewers, editors, and readers about the strength of the evidence presented in empirical articles. Given that current research may not be optimally advancing knowledge, journal editors published statements that set standards designed to improve the integrity of published findings.

Still other articles have taken a more sanguine view about science and the current sense of crisis (e.g., Maxwell et al. 2015, Stroebe 2016). These treatments have reinterpreted QRPs and

[1] Readers can find both overlapping and also relatively different perspectives on many of the issues we treat in the review in this volume by Nelson et al. (2018).

the replication crisis as unfortunate but relatively natural features of the scientific enterprise, ones that must be policed and sanctioned but that are hardly worth the disciplinary panic that has ensued. Methodologists within this perspective have formally analyzed apparently surprising replication results and concluded that they are neither surprising nor difficult to interpret.

ORIGINS: HOW DID SCIENTIFIC CONVENTIONS FOR EVALUATING EVIDENCE EVOLVE?

Beginning with the establishment of Wilhelm Wundt's psychology laboratory in 1879, knowledge construction in psychology has been associated with the scientific method. Basic textbooks in psychology are packed with findings based on empirical evidence, which is itself typically embedded in a scaffold of psychological theory that fills gaps in the evidence and suggests new avenues of research. Theories are typically developed inductively from empirical patterns but then provide deductive special cases that can be further evaluated to enhance knowledge construction.

In his seminal work *A System of Logic*, Mill [2008 (1843)] defined the philosophical basis for the true experiment as the method of differences. He posited that, when two groups were exactly equivalent in all respects except that one received a treatment, any posttreatment differences were logically emergent from the treatment. The problem for Mill and for subsequent laboratory and natural scientists was how to make groups exactly equal pretreatment. For 75 years following Mill's work, equating was achieved using systematic designs that mechanically equated on obvious confounding factors (Box 1978, pp. 144–45).

As an agricultural scientist, Fisher was concerned about natural variation that obscured inferences made by direct observations. His insight (e.g., Fisher 1925) was to use random assignment to exactly equate groups across replications so that the effect of a treatment could be explicitly studied. His logical expansion of Mill's method [combined with work by Neyman & Pearson (1928, 1933)] became NHST, which requires scientists to work within a falsification context: How likely are the data under a null hypothesis in which the scientists' theories are incorrect? If the data are unusual under the null hypothesis, we reject the null hypothesis, whereas if the data are not unusual, then the null hypothesis is not rejected.

Fisher proposed defining "unusual" as less than or equal to 1 in 20 ($\alpha = 0.05$), although he had no intention for $\alpha = 0.05$ to become an industry standard: "If one in twenty does not seem high enough odds, we may … draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point)" (Fisher 1925, p. 504). Setting α allows tuning for the necessary trade-off between Type I and Type II errors, though modern researchers seldom consider using this tuning potential. Obviously, if replicable results were the only research goal, setting α lower (say, $\alpha = 0.001$) would reduce Type I errors. The result would be fewer Type I errors but increased Type II errors and reduced power if the study design were held constant.

Fisher was a scientist studying agricultural yield—for example, attempting to optimize the effect of fertilizer and planting practices. He naturally used replication across individual plants, plots, elevation levels, and rainfall to study variation in crop yield. To be clear, Fisher's original concern was never with the question of whether the original finding replicates. Rather, he viewed replication as a logical mechanism to achieve the goal, originally stated by Mill, of equating two groups prior to treatment—the equating occurred probabilistically across replications.

We believe that Fisher would bring similar concerns to evaluating the current replication crisis. Likely, he would simply suggest that when results do not replicate, science is proceeding as it should in a self-correcting manner. Similarly, when results do replicate, continued caution and scrutiny are still appropriate. However, Fisher and his colleagues were concerned with logical and statistical issues rather than fraud, deceit, and carelessness.

CONCERNS: WHAT ARE THE SPECIFIC PROBLEMS WITH SCIENTIFIC PRACTICES IN PSYCHOLOGY?

Among the factors that diminish accumulation of knowledge, outright fraud is the most serious, but it also appears to be low in prevalence. We do not devote much of this review to the prevention of outright fraud, as there are legal and professional tools available to do this. However, we note that a colleague who inexplicably violates the trust of the scientific enterprise primes the fear that we must be suspicious of others. The overgeneralization of such suspicions can be pernicious.

Fraud aside, several authors have written about ways in which researchers systematically increase Type I error rates and overstate the magnitude of effects. Researchers have identified a set of QRPs that are likely to lead to scientific claims that are false or too strong. These practices are often encouraged in a motivational environment where hiring, tenure, promotion, and grants are heavily influenced by numbers of publications in top-tier journals (Nosek et al. 2012). These journals aim to publish important, novel, and theoretically motivated findings that have a compelling narrative.

Kerr (1998) described a pattern of formulating hypotheses after research results are known (HARKing), in which the research appears more confirmatory than it actually was. For example, suppose a study yields few statistically significant results, but one significant result was an unanticipated interaction. A skeptical view would consider the interaction to be likely due to sampling variation, and a correction for multiple analyses of the data might show that the result is not unusual. However, suppose that the author finds a literature that would have actually predicted the interaction and then writes the introduction to suggest that the study was designed to test this specific interaction. Even a skeptic might be convinced that the evidence was compelling. Kerr invites the skeptic to think again.

HARKing confuses exploratory and confirmatory studies and presents results from the former as though they were the latter. Exploratory studies are an important component of most research programs and can lead to discoveries of interesting associations, complicated multivariate structures, and rare but important occurrences (Ledgerwood et al. 2017). Statistical methods for data exploration are well known (Tukey 1977), and advances in machine learning and other exploratory data mining tools are extending the reach and power of exploration (e.g., McArdle & Ritschard 2014). However, insights from exploration should be confirmed in new data: "Exploratory data analysis is detective in character. Confirmatory data analysis is judicial or quasi-judicial in character" (Tukey 1977, p. 3). NHST is designed to control false positive rates for confirmatory studies under the assumption that hypotheses and analytic procedures are defined ahead of time. HARKed results typically violate the assumption of full disclosure.[2]

Vul et al. (2009) described a pattern of reporting functional magnetic resonance imaging (fMRI) results that also mixed exploratory analysis and confirmatory claims. They noted that many fMRI studies reported extremely high correlations (>0.8) between brain activation and individual differences and speculated that at least some of the high correlations were due to initial exploration and identification of brain regions of interest that were subsequently treated as if they were known. Vul et al. (2009) showed that, if the same data were used to find promising associations and then to test those associations, biased correlation estimates and inflated Type I error rates would result.

Simmons et al. (2011) described ways in which an investigator motivated to obtain publishable findings might report a result with a higher Type I error rate than the nominal test statistic

[2] Investigators who make an unexpected finding as they explore data and then disclose that many tests were used to uncover the finding can use the modern multiple comparison methods of Benjamini & Hochberg (1995) to reduce the likelihood of false discovery.

suggests. The key QRPs they highlighted were (*a*) using more than one dependent variable that reflects the outcome of interest, (*b*) carrying out interim data analysis during the data collection and stopping the study when a desired finding becomes significant, (*c*) carrying out multiple analyses with different covariates as ancillary variables, and (*d*) dropping groups or levels to focus on a larger effect in a subset of the data. These QRPs can lead to grossly inflated Type I error rates, even without HARKing unexpected findings.

In addition to QRPs that introduce uncontrolled increases in Type I error or inflation of estimated effect sizes, statistical errors can lead to additional bias. Bakker & Wicherts (2011) checked 281 articles and found statistical errors in 18% of them. Although these errors might have simply added white noise to the literature, the authors found that the vast majority made the results more apparently significant and thereby more reportable in journal articles.

False positive results are not the only kinds of errors that are important to address in psychology. One historically important QRP is to carry out studies with inadequate statistical power to test interesting effects. For example, Rossi (1990) reported that, in a survey of published studies, average power to detect what Cohen (1988) called a medium effect (e.g., $d = 0.50$) was 57%. There is no evidence that this high Type II error rate has improved in the 25 years since the Rossi report. The problem of high false negative rates is compounded by the habit of some recent researchers to conclude that an effect is absent if a test is nonsignificant. Confidence bounds around parameter estimates are infrequently used to characterize nonsignificant findings as either informative about a small or absent effect or indicative of an inconclusive study design (Cumming 2014).

PRACTICAL RESPONSES: WHAT PROCEDURAL STEPS HAVE BEEN TAKEN TO ADDRESS THESE PROBLEMS?

