Peer Review Guidelines Promoting Replicability and Transparency in Psychological Science

William E. Davis, Wittenberg University, davisw4@wittenberg.edu

Roger Giner-Sorolla, University of Kent, rsg@kent.ac.uk

D. Stephen Lindsay, University of Victoria, slindsay@uvic.ca

Jessica P. Lougheed, Purdue University, jlougheed@purdue.edu

Matthew C. Makel, Duke University, mmakel@tip.duke.edu

Matt E. Meier, Western Carolina University, mmeier@wcu.edu

Jessie Sun, University of California, Davis, jesun@ucdavis.edu

Leigh Ann Vaughn, Ithaca College, lvaughn@ithaca.edu

John M. Zelenski, Carleton University, john_zelenski@carleton.ca

Revision, 31 August 2018

Abstract

More and more psychological researchers have come to appreciate the perils of common but poorly justified research practices, and are rethinking commonly held standards for evaluating research. As this methodological reform expresses itself in psychological research, peer reviewers of such work must also adapt their practices to remain relevant. Reviewers of journal submissions wield considerable power to promote methodological reform, contributing to the advancement of a more robust psychological literature. We describe concrete practices that reviewers can use to encourage transparency, intellectual humility, and more valid assessments of methods and statistics.

Author Notes and Disclosures (in lieu of Method section)

Address correspondence to Roger Giner-Sorolla at R.S.Giner-Sorolla@kent.ac.uk.

Conflicts of Interest: The author(s) declared no conflicts of interest with respect to the authorship or the publication of this article.

Author Contributions: Authors are listed in alphabetical order. All authors contributed to idea generation and to an initial collaborative draft. Further refinement of this draft proceeded with the input of all authors, but with most of the rewriting coming from DSL and RG-S.

Peer Review Guidelines Promoting Replicability and Transparency in Psychological Science

Psychological science is undergoing a "renaissance" (Nelson, Simmons, & Simonsohn, 2018) or "credibility revolution" (Vazire, 2018) in understanding statistical inference, in standards for methodological rigor, and in expectations of what should be reported in scientific communications. These developments have come with a realization that previous standard practices, most notably the focus on multiple conceptual replications in a single research article, were not enough to ensure replicable and robust science. There is a growing call to raise the field's standards (Vazire, 2018), and this in turn will require access to more details of a study's methods, analyses, and data than was previously typically provided—information that is still often omitted from reports.

Our aim in this paper is to provide recommendations for reviewers to promote transparency, statistical rigor, and intellectual humility in research publications. Well-informed peer reviewers help journal editors make better decisions, not only about *whether* a piece of research should be published, but also about *how* the work is reported if it is published. Reviewers can influence reporting practices by requesting the transparency necessary for all readers to assess the quality of the evidence and the validity of conclusions in the paper (Morey et al., 2016; Vazire, 2017). Our advice applies particularly to quantitative research in psychology, but is also relevant to research in other fields of science, especially those that use inferential statistics.

This paper grew out of a workshop, "How to Promote Transparency and Replicability as a Reviewer," at the 2017 meeting of the Society for the Improvement of Psychological Science (SIPS). Workshop participants (including this paper's authors) read existing advice on reviewing provided for the occasion by 22 journal editors (available at https://osf.io/hbyu2 and https://osf.io/swgyz), Roediger's (2007) 12 Tips for Reviewers in the *APS Observer*, a

chapter on reviewing by Tesser and Martin (2006), and an excerpt from Commitment to Research Transparency (Schönbrodt, Maier, Heene, & Zehetleitner, 2015). Workshop members then put together a set of new recommendations aimed at promoting transparency and replicability. This article will first explain some of the issues underlying our advice, then present our recommendations.

The New Approach to Statistical Inference and Reporting

Most empirical papers in psychology use null hypothesis significance testing (NHST) as a metric of evidence. In NHST, inferential analyses such as t-tests yield estimates of the probability (*p*) of the obtained result (or a more extreme result) occurring by chance under the null hypothesis of no effect. If *p* is low enough, usually under the conventional $p < .05$ threshold, the result is deemed "statistically significant." Significance can be taken as a heuristic indicating that the direction of the effect in the sample is likely to be the same as in the population (Krueger & Heck, in press). However, problematic practices call into question the usual ways in which statistical significance, in particular the criterion of $p < .05$, has informed publication decisions.

NHST is accurate only in confirmatory research, in which the research specifies the hypotheses to be tested and the method of testing *before* examining the data (Simmons, Nelson, & Simonsohn, 2011). But in practice, researchers sometimes decide which analyses to run based on which tests produce the most favorable results, and then report those analyses *as if* they had been planned in advance. Similarly, researchers sometimes adjust their procedures while analyzing their data (e.g., dropping some subjects, observations, dependent variables, or conditions; adding covariates; transforming measures) and fail to report these adjustments. All these practices may reflect a desire for brevity and a stronger narrative—spurred as much by editorial standards as by the authors themselves.