The reported low replication rate in the Open Science Collaboration studies, as well as extensive discussions of QRPs, has led to constructive suggestions for improving scientific procedures in psychology. One useful thread of discussion concerns the definition and meaning of replication. A direct or exact replication is a new study that employs the same procedure, materials, measures, and study population as the original study. The sample size need not be the same in an exact replication; indeed, a power analysis will usually suggest a larger sample (Brandt et al. 2014). A systematic replication is a direct replication in which some ancillary features, such as the order of the presentation of stimuli, are different from the original. A conceptual replication is intentionally different from a direct replication and is designed to assess generalizability, as well as veracity, of a result. It may involve a similar but not identical intervention, alternate measures of the outcome, or samples from a distinctly different population or era (Fabrigar & Wegener 2016, Ledgerwood et al. 2017). The distinctions between exact, systematic, and conceptual replications represent a continuum.

All types of replication are informative—the question is which gets priority. When a conceptually replicated effect is statistically significant, it shows that the phenomenon is more general (Stroebe 2016). However, if the effect does not replicate, the investigator does not know if the original finding is suspect or if it does not extend to other measures, procedures, or contexts (Fabrigar & Wegener 2016).

Other commentaries have suggested that we rethink the twentieth-century assumption that individual scientists can be trusted to report the relevant results and theorizing that led to the studies. Instead of trust, the theme of the new norms is openness. In 2013, Brian Nosek led a group that established the Center for Open Science and the Open Science Framework (OSF), a nonprofit organization that facilitates the sharing of study plans, materials, and documents. The OSF promotes scientific quality by facilitating a sequence of steps: (*a*) Preregister confirmatory hypotheses before data are collected, including outcomes and analyses; (*b*) preregister study design, including procedures, measures, and planned sample size; (*c*) make data and analysis syntax available so that others can reproduce the results. The OSF provides a pragmatic infrastructure for implementing the suggested norms (Nosek et al. 2015), building on a tradition of preregistration in health studies (De Angelis et al. 2005).

Preregistering hypotheses makes it possible to rule out HARKing and also guards against many of the QRPs identified by Simmons et al. (2011) and others (Nosek & Lakens 2014, Wagenmakers et al. 2012). Requiring explanation of the sample size rationale both encourages formal power analyses and guards against investigators stopping the study when desired results are obtained. Providing the original data used to make scientific claims guards against incorrect inferences that are based on misspecified statistical analyses. In our opinion, the preregistration movement and the OSF support for it have been the most important outcomes of the replication crisis.

A number of other contributions to the replication literature have emphasized traditional approaches to strong science. Some have recommended increased sample sizes (Button et al. 2013, Francis 2012a, Maxwell et al. 2015, Perugini 2014), and some have even suggested specific target sample sizes of approximately 90 per group (Vazire 2016). Others have advocated for increased attention to measurement error (Asendorpf et al. 2013) and to the quality of peer review (Coyne 2016, Nosek & Bar-Anan 2012) and for flexibility regarding the traditional significance level of $p < 0.05$ (Vazire 2016). Among the most controversial suggestions is that a cadre of researchers should act to identify other researchers whose results seem to replicate with lower frequency than the norm (Francis 2012c, Simonsohn 2013).

Editors of various journals have been processing these recommendations and have issued statements about accountability and quality control mechanisms For example, *Psychological Science* instituted badges to mark articles that implement the new norms of preregistration, open access to materials, and open access to data (Kidwell et al. 2016). We discuss these editorial steps further in the final section of this article.

STATISTICAL RESPONSES: HOW CAN STATISTICAL THEORY HELP ADDRESS THE PROBLEMS?

While editors and leaders pondered policies to improve the quality of psychological research, quantitative psychologists and statisticians studied the formal logic and mathematics of the replication process. In this section, we review four aspects of this work: (*a*) the use of refined power analysis for replication studies, (*b*) the role of meta-analysis in understanding replication variation, (*c*) the promise of Bayesian analysis for understanding replication variation, and (*d*) the use of resampling methods.

Planning Power for a Replication Study

When the Open Science Collaboration planned the replication studies of 100 sampled findings, they reported that the new studies had 0.92 power on average. These calculations were undoubtedly based on the effect sizes of the original reports. The sample sizes of the replication studies were typically larger than in the original report, with a median sample size more than 25% larger than the original (Open Sci. Collab. 2015). Even with larger samples, those replication studies were likely underpowered. McShane & Böckenholt (2014) noted that any meta-analysis reveals effect size heterogeneity; different studies report different effect sizes. This heterogeneity may be due to changes in setting, procedures, and subject characteristics. When the variability is taken into account in the power analysis, the replication sample needs to be larger, often considerably so.

Maxwell et al. (2015) came to the same conclusion but added concerns about statistical bias and estimation variation in the effect estimate. The typical power analysis for a close replication is a conditional power analysis, which treats the effect size as known and fixed. Maxwell et al. (2015) argued that the correct analysis takes into account the uncertainty of the effect size distribution. Describing the possible range of effect sizes is challenging because of publication bias (Francis 2012b) as well as the sampling variation of any estimate of an effect size. To illustrate publication bias, suppose that a study is theoretically justified, the hypothesis is preregistered, and the analysis is planned and properly executed, but that statistical power is only 50%. A lucky investigator would obtain a sample effect size that is randomly larger than the population value and would therefore publish, whereas an unlucky investigator would obtain an effect size that is smaller and nonsignificant due to sampling fluctuations alone. The latter result would historically not be published, and so the published effects would be biased relative to the sampling distribution. Publication bias accounts for at least part of the observation that effect sizes in the Open Science Collaboration replication studies were about half the size of the original studies (Open Sci. Collab. 2015).

In addition to publication bias, there is sampling variation in the reported effect size. This variation is larger as the sample size of the original study gets smaller. What is often not appreciated is that the impact on power of a small effect size change is not linear. **Figure 1** shows the sample size that is needed to obtain 95%, 90%, and 80% power for a range of effect sizes, measured as Cohen's *d*. This figure shows that, if one reduces an effect size from 0.8 to 0.7, then the required total sample size for 95% power increases by 26 (from 84 to 110); if the effect size is reduced from 0.4 to 0.3, then the sample size increases by 252 (from 328 to 580). The change in required *n* is even more dramatic if the effect size is reduced from 0.3 to 0.2. Taking the midpoint of a

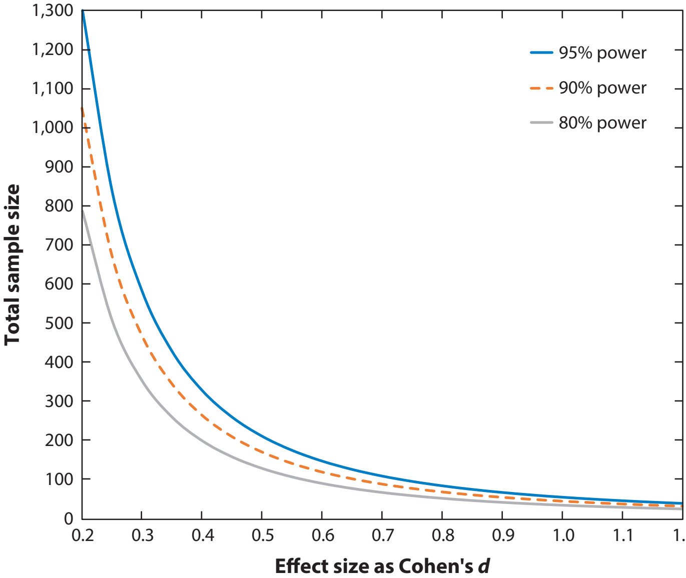

Figure 1

Power analysis showing the total sample size needed to achieve 95%, 90%, and 80% power for a two-group t test as a function of assumed effect size in Cohen's *d* metric. Power is calculated assuming equal sample sizes in both groups and a two-tailed test with $\alpha = 0.05$. Note that small differences in effect sizes in the $d = 0.2$–0.6 range have large effects on required sample size.

range of possible effect sizes is likely to lead in the long run to an underpowered study because of this nonlinear pattern. A correction for sampling variation involves computing power from the lower bound of a confidence interval on the effect size, as Maxwell et al. (2015) demonstrated. A number of statisticians have studied, from a frequentist perspective, the problem of describing the distribution of the unknown effect size to obtain an optimal predictive power estimate (Anderson & Maxwell 2017, McShane & Böckenholt 2014, Perugini et al. 2014, Taylor & Muller 1996). A promising alternate approach is to take a Bayesian perspective, which we discuss below.

Even if an appropriate power analysis is carried out and the replication study is properly powered, Maxwell et al. (2015) pointed out that it is often not possible to strongly infer that there is no effect. The replication study can produce results that are consistent with some effect size of interest, as well as results consistent with absolutely no effect. To make a strong claim that there is no meaningful effect from a frequentist perspective, the replication study must be designed as an equivalence study (Rogers et al. 1993, Seaman & Serlin 1998). Equivalence studies are used to show that a generic drug has the same effectiveness as the original patented drug. They require specifying an effect interval around the null value that is considered to be virtually (if not exactly) null and then showing that the new data are only consistent with effect sizes in the null-equivalent interval. To make a strong equivalence statement, the sample size typically needs to be very large. Maxwell et al. (2015) provided an illustration where the required sample sizes were between 1,000 and 10,000 per group, depending on how sure the scientist wants to be that no small effect is present. Given those requirements, Maxwell et al. (2015) recommended that psychologists cease their focus on single replications and instead use multiple studies to obtain a range of new effect sizes. As multiple studies are completed and reported, the focus can shift to estimating effect size distributions, rather than binary decisions.

Maxwell et al. (2015) reinterpreted the Open Science Collaboration's finding that only 36 of 100 replication efforts were significant as fully consistent with methodological expectations. Although a successful replication rate of 36% was surprising to many, Maxwell et al. (2015) demonstrated that the finding does not signal a crisis. Instead, we need to focus on understanding distributions of effect sizes rather than on trying to obtain closure on whether an effect is true or false from a single study. The shift in focus from testing of hypotheses to estimation of effects makes meta-analysis and Bayesian methods particularly useful, as we discuss next.

Meta-Analysis of Multiple Replications

Meta-analytic techniques provide key tools for analyzing and interpreting distributions of effects from replication studies. From a meta-analytic perspective, finding that two studies have different outcomes is not a problem. Rather, it is a common result as evidence accumulates (Schmidt & Oh 2016). Many authors addressing statistical issues related to replication have noted that meta-analysis is an important analytic tool (e.g., Braver et al. 2014, Cumming 2014, Fabrigar & Wegener 2016, Maxwell et al. 2015, McShane et al. 2016).

Meta-analysis allows not only estimation of the average and variation of effect sizes but also substantive exploration of factors that moderate the size of the effects. In the context of meta-analysis, the original study now has status weighted according to its sample size (as do the other studies), and the unnatural focus on the single original study is avoided. Treating study results from multiple studies as random effects, generating effect size distributions rather than assumed fixed values, has often been advised (see, for example, Borenstein et al. 2009). However, an important caveat is that traditional random effect models require more than a few replications. Many statisticians would recommend 20 or more studies to obtain stable variation estimates, but Bayesian methods, described below, may be used with a smaller number when an appropriate prior distribution can be specified.

An important future direction for meta-analysis is the consideration of the population of situations and studies that are characterized by the average effect size estimates. McShane et al. (2016) considered how selection effects can lead to misleading conclusions about overall effects. Bonett (2009) considered an alternative to a random effect model that does not require envisioning a meaningful superpopulation of effects. Replication projects (Simons et al. 2014, reviewed in detail below) provide a meaningful framework for considering effect estimates obtained across independent labs to be samples from a population of replications; they also avoid the file drawer problem, whereby nonsignificant findings are not reported publically. Special efforts must be taken to assure that meta-analysis is not carelessly used to summarize findings that are biased by p-hacking (Braver et al. 2014). The application of sophisticated meta-analytic methods to data from coordinated preregistered studies of specific effects has great promise for knowledge production in psychology.

In addition to meta-analysis of studies, there is a literature on meta-analysis of individual data (Stewart & Parmar 1993), also called integrative data analysis (Curran & Hussong 2009). This approach is worth considering when individual data from studies are available and when there are few replications that can be combined with the original study. If the studies can be considered comparable, then the combined sample will have more power than the original study, and the precision of the combined estimates will increase. This approach also allows flexibility in examining the impact of different statistical models and adjustments. If individual data are available for a large number of studies, then the multilevel model can include random effects for the treatment.

The Promise of Bayesian Analysis

Although most modern treatments of psychological statistics are frequentist rather than Bayesian, the original methods in statistics were firmly Bayesian; the publication of Bayes' Theorem (posthumously) in 1761 and subsequent contributions from Laplace are often identified as the beginning of the field of statistics (see Fisher 1958, Stigler 1986). It was more than 150 years later that Fisher, Neyman, and Pearson developed the components of the frequentist school of statistics, which, once blended, became NHST. Despite early skepticism, their methods were ultimately so influential that Bayesian thinking virtually disappeared for several decades, until a revival began in the 1950s (see Zabell 1989). Fisher was dismissive of the Bayesian perspective early in his career, but became ever more appreciative of inverse probability theory (Aldrich 2008).

Inverse probability refers to evaluating the probability of a hypothesis given the data. In contrast, frequentist NHST procedures evaluate the probability of the data given the null hypothesis. Even frequentists believe the Bayesian to be the more natural scientific formulation (e.g., Cohen 1994, Maxwell et al. 2015) and to be ideally suited to addressing issues of replication. At a conceptual level, a Bayesian blends prior knowledge (a prior distribution) with the data to arrive at revised knowledge (a posterior distribution). Specifically, a Bayesian takes past results, combines them with one or more efforts to replicate those results, and arrives at an updated opinion. The explicit incorporation of emerging data into scientific assessment and interpretation of phenomena is one reason that the Bayesian approach is both useful and appealing (Kruschke & Liddell 2017).

Another advantage of the Bayesian approach for replication is the way that it directly deals with the plausibility of the null hypothesis. Gallistel (2009, p. 439) noted, "Conventional statistical analysis cannot support [the null hypothesis]; Bayesian analysis can." In the frequentist school, the null is either rejected or not rejected. Bayesian methods support a more formal approach to direct evaluation of the null hypothesis. Wagenmakers et al. (2016) provided a clear explanation of how the Bayes factor—the ratio of the probability of the null hypothesis to the probability of a competing hypothesis—can quantify evidence for the absence of an effect relative to a hypothesized effect. When the Bayes factor is greater than 1.0, the null hypothesis is more likely, given the data, than the alternative. When the Bayes factor is less than 1.0, the alternative is more likely.

Bayesian statisticians have developed new methodology to study replication efforts. Verhagen & Wagenmakers (2014) developed a null and alternative hypothesis framework explicitly directed to test whether the effect in the new data is similar to what was found in the initial study or whether the new effect is absent. Of additional note is that Verhagen & Wagenmakers (2014) used a bootstrap to obtain a Bayesian posterior distribution, and they also compared their method to the partial Bayes factor, based on a cross-validation methodology in a Bayesian context.

The promise of Bayesian analysis is especially appreciated by researchers who are interested in estimation, such as interval estimation and meta-analysis (Kruschke & Liddell 2017, Scheibehenne et al. 2016). Bayesian estimation of the posterior distribution of parameter values given new data is informative and can be estimated using numerical methods that are becoming widely available (see, e.g., Kruschke 2015). The posterior distribution can be graphed, and the likelihood of the different values (given the data) can be assessed. The interval where the values are most likely is called the highest density interval or the credible interval. Although it is sometimes compared to frequentist confidence intervals, this interval contains richer information about the likelihood of various values than confidence intervals. When Bayesian methods are used in meta-analysis, both between-study and within-study variations contribute to the overall posterior distribution. Kruschke & Liddell (2017) demonstrated how meta-analysis forest plots can be modified to show more information about the posterior distributions within and between studies.

Resampling Methods

Resampling methods (e.g., Beasley & Rodgers 2009, Efron 1979) and cross-validation (for a detailed treatment, see Efron & Gong 1983) are useful statistical tools to bring to the study of replication results. Both methods are based on a replication perspective. Resampling methods test hypotheses using the data to build empirical sampling distributions, rather than theoretical sampling distributions. Cross-validation approaches use multiple samples within a single study; the first sample (the estimation or training sample) is used to identify models in an approximately exploratory context, whereas the second sample (often the calibration sample) is used to further investigate hypotheses in a confirmatory sense. However, resampling methods usually consider the research context where the data were collected to be fixed, and, thus, that source of variation is ignored by these methods.