This sort of flexible, post hoc approach to NHST has been common practice in many areas of psychology (John et al., 2012). Unfortunately, these practices make *p* values misleading. Different critics have used different terms to highlight various aspects of the problem (e.g., HARKing, Kerr, 1998; researcher degrees of freedom, Simmons et al., 2011; *p*-hacking, Simmons, Nelson, & Simonsohn, 2012; the garden of forking paths, Gelman & Loken, 2014; questionable research practices, John, Loewenstein, & Prelec, 2012). Regardless of terminology, these practices can exaggerate estimates of the sizes of effects and inflate the risk of falsely rejecting the null hypothesis. When "significant" *p* values obtained via undisclosed flexibility are presented as if they arose from planned tests of hypotheses, readers are likely to conclude that the evidence is stronger than it actually is.

It is good and proper for researchers to conduct exploratory research as well as hypothesis-testing research. Poking around in one's data, speculating about unexpected patterns, is a great way to generate ideas. For conducting such exploratory analyses, confidence intervals and estimates of effect size are useful tools (e.g., McIntosh, 2017). But NHST *p* values become meaningless when the data drive decisions about which tests to run and how to run them, because more risks have been taken than the *p* value takes into account. At a minimum, reviewers and readers need to know how researchers made their data-analysis decisions.

Vazire (2017) drew an analogy between readers of science articles and used-car shoppers: Transparent reporting puts readers in a better position to tell the difference between "lemons" and trustworthy findings. One powerful tool for promoting such transparency is a preregistered research plan (preregistration; see Lindsay, Simon, & Lilienfeld, 2016; van't Veer & Giner-Sorolla, 2016). Preregistration makes clear which aspects of a study and its analyses were planned in advance of data collection. Openly sharing data and materials (e.g., tests, stimuli, programs), and explicitly declaring that methodological details have been

completely reported (e.g., the Simmons et al., 2012, “21-word solution”), can also help readers to assess the evidence value of an empirical report.

To allow for correction of mistakes in reporting and for exploration of alternative analyses and explanations, transparency requires that researchers make raw data available to other researchers, along with codebooks and analysis scripts. Despite protocols requiring such sharing for verification (e.g., Section 8.14 of the American Psychological Association’s ethical principles, http://www.apa.org/ethics/code/), the availability of data has often been poor (e.g., Wicherts, Borsboom, Katz, & Molenaar, 2006). Finally, authors can also advance transparency by providing more comprehensive descriptive statistics, such as data graphs that show the distribution of scores.

Making defensible claims in research reports also entails intellectual humility about the limitations of one’s own perspective and findings (Samuelson et al., 2015). Scientific claims require a realistic perspective on the generalizability of one’s own research and views. In moving from a standard that prioritizes novelty to one that emphasizes robustness of evidence, claims about the importance of any one study or series of studies should be limited, and replications should be encouraged. Researchers should also strive to be aware of the assumptions they bring to conducting and evaluating research—for example, ideas about what constitutes a “standard” or “unusual” sample (see Henrich, Heine, & Norenzayan, 2010) or preconceptions about research that has political implications (Duarte et al., 2015).

Over the past decade, some journals in psychology and other fields have adopted more open reporting requirements such as those outlined in the Transparency and Openness Promotion (TOP) guidelines (Nosek et al., 2015; https://cos.io/our-services/top-guidelines/). Over 5,000 journals and organizations have become signatories of the TOP guidelines, and over 850 journals have implemented the standards. However, many journals have not changed their policies, and editors and reviewers vary in implementing these reforms. Our

aim with the following recommendations is to provide concrete guidelines showing how you, as a peer reviewer of empirical research articles, can encourage transparency, statistical rigor, and intellectual humility. We organize these guidelines, roughly, in the order they will come up as you deal with a review. Appendix A gives a slightly reorganized outline of our advice that can be used as a checklist during the review process.

Preparing to Review

Know your stuff. To be able to understand and communicate criticism of problems in research you review, ensure you have a solid grasp of the key statistical issues. Appendix B lists selected educational resources, with more specific explanations in the section on evaluating research. Although specific statistics applications vary across fields, good reviewers should sharpen their understanding of the following issues that often are forgotten after postgraduate statistical training:

- The logic of NHST: If you understand why the *p*-value is not itself "the probability that the null hypothesis is true" (e.g., Cohen, 1994), you have come farther than many.
- The need for a priori specification of hypothesis tests, and methods used to control selective reporting, such as pre-registration, openness about exploratory analyses, methodological disclosure statements (Simmons, Nelson & Simonsohn, 2012), and open materials.
- Assumptions underlying frequently used statistical tests in your research area, and in particular, knowing when the test is not robust to violations.

As a source of inspiration, the APA's Journal Article Reporting Standards (JARS; 2018) lists desirable features for reporting in all types of research article, including qualitative, meta-analytic, and mixed methods. Using JARS as a checklist, you can look for the methodological and statistical considerations that are particularly important to report in your area of research, and carry out further reading to ensure you understand their rationale.