BLENDING PSYCHOLOGY WITH STATISTICS: HOW CAN PSYCHOLOGICAL THEORY HELP INFORM EFFECT SIZE VARIATION?

The statistical responses to the replication crisis emphasize the fact that effect size values needed for power analyses come from distributions of effect sizes. Meta-analyses can show how much variation there is in these distributions but cannot necessarily distinguish between the sources of variation—effect heterogeneity (McShane & Böckenholt 2014) or estimation variability (Maxwell et al. 2015). Even if meta-analyses are not available to inform a researcher about how much variation to expect, it is important to recognize that variation is likely to exist. One way to think about this variation is to use psychological theory to anticipate factors that might moderate the effect of central interest.

Modeling Effect Variation

To illustrate how one can conceive of effect variation, consider a loud sound right behind you, such as a firecracker or a gunshot. It would be a rare person who would not jump, and we speculate that the effect of this sonic intervention would be in the order of $d = 2$ or larger. However, it is possible to think of circumstances where the effect would be attenuated or enhanced. We consider four classes of such moderating factors: (*a*) the strength of the intervention, (*b*) the choice of outcome, (*c*) characteristics of the participants, and (*d*) the setting and context of the study (see also Asendorpf et al. 2013).

Intervention strength. The intervention can vary in intensity and purity. In the case of the firecracker, its size and location would be associated with the intensity of the sound. In some instances, a single intervention ends up having a distribution of effect sizes because replications of stimuli involve different words, phrases, pictures, or confederates, and each instance of the stimulus has a possibly different intervention effect. When the stimuli or delivery mechanisms of interventions are sampled, they must be considered random effects in the analysis (Westfall et al. 2015).

Outcome. Interventions can have impacts on a range of outcomes. For example, the loud popping sound would affect skin response, blinking response, and the neural networks that process the sensation. An intervention for depression might reduce symptoms as reflected in either brief psychological screening measures or extensive psychiatric diagnostic protocols (Tackett et al. 2017). Often, investigators choose an outcome that is easily or affordably assessed even if it might be less affected by the intervention.

Participant characteristics. The characteristics of the participants can have important moderating effects on interventions. Someone with auditory impairment would not react to a gunshot as much as someone with normal hearing. A veteran with posttraumatic stress disorder might have an enhanced reaction. Moderation of effects by participant characteristics is common. Relationship researchers, for example, can show that daily partner criticism has a larger effect on people with attachment anxiety relative to persons who are securely attached (Overall et al. 2014). Psychopathology research has studied so-called personalized medicine, where a patient's genetic profile is often related to responsiveness to a pharmacologic intervention (Ozomaro et al. 2013).

Study setting and context. Studies may have effects that vary because of features of the laboratory settings, such as sound isolation, lighting, color scheme, and temperature. An important setting feature is whether the experiment is carried out in a controlled lab or in an uncontrolled home setting using online technology. Another feature of the setting is the expertise of the research group; a replication study carried out by an experienced research team might obtain a different result than one by an undergraduate class (Bench et al. 2017). Related to setting are the broader social context and whether that context is likely to moderate the outcome. Van Bavel et al. (2016) showed that contextual sensitivity was an important predictor of whether the replication result matched the original result.

Equation 1 represents a formal specification of the effect size in a specific study, which we represent as Δ, with subscripts for intervention type (i), outcome (k), participant type (p), and setting factors (s), which can modify the effect in a specific circumstance:

$$\Delta_{\mathrm{ikps}} = \mu + \xi_{\mathrm{i}} + \theta_{\mathrm{k}} + \pi_{\mathrm{p}} + \eta_{\mathrm{s}} + \xi\theta_{\mathrm{ik}} + \xi\pi_{\mathrm{ip}} + +\xi\eta_{\mathrm{is}} + \cdots + \pi\eta_{\mathrm{ps}} + \varepsilon_{\mathrm{ikps}}. \qquad 1.$$

This equation represents effect variation as main effects of the type of modifier and as two-way interactions of those effects, but one can model an even longer list of higher interaction terms. Equation 1 is inspired by generalizability theory (Cronbach et al. 1972), which describes various sources of variance of measurements in an extension of classical test theory. Consistent with this approach, we assume a universe of persons, settings, and constructs to describe an average effect size, and μ describes that average (mean). We define the other terms as random effects that can increase or decrease the size of the intervention effect as a function of intervention type (ξ_{i}), outcome specification (θ_{k}), participant type (π_{p}), setting or context (η_{s}), or their interactions. All the distributions of the random effects are assumed to be centered at zero. We also assume that all factors are independent, insofar as each can be specified separately in an experimental design. Finally, we add a residual term ($\varepsilon_{\mathrm{ikps}}$) to represent a portion of a study effect that is real but transient in time, sample, and context. If researchers design a study without attending to these effect modifiers, then the expected variance of Δ_{ikps} is the sum of the variances of all the effect modifier factors. The expected value of the effect in a specific study is μ plus the specific values of ξ, θ, π, and η, which correspond to the intervention, outcome, sample, and setting choices, respectively.

Equation 1 represents effect heterogeneity rather than sampling variation in the estimated effect. It can be used to represent the effect variation that occurs when a new study design moves from an exact or close replication to a conceptual replication. In an exact replication, investigators attempt to restrict all the moderating factors to be the same. In a conceptual replication, investigators might vary the implementation of the intervention, the primary outcome measure, the characteristics of the participants, and the setting where the study takes place (Brandt et al. 2014, Fabrigar & Wegener 2016). If care is not taken to consider the possible moderating factors, then the study might be implemented with an essentially arbitrary design regarding effect moderators.

Predicting Where the Effect Is Greatest

Rather than studying an intervention in an arbitrarily chosen but convenient design, consider a design where the investigator carefully specifies a combination of (*a*) intervention, (*b*) outcome, (*c*) participant type, and (*d*) study setting or context and where there is a theoretically based sweet spot for the effect. The ideal occurs when the effect is as large as our hypothetical firecracker effect. As a number of authors have argued in recent years (Edwards & Berry 2010, Fiedler 2017, Stroebe 2016), psychological theories can provide fertile ground for making predictions about mechanisms and contexts where large effects will occur. Theory can become proactive, in that it can be used to develop more nuanced theoretical propositions and more effective interventions.

However, we are not proposing that researchers find a constrained set of conditions where an effect is large and then make unqualified claims about the generality of the effect (Meiser 2011). Instead, we recommend that the effect space represented by Equation 1 be used to create a prototype effect and that generalizations be tested by explicitly changing the sources of effect variation in a series of conceptual replications. The program of research that examines boundaries and effect moderators will require larger sample sizes when the effect is diminished by a moderator or diluted by variation in subject or intervention characteristics. Considering the factors that can make intervention effects larger or smaller is particularly important when scientific findings are used as a basis for applications to clinical, health, or organizational psychology. In prevention science, for example, universal interventions such as education or training programs are administered to a large number of people with the knowledge that the average effect is small but that the exposure makes the intervention worthwhile. In contrast, targeted interventions might be limited to persons with risk factors such as obesity or smoking history. In that limited group, the average effect size might be quite large, justifying the focus on the special group. In the latter case, it might not be necessary for the initial researcher to show that the specific intervention is applicable to a broad population (see, for example, Offord et al. 1998).

When researchers think vaguely about the effect size and ignore features of their design that are likely to dilute the effect, their study can regress to the average effect, such as the value of $d = 0.41$ obtained from meta-analyses of all studies in a field (Richard et al. 2003). If a researcher plans a generic study design for a generic outcome, as in this case, the required sample size to obtain appropriate power may be 155 per group (LeBel et al. 2017). Such sample sizes make controlled laboratory studies difficult and may push investigators to use online testing with participants of variable quality (Paolacci & Chandler 2014). The shift from a setting that provides precision to one that leads to more noise might lead to effects that are even smaller than the average (for related points, see Finkel et al. 2017).

When effect heterogeneity is considered explicitly, it is possible to form a principled answer to the often-asked question: Is it better to carry out one study with $N = 400$ or ten studies, each with $N = 40$? If effect heterogeneity is considered likely, then many smaller studies done at different times and in collaboration with other labs will be more informative about the heterogeneity than a single large study, although the smaller studies will individually be less precise.

As noted above, Maxwell et al. (2015) explained that the relationship between effect size and power is nonlinear (see **Figure 1**). If subfields of psychology begin to accumulate information about effect sizes, as represented by Equation 1, researchers can specify a design where effects can be reliably observed. One problem with designing a study using the sweet spot and then finding a significant result in a relatively small sample is that skeptics will complain that the result is either a fluke or a product of HARKing or other QRPs. To address this skeptical concern, the researcher who uses principled theory to predict an effect (large or otherwise) should preregister that theoretical thinking.

BROADER SOLUTIONS: HOW CAN REPLICATIONS BECOME MORE COMMON IN PSYCHOLOGY?

More than forty years ago, Greenwald (1976) laid out his policies as incoming editor of the *Journal of Personality and Social Psychology* (JPSP): "There may be a crisis in personality and social psychology, associated with the difficulty often experienced by researchers in attempting to replicate published work" (Greenwald 1976, p. 2). Obviously, references to the urgency of the recent crisis are overstated. Greenwald recommended that authors provide evidence that a finding can be reproduced when exactly the same procedures are applied and also when a conceptually related hypothesis is tested. His suggestions led to the multiple study tradition in JPSP, designed to lead naturally to replication and the reduction in published Type I errors. As prescient as Greenwald was, the field did not yet understand that replication studies should be reported regardless of whether the new result was statistically significant. Often, there was a naive expectation that all of the reported studies should have results that were statistically significant.

Several authors noted that obtaining a series of significant findings when the underlying effect size is medium or small and power is limited is highly unlikely (Francis 2012c, Schimmack 2012, Simonsohn 2013). Although some attribute these unlikely events to QRPs on the part of the authors, the explanation cannot usually be known with any certainty. Reviewers and journal editors in the past might have suggested that nonsignificant results or studies be removed from the report to clarify the story. Whatever the reason, a pattern of four or five significant results related to an effect that is not very large in studies with small to medium power does not build confidence that the result will replicate in other laboratories. As Schimmack (2012) noted, replication results that are too consistently significant may lead, ironically, to disbelief.

In recent years, replication studies carried out in independent laboratories have been emphasized. One noteworthy model of a program to promote close replications is the Registered Replication Report (RRR) mechanism initiated by the Association for Psychological Science (Simons et al. 2014). These reports are the culmination of collaborative research carried out by multiple independent research teams to examine a specific research finding through multiple replication studies. The collaborative project is proposed and approved before the replication studies are carried out. The journal commits to publish a well-prepared RRR report regardless of findings. Once the RRR is accepted, the lead scientist develops a protocol and invites participation from a number of independent research groups. This protocol is sent to the original scientist for comment and review. The goal is to make the replication process constructive rather than antagonistic. When the multiple studies have been completed, a meta-analysis of the results is prepared that allows readers to form a conclusion about the size of the hypothesized effect, as well as the degree of effect heterogeneity.

At the time of this writing, five RRR reports have been published. We review two of these. One team (Alogna et al. 2014) replicated a study by Schooler & Engstler-Schooler (1990) that documented a so-called verbal overshadowing effect on visual memories. The original study found that, when subjects provided a verbal description of a target person seen on a video, they were 25% less likely to identify that target person in a subsequent identification test than subjects in a control condition. In the RRR project, two waves of studies were done. The first apparently misrepresented the original procedure, and so we focus on the second (RRR2). The RRR2 protocol was specific, requiring that subjects be White undergraduates and that they come to a lab room to participate. There were 22 research groups from 8 countries participating in RRR2, and sample sizes ranged from 33 to 83, with a median of 52. The aggregated results were analyzed using a random effect meta-analysis. Although only 8 out of the 22 studies (36%) found significant results as individual studies, the aggregated effect (0.16) was statistically significant (95% CI: 0.12, 0.20), and 100% of the results were in the predicted direction. The report concluded that the RRR2 studies "provide clear evidence for verbal overshadowing" (Alogna et al. 2014, p. 571). The estimated effect size was reduced by the RRR2 effort, and the precision of the estimate was much improved.

None of the other four RRR projects provided clear support for the original finding. For example, an RRR report by Hagger & Chatzisarantis (2016) examined ego depletion effects using a protocol carried out on computers by Sripada et al. (2014). Ego depletion occurs when finite self-control resources are used up by initial demands and subsequent self-control is hindered by the depletion of those limited resources. The RRR project involved 23 laboratories in 11 countries and a total of 2,141 respondents. Only 2 of the 23 studies were significant in the predicted direction, and one study yielded a significant result in the opposite direction. Based on a random effect meta-analysis, the average effect size was 0.04 (95% CI: −0.07, 0.15). Not only did the interval include zero, the upper bound was far from the effect size of the original study. The authors of the RRR concluded, "Results from the current multilab registered replication of the ego-depletion effect provide evidence that, if there is any effect, it is close to zero" (Hagger & Chatzisarantis 2016, p. 558).

The RRRs provide a model of how to generate close replications across labs. We believe that findings will be more likely to hold in such rigorous tests if they are first replicated in the original lab. Perhaps the best way to provide an exact replication is for the original investigators to build the replication into the original report (as in cross-validation). For example, if investigators have to make analytic decisions, such as refining measures or determining the most appropriate covariates, before carrying out a confirmatory analysis, they could randomly split the sample into a training sample and a confirmatory sample. The second sample would be set aside until the analytic decisions were made and then used to provide an exact replication of the initial analysis. The effect sizes in the second analysis will typically be smaller than in the first, and this provides important information about effect sizes for later research.

Neuroscientists often provide exact replications of their findings, either with new experiments on the same subjects (Overath et al. 2015) or with replications with new participants (Ding et al. 2016). Moreover, within-subject results are often shown in neuroscience journals at the individual level and support evaluation of whether an effect is found in literally every subject (Ding et al. 2016). For example, Overath et al. (2015) reported fMRI measurements of specific regions of the superior temporal sulcus (STS) that were more responsive to auditory stimuli composed of packages (quilts) of 960-ms segments of human speech than to packages of 30-ms segments. The longer segments contained sound that could be identified as speech, whereas the shorter segments were too short to be identified. No similar association to segment variation was found in early auditory cortex responsiveness. Findings were replicated on different days for up to four repeat fMRI sessions for the 15 subjects. These sessions included different control conditions, but the condition on STS variation with segment length was repeated each time. Although the authors did not report effect size in standardized measures, their figures show that the effect was very large ($d > 2.5$ to compare the 30-ms and 960-ms conditions). Effects of this size are apparent in graphical representations, as well as through formal statistical analysis. The participants are not sampled from a known population, and their representativeness cannot be assured, but the effect was found for all the participants, suggesting that it would likely replicate in other laboratories.

Ledgerwood & Sherman (2012) encouraged researchers in all areas of psychology to avoid making claims based on a single study. They reminded researchers that careful documentation of a scientific finding is akin to the strategy of the tortoise rather than that of the hare in the race to produce scientific knowledge. If an interesting result has a medium, rather than a large, effect size, it will take many more subjects and additional studies to establish that result. When the researcher finds that the interesting finding goes away with repeated replications, there may be no publication to reward the researcher for the effort. Although the failure to establish a predicted effect may slow the career of the individual scientist, the suppression of reports that turn out to be Type I errors may speed up the accumulation of knowledge in the field as a whole. Meanwhile, the strategy, discussed in the previous section, of identifying the set of circumstances where an intervention can produce a large effect and can be documented with a relatively small sample may help mark progress even for the individual scientist.

A novel strategy for testing researchers' favorite hypotheses before publication was reported by Schweinsberg et al. (2016). Called the Pipeline Project, this group reported results from a prepublication independent replication (PPIR) effort to test 10 hypotheses that were being examined (in the pipeline) by the Uhlmann lab. These focused on person-centered morality (six hypotheses), morality and markets (two hypotheses), and reputation management (two hypotheses). Rather than the Uhlmann lab publishing their findings and then inviting replications by others, the PPIR approach established the effects before initial publication with the help of 25 independent labs, which collected data on more than 5 of the 10 target hypotheses. The results were reported as a meta-analysis of the replication studies using both frequentist and Bayesian methods (Verhagen & Wagenmakers 2014). Six of the original study results had unqualified support from the replication, two others had qualified support, and two others were inconsistent with the PPIR results. Bayesian analyses quantified the evidence for the alternative hypothesis relative to the null hypothesis rather than simply providing a binary score card. This crowdsourcing approach to replication prior to publication is a promising new direction. Similar collaborative work is being done by the Many Labs project (Ebersole et al. 2016, Klein et al. 2014).

UNINTENDED SIDE EFFECTS: DO NEW NORMS AND PROCEDURES RESULT IN COLLATERAL DAMAGE TO SOME SCIENTISTS?

As calls have been made to change the way science is conducted in psychology by preregistering designs and analyses and increasing sample sizes, some authors have noted what might be called collateral damage. The three types of damage that have been identified are (*a*) slowing and ultimate reduction of new findings and phenomena, (*b*) penalizing different subfields with the imposition of one-size-fits-all norms, and (*c*) discouraging young scientists from staying in the field because of the higher bar for publication and professional advancement.

Finkel et al. (2015, 2017) argued that the new conventions and norms overemphasize reducing false positive errors at the expense of allowing false negative errors. In the effort to minimize reports of findings that do not replicate, the norms can discourage discovery, innovation, and astute observation. Preregistration and other efforts that minimize HARKing may decrease reports of legitimate insights that occur when examining data. The new norms suggest that such new data-driven exploratory findings should be checked using independent data from a new confirmatory study, but Finkel et al. (2015, 2017) pointed out that obtaining new data in fields such as relationship science, health psychology, or neuroscience (among others) might take years rather than months. They also argued that the current emphasis is on being skeptical about a purported interesting finding rather than about a report that an interesting prediction was not found. Baumeister (2016) even suggested that neglect of discovery in psychology will lead to the field becoming boring.

Although the issues of replication in science go well beyond psychology (Dickersin & Rennie 2012, Ioannidis 2005), many of the advances in open science have come from psychological scientists. Especially active have been those in social psychology, which has been singled out as having lower rates of replication (Open Sci. Collab. 2015). The norms for sample size, multiple close replications of a finding, preregistration, data sharing, and material sharing have received particular attention from social psychologists and from methodologists who do no empirical research themselves. Finkel et al. (2015) made a compelling argument that these norms should not be rigidly applied to all psychology subfields. Rules of thumb for minimum sample size do not take into account the time and expense of data collection, the statistical power afforded by well-specified longitudinal models, or effect size variability across different subareas. Finkel et al. (2015) also argue that sharing original data of video recordings may not be possible and that sharing all research materials in longitudinal studies would place an extreme burden on investigators. These concerns are not limited to relationship science. Similar concerns could apply to studies of motor development (e.g., Adolph et al. 2012), clinical science (e.g., Tackett et al. 2017), and neuroscience (e.g., Overath et al. 2015). If the suggestions and norms become calcified rules and regulations, progress in some areas of psychology could be diminished.

The shifting publication rules and extra effort needed to document designs and data before making them public or open have led to concerns about the pipeline of graduate students, postdocs, and untenured junior faculty (Baumeister 2016). Even if we agree that new procedures lead to fewer false positive claims, they may also suppress discovery of true positive results, and they take longer to implement—particular problems for graduate students, postdocs, and junior faculty. Finkel et al. (2017) argued that researcher resources should be considered a fixed quantity; thus, larger sample size requirements will necessarily result in fewer new studies and publications. Search and promotion committees at institutions need to recognize the costs of increased documentation and sample sizes and account for these when making hiring or promotion decisions.

CONCLUDING COMMENTS: WHAT IS THE TAKE-HOME MESSAGE ABOUT HOW PSYCHOLOGICAL SCIENCE CAN SPEED KNOWLEDGE CONSTRUCTION?

The most recent replication crisis has led to a number of important insights and practices for psychological sciences. First, just because a finding has been claimed on the basis of a statistical test (NHST or Bayesian) does not mean that a new study will obtain the same result. Second, multiple replication studies of important findings advance knowledge by affirming findings, identifying boundary conditions, or showing legitimate lack of replication. Third, openness with regard to the documentation of the development of ideas, the exploration of preliminary data, the analysis of confirmatory data, and the data themselves will help allay concerns about fraud, QRPs, and human error. Fourth, the community of scientists should support one another and the scientific enterprise by providing (*a*) resources for open preregistration of ideas, research plans and data; (*b*) new statistical methodology for gaining insights from existing data and for planning new informative data collection; and (*c*) collaborative crowdsourcing resources for replicating important scientific claims. Fifth, treating the dialectical relationship between exploratory and confirmatory research honestly and seriously can further the goals of identifying legitimate and meaningful knowledge production. In short, the replication crisis can be reframed as a mandate from the research community to engage in methodologically sound, ethically driven research in which probabilistic decisions are made explicitly and respected by an open research process.

To promote consistent and high-quality research, a number of groups have made lists of best practice guidelines (Asendorpf et al. 2013, Brandt et al. 2014, Fiedler 2017, Munafò et al. 2017, Nosek et al. 2015, Open Sci. Collab. 2017), and themes from these lists have been emphasized by journal editors (Brown et al. 2014, Giner-Sorolla 2016, Kawakami 2015, Lindsay 2015, Vazire 2016). We have collected a large number of these recommendations in **Table 1**, organized into sections related to (*a*) initial steps, (*b*) conduct of the study and analysis, (*c*) preparation of the scientific report, and (*d*) open science. The last section includes archiving data, analysis scripts, and materials, but implied in all sections is the widely endorsed emphasis on openness.

We are especially impressed by the resources that are being provided by the Center for Open Science through the OSF to facilitate openness in phases of scientific discovery and empirical confirmation. Some researchers were initially concerned that establishing norms of preregistration would stifle creativity and exploration, but the OSF has adapted its resources to make it possible to archive different types of plans and ideas and to control the degree to which these notes and plans are made public. OSF provides a virtual laboratory notebook where scientists can make lasting notes in ink (not pencil) about what ideas they are pursuing, the kinds of data required, and the requisite steps necessary to achieve their scientific goals. If the researcher is undecided about which measure to use or does not know if a procedure borrowed from one research context will work in a novel context, these reflections can be noted in the OSF for private use initially and for public disclosure after research progress has been made. To some researchers, this recording may seem obsessive and time consuming; nevertheless, we endorse this systematic and thoughtful approach, and we expect that many journals will adopt the standards that are emerging from the OSF (Nosek et al. 2015).

Table 1 explicitly encourages systematic thinking, whether the research is in an exploratory, confirmatory, or mixed mode. In the left-hand column, we list recommended activities, and in the right-hand column, we indicate whether the recommendation is relevant to confirmatory or exploratory research modes; we were struck by how often the recommendation is relevant for both confirmatory and exploratory research. The "Initial steps" heading emphasizes planning, in terms of research goals, study design, attention to reliable measurement, and setting acceptable Type I

Table 1 Recommendations for research practices designed to speed knowledge construction in psychology and to reduce concerns about replication success in both exploratory (E) and confirmatory (C) studies

Recommendation	Study type
Initial steps	
Articulate goals of research	E, C
Consider power-enhancing designs	E, C
Set Type I error rate for multiple tests	C
Address measurement error	E, C
Provide detailed power analysis	C
Set data collection start and stop rules	C
Preregister hypotheses	C
Preregister designs	E, C
Preregister materials	E, C
Preregister necessary exploratory steps	E, C
Preregister confirmatory analyses	C
Conduct of study and analysis	
Set aside data for confirmation	E, C
Report confidence and credibility intervals	E, C
Avoid binary statistical decisions	E, C
Justify restrictions on the sample	E, C
Enumerate possible effect modifiers	E, C
Model effect variation	E, C
Scientific report	
Report all confirmatory preregistered results	C
Report exploratory analyses	E, C
Be flexible about $p < 0.05$	E, C
Report negative findings	E, C
Disclose ad hoc decisions	E, C
Model causal process	C
Direct replication and cross-validate	C
Report log of all analyses done in online appendix	E, C
Open science	
Share primary data	E, C
Share analysis syntax	E, C
Provide relevant research materials	E, C
Collaborate to replicate	E, C
Add disclosure statement to reports	E, C
Use open-source software	E, C

error rates. In this phase, important decisions must be made, in confirmatory studies, about whether it is possible to find a constellation of intervention methods, outcome measurements, participant characteristics, time frames, and study settings that illustrate the prototype of the effect with a large effect size. Alternatively, decisions might be made to explore the generality of an effect already established under specific conditions. These decisions, as well as those about whether the study design involves within- or between-person effects and assessments about the range of likely effect sizes, have important implications for an appropriate power analysis. As these decisions are made, we recommend that they be registered (or revised) within the OSF system.

Once data are collected and available, analyses that are both exploratory and confirmatory will be carried out. If the data are plentiful, it is ideal to randomly set aside a portion of the data for a confirmation study in a cross-validation. Even with a confirmatory study, decisions often need to be made about which data are retained or set aside because of quality concerns, and whether certain items or measures display expected validity patterns. These decisions should be systematic, but researchers should avoid rules based on simple binary significance tests (e.g., Shrout & Yip-Bannicq 2017) and instead use confidence or credible intervals, recognizing that data are sometimes precise and sometimes indeterminate. A thorough analysis of both confirmatory and exploratory data evaluates whether associations and patterns are consistent across important subsets of the sample, such as males and females.

The actions under the "Open science" heading of **Table 1** emphasize full disclosure of all results, whether tests are statistically significant or not, and all ancillary analyses. If validation data are set aside, the replication results should be reported regardless of whether they support or conflict with the original claim. Even when a true effect is present, it is unlikely that objective replications will lead to tidy results (Schimmack 2012), and the presentation of all data allows other researchers a better understanding of the likely effect variation. One of the benefits of the recent replication publications is that editors and reviewers now expect more complicated reports. Another benefit of science in the twenty-first century is that many details can be archived in online supplements (see Nosek & Bar-Anan 2012 for one vision of the future).

Related to creating complete reports of planned and exploratory analyses and results is the archiving of original data, analysis syntax, descriptions of relevant measures, and descriptions of protocols. Many experts recommend that, when possible, analysis syntax for open-source software systems be used, rather than expensive commercial programs that are not always available to interested researchers. These materials can be archived in the OSF system.

Recommendations in **Table 1** and in similar lists should not be applied as general and rigid rules for psychological research. In particular, the practice of automatic rejection of studies because sample sizes are judged to be too small using a heuristic rule is the opposite of the thoughtful and critical practices that we hope are emerging from the replication storms of 2011–2016. That having been said, a careful consideration of effect sizes as distributions rather than as fixed points will often necessitate larger sample sizes in years to come.

Some commentators have indicted psychological science for too many false claims and shoddy practices. Our view is that recent attention to replication in particular and knowledge generation more generally has led to remarkable and positive effects. This attention has generally sharpened our methodological efforts, opened laboratory settings for unprecedented and positive scrutiny, created expansive collaborative efforts and new methodological tools to combine results across studies, and allowed thoughtful exchange across disciplinary boundaries that has moved psychological practice and research substantially forward. The future of psychological science is bright.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank those who gave comments on earlier versions of this review, including participants in the Quantitative Methods Colloquium at Vanderbilt University, Couples Lab at New York University, Research Center for Group Dynamics at the University of Michigan, and the following individuals: Corina Berli, William Brady, Barry Cohen, Paul Eastwick, John Kruschke, Alison Ledgerwood, Uli Schimmack, Barbara Spellman, Gertrude Stadler, Rugile Tuskeviciute, Jay Van Bavel, Julian Wills, and Marika Yip-Bannicq.

LITERATURE CITED

Adolph KE, Cole WG, Komati M, Garciaguirre JS, Badaly D, et al. 2012. How do you learn to walk? Thousands of steps and dozens of falls per day. *Psychol. Sci.* 23(11):1387–94

Aldrich J. 2008. R. A. Fisher on Bayes and Bayes' theorem. *Bayesian Anal.* 3(1):161–70

Alogna VK, Attaya MK, Aucoin P, Bahník Š, Birch S, et al. 2014. Registered replication report. *Perspect. Psychol. Sci.* 9(5):556–78

Anderson SF, Maxwell SE. 2017. Addressing the "replication crisis": using original studies to design replication studies with appropriate statistical power. *Multivar. Behav. Res.* 52:305–24

Asendorpf JB, Conner M, de Fruyt F, de Houwer J, Denissen JJA, et al. 2013. Recommendations for increasing replicability in psychology. *Eur. J. Personal.* 27(2):108–19

Bakker M, Wicherts JM. 2011. The (mis)reporting of statistical results in psychology journals. *Behav. Res. Methods* 43(3):666–78

Baumeister. 2016. Charting the future of social psychology. *J. Exp. Soc. Psychol.* 66:153–58

Beasley WH, Rodgers JL. 2009. Resampling theory. In *Handbook of Quantitative Methods in Psychology*, ed. R Millsap, A Maydeu-Olivares, pp. 362–86. Thousand Oaks, CA: Sage

Bench SW, Rivera GN, Schlegel RJ, Hicks JA, Lench HC. 2017. Does expertise matter in replication? An examination of the reproducibility project: psychology. *J. Exp. Soc. Psychol.* 68:181–84

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57(1):289–300

Bhattacharjee Y. 2013. The mind of a con man. *The New York Times Magazine*, Apr. 28

Bonett DG. 2009. Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychol. Methods* 14(3):225–38

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2009. *Introduction to Meta-Analysis*. Hoboken, NJ: Wiley

Box JF. 1978. *R. A. Fisher: The Life of a Scientist*. Hoboken, NJ: Wiley

Brandt MJ, IJzerman H, Dijksterhuis A, Farach FJ, Geller J, et al. 2014. The replication recipe: What makes for a convincing replication? *J. Exp. Soc. Psychol.* 50:217–24

Braver SL, Thoemmes FJ, Rosenthal R. 2014. Continuously cumulating meta-analysis and replicability. *Perspect. Psychol. Sci.* 9(3):333–42

Brown SD, Furrow D, Hill DF, Gable JC, Porter LP, Jacobs WJ. 2014. A duty to describe. *Perspect. Psychol. Sci.* 9(6):626–40

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14(5):365–76

Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Abingdon, UK: Routledge. 2nd ed.

Cohen J. 1994. The Earth is round ($p < 0.05$). *Am. Psychol.* 49(12):997–1003

Coyne JC. 2016. Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychol.* 4(1):28

Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. 1972. *The Dependability of Behavioral Measurements*. Hoboken, NJ: Wiley

Cumming G. 2014. The new statistics. *Psychol. Sci.* 25(1):7–29

Curran PJ, Hussong AM. 2009. Integrative data analysis. *Psychol. Methods* 14(2):81–100

De Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, et al. 2005. Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. *N. Engl. J. Med.* 352(23):2436–38

Dickersin K, Rennie D. 2012. The evolution of trial registries and their use to assess the clinical trial enterprise. *JAMA* 307(17):1861–64

Ding N, Melloni L, Zhang H, Tian X, Poeppel D. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat. Neurosci.* 19(1):158–64

Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, et al. 2016. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* 67:68–82

Edwards J, Berry J. 2010. The presence of something or the absence of nothing: increasing theoretical precision in management research. *Organ. Res. Methods* 13(4):668–89

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7(1):1–26

Efron B, Gong G. 1983. A leisurely look at the bootstrap, jackknife, and cross-validation. *Am. Stat.* 37:36–48

Fabrigar LR, Wegener DT. 2016. Conceptualizing and evaluating the replication of research results. *J. Exp. Soc. Psychol.* 66:68–80

Fiedler K. 2017. What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspect. Psychol. Sci.* 12(1):46–61

Finkel EJ, Eastwick PW, Reis HT. 2015. Best research practices in psychology: illustrating epistemological and pragmatic considerations with the case of relationship science. *J. Personal. Soc. Psychol.* 108(2):275–97

Finkel EJ, Eastwick PW, Reis HT. 2017. Replicability and other features of a high-quality science: toward a balanced and empirical approach. *J. Personal. Soc. Psychol.* 113(2):244–53

Fisher RA. 1925. *Statistical Methods for Research Workers*. Edinburgh, Scotl.: Oliver & Boyd

Fisher RA. 1958. The nature of probability. *Centen. Rev. Arts Sci.* 2:261–74

Francis G. 2012a. The psychology of replication and replication in psychology. *Perspect. Psychol. Sci.* 7(6):585–94

Francis G. 2012b. Publication bias and the failure of replication in experimental psychology. *Psychon. Bull. Rev.* 19(6):975–91

Francis G. 2012c. Too good to be true: publication bias in two prominent studies from experimental psychology. *Psychon. Bull. Rev.* 19(2):151–56

Gallistel CR. 2009. The importance of proving the null. *Psychol. Rev.* 116(2):439–53

Giner-Sorolla R. 2016. Approaching a fair deal for significance and other concerns. *J. Exp. Soc. Psychol.* 65:1–6

Greenwald AG. 1976. An editorial. *J. Personal. Soc. Psychol.* 33(1):1–7

Hagger MS, Chatzisarantis NLD. 2016. A multilab preregistered replication of the ego-depletion effect. *Perspect. Psychol. Sci.* 11(4):546–73

Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Med.* 2(8):e124

Kawakami K. 2015. Editorial. *J. Personal. Soc. Psychol.* 108(1):58–59

Kerr N. 1998. HARKing: hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* 2(3):196–217

Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, et al. 2016. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLOS Biol.* 14(5):e1002456

Klein RA, Ratliff KA, Vianello M, Adams RB, Bahnik S, et al. 2014. Investigating variation in replicability: a "many labs" replication project. *Soc. Psychol.* 45(3):142–52

Kruschke JK. 2015. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan*. Burlington, MA: Academic. 2nd ed.

Kruschke JK, Liddell TM. 2017. The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.* 24:1–29

LeBel EP, Campbell L, Loving TJ. 2017. Benefits of open and high-powered research outweigh costs. *J. Personal. Soc. Psychol.* 113(2):254–61

Ledgerwood A, Sherman JW. 2012. Short, sweet, and problematic? The rise of the short report in psychological science. *Perspect. Psychol. Sci.* 7(1):60–66

Ledgerwood A, Soderberg C, Sparks J. 2017. Designing a study to maximize informational value. In *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency*, ed. MC Makel, JA Plucker, pp. 33–58. Washington, DC: Am. Psychol. Assoc.

Lindsay DS. 2015. Replication in psychological science. *Psychol. Sci.* 26(12):1827–32

Maxwell SE, Lau MY, Howard GS. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* 70(6):487–98

McArdle JJ, Ritschard G. 2014. *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*. Abingdon, UK: Routledge

McShane B, Böckenholt U. 2014. You cannot step into the same river twice: when power analyses are optimistic. *Perspect. Psychol. Sci.* 9(6):612–25

McShane BB, Böckenholt U, Hansen KT. 2016. Adjusting for publication bias in meta-analysis. *Perspect. Psychol. Sci.* 11(5):730–49

Meiser T. 2011. Much pain, little gain? Paradigm-specific models and methods in experimental psychology. *Perspect. Psychol. Sci.* 6(2):183–91

Mill JS. 2008 (1843). *A System of Logic*. London: Longmans, Green

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, et al. 2017. A manifesto for reproducible science. *Nat. Hum. Behav.* 1(1):21

Nelson LD, Simmons J, Simonsohn U. 2018. Psychology's renaissance. *Annu. Rev. Psychol.* 69. In press

Neyman J, Pearson ES. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika* 20A(1/2):175–240

Neyman J, Pearson ES. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 231:289–337

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, et al. 2015. Promoting an open research culture. *Science* 348(6242):1422–25

Nosek BA, Bar-Anan Y. 2012. Scientific utopia: I. Opening scientific communication. *Psychol. Inq.* 23(3):217–43

Nosek BA, Lakens DD. 2014. Registered reports: a method to increase the credibility of published results. *Soc. Psychol.* 45(3):137–41

Nosek BA, Spies J, Motyl M. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* 7(6):615–31

Offord DR, Kraemer HC, Kazdin AE, Jensen PS, Harrington R. 1998. Lowering the burden of suffering from child psychiatric disorder: trade-offs among clinical, targeted, and universal interventions. *J. Am. Acad. Child Adolesc. Psychiatry* 37(7):686–94

Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716

Open Sci. Collab. 2017. Maximizing the reproducibility of your research. In *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, ed. SO Lilienfeld, ID Waldman, pp. 3–21. New York: Wiley

Overall NC, Girme YU, Lemay J, Edward P, Hammond MD. 2014. Attachment anxiety and reactions to relationship threat: the benefits and costs of inducing guilt in romantic partners. *J. Personal. Soc. Psychol.* 106(2):235–56

Overath T, McDermott JH, Zarate JM, Poeppel D. 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci.* 18(6):903–11

Ozomaro U, Wahlestedt C, Nemeroff CB. 2013. Personalized medicine in psychiatry: problems and promises. *BMC Med.* 11(1):132

Paolacci G, Chandler J. 2014. Inside the Turk: understanding Mechanical Turk as a participant pool. *Curr. Dir. Psychol. Sci.* 23(3):184–88

Perugini M, Gallucci M, Costantini G. 2014. Safeguard power as a protection against imprecise power estimates. *Perspect. Psychol. Sci.* 9(3):319–32

Richard FD, Bond CF, Stokes-Zoota JJ. 2003. One hundred years of social psychology quantitatively described. *Rev. Gen. Psychol.* 7(4):331–63

Rogers JL, Howard KI, Vessey JT. 1993. Using significance tests to evaluate equivalence between two experimental groups. *Psychol. Bull.* 113(3):553–65

Rossi JS. 1990. Statistical power of psychological research. *J. Consult. Clin. Psychol.* 58(5):646–56

Scheibehenne B, Jamil T, Wagenmakers E. 2016. Bayesian evidence synthesis can reconcile seemingly inconsistent results. *Psychol. Sci.* 27(7):1043–46

Schimmack U. 2012. The ironic effect of significant results on the credibility of multiple-study articles. *Psychol. Methods* 17(4):551–66

Schmidt FL, Oh I. 2016. The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Arch. Sci. Psychol.* 4(1):32–37

Schooler JW, Engstler-Schooler TY. 1990. Verbal overshadowing of visual memories: Some things are better left unsaid. *Cogn. Psychol.* 22(1):36–71

Schweinsberg M, Madan N, Vianello M, Sommer SA, Jordan J, et al. 2016. The pipeline project: prepublication independent replications of a single laboratory's research pipeline. *J. Exp. Soc. Psychol.* 66:55–67

Seaman MA, Serlin RC. 1998. Equivalence confidence intervals for two-group comparisons of means. *Psychol. Methods* 3(4):403–11

Shrout PE, Yip-Bannicq M. 2017. Inferences about competing measures based on patterns of binary significance tests are questionable. *Psychol. Methods* 22(1):84–93

Simmons J, Nelson L, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22(11):1359–66

Simons DJ, Holcombe AO, Spellman BA. 2014. An introduction to registered replication reports at *Perspectives on Psychological Science*. *Perspect. Psychol. Sci*. 9(5):552–55

Simonsohn U. 2013. Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychol. Sci.* 24(10):1875–88

Sripada C, Kessler D, Jonides J. 2014. Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychol. Sci.* 25(6):1227–34

Stewart LA, Parmar MKB. 1993. Meta-analysis of the literature or of individual patient data: Is there a difference? *Lancet* 341(8842):418–22

Stigler SM. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard Univ. Press

Stroebe W. 2016. Are most published social psychological findings false? *J. Exp. Soc. Psychol.* 66:134–44

Tackett JL, Lilienfeld SO, Johnson SL, Krueger RF, Miller JD, et al. 2017. It's time to broaden the replicability conversation: thoughts for and from clinical psychological science. *Perspect. Psychol. Sci.* 12(5):742–56

Taylor DJ, Muller KE. 1996. Bias in linear model power and sample size calculation due to estimating noncentrality. *Commun. Stat. Theory Methods* 25(7):1595–610

Tukey JW. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley

Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA. 2016. Contextual sensitivity in scientific reproducibility. *PNAS* 113(23):6454–59

Vazire S. 2016. Editorial. *Soc. Psychol. Personal. Sci.* 7(1):3–7

Verhagen J, Wagenmakers E. 2014. Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* 143(4):1457–75

Vul E, Harris C, Winkielman P, Pashler H. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4(3):274–90

Wade N. 2010. Inquiry on Harvard lab threatens ripple effect. *The New York Times*, Aug. 12

Wagenmakers E, Verhagen J, Ly A. 2016. How to quantify the evidence for the absence of a correlation. *Behav. Res. Methods* 48(2):413–26

Wagenmakers E, Wetzels R, Borsboom D, van der Maas HLJ, Kievit RA. 2012. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* 7(6):632–38

Westfall J, Judd CM, Kenny DA. 2015. Replicating studies in which samples of participants respond to samples of stimuli. *Perspect. Psychol. Sci.* 10(3):390–99

Zabell S. 1989. R. A. Fisher on the history of inverse probability. *Stat. Sci.* 4(3):247–56