Reading and Evaluating the Paper

Evaluate statistical logic and reporting. You might think that all editors of scientific journals in psychology are statistically savvy, but you'd be wrong. Unfortunately, it is possible to become an eminent scholar and gatekeeper in psychology while one's statistical knowledge stays focused on the skills that help get articles published, rather than on statistical best practices. Even if journals espouse improved statistical standards or refer back to general guidelines, such as those in the *Publication Manual of the American Psychological Association* (American Psychological Association, 2010), editors do not always enforce such guidelines before sending the manuscript to reviewers. It is often up to you, the reviewer, to insist on complete statistical reporting for the sake of transparency.

Of course, editors and authors may privilege other goals above full statistical reporting, such as manuscript readability or word count limits. Your suggestions for increasing the amount of reporting should take into account what is possible at the journal, as specified in its submission guidelines, which should be available on the website (sometimes known as "Guide for Authors" or "Instructions for Authors"). Limitations caused by word counts, for example, can be overcome by adding details in supplementary online materials (which many journals now offer) or on public repositories such as the Open Science Framework (http://osf.io).

Beyond enforcing the journal's own standards, the issues you look for will depend on your own knowledge and preparation. Here are several frequently encountered issues:

- Many psychology studies cannot obtain precise results because they do not have sufficient sample size to provide accuracy in parameter estimation (AIPE; Maxwell, Kelly, & Rausch, 2008; see also Cumming 2014). That fact has been known for decades (Cohen, 1962), but only recently has awareness become widespread. Accuracy allows inference to go beyond a merely directional finding, allowing comparison of the finding's

effect size to other known influences on the outcome, and evaluating it as a potential basis for real-world applications. Precision for planning, AIPE, and statistical power analysis can all help readers judge the sensitivity of methods, which has implications for interpreting both positive and null results. All these techniques are preferable to criticizing a study based on your idea of what a "low N" looks like. Some methods, such as repeated-measures designs, can yield precise results or high power with a surprisingly low number of participants (Smith & Little, 2018).

- Effect sizes, and related statistics such as confidence intervals, are important adjuncts to significance tests that help readers interpret data more fully, especially when samples are unusually large or small (Cumming, 2014; Howell, 2010). Even if effect sizes are reported in results sections, check to see that the discussion of results takes into account their magnitude and precision, rather than only basing conclusions on the *p*-value.
- Power analysis tests the likelihood of rejecting the null hypothesis if the alternative hypothesis is true, a reporting feature that journals are increasingly requiring. Not all power analyses are equal, though. Post-hoc power analyses, for instance, are uninformative, being merely a function of the *p*-value (Goodman & Berlin, 1994). Best practice is to base the sample size on a reported *a priori* power analysis, including a rationale for deciding the expected effect size that must be input to these calculations (e.g., prior literature, or estimates of the typical effect size for the field and methodology if studying an entirely new effect). If power analysis was not done a priori, you can still request a sensitivity power analysis that outputs minimum effect sizes that the study could have detected, based on actual *N* and one or more levels of desired power (Lakens, 2014). A study that can only detect a conventionally "large" effect at 80% power is not well powered to detect the small and medium-sized effects that are more characteristic of

many areas of psychological research. For further reasons to prefer well-powered research, see the section below, "Evaluate sensitivity as well as validity."

- "Optional stopping" refers to the practice of deciding whether to stop or extend data collection based on the outcome of a hypothesis test on preliminary data. Researchers might plan to stop data collection after a certain number of cases if the hypothesized effect is then statistically significant, and to continue data collection if it is not. This procedure might continue until a criterion significance is reached, or until a maximum number of cases has been reached. Optional stopping can be acceptable if the researcher adjusts the alpha level accordingly (e.g., Lakens, 2014; Sagarin, Amber, & Lee, 2014) or uses appropriate Bayesian analyses (e.g., Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). However, using the unadjusted .05 threshold with optional stopping inflates the Type I error rate. As a reviewer, it is hard to detect unreported stopping rules, but you can look for or request a disclosure statement that explicitly describes how sample size was determined at each stage (Simmons et al., 2012).
- Descriptive statistics, such as cell *n*, means, standard deviations, and correlations between multiple measures, are sometimes omitted from more advanced statistical reports. Insist on seeing them anyway, because they may reveal underlying problems that qualify the fancier analysis. For example, means might be high on the scale and low in variance (floor/ceiling effect), violating the assumptions of the statistical test; or two variables might be so highly correlated (e.g., .8 or above) that drawing distinctions between them is problematic. And if a complex, multi-variable model gives results that appear at odds with the basic zero-order correlations in the data, it is important to understand why.
- Basic statistical errors are surprisingly common in published research (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Being roughly familiar with the formulas for degrees of freedom in commonly used statistical tests (e.g., Howell, 2010)

can help you detect discrepancies between reported participant numbers and the actual numbers tested. There are also tools for checking whether the decimal places of reported means are impossible to obtain given the reported numbers in a condition (e.g., Brown & Heathers, 2016). Both problems may point to undisclosed missing or excluded data. You might also want to run StatCheck (http://statcheck.io/; Epskamp & Nuijten, 2016) on papers you review. This free program detects discrepancies between some of the most common inferential statistical indices (e.g., F, r, t, z), the reported degrees of freedom, and the reported *p* value.

Assess any preregistrations. As noted earlier, it has been common practice in psychology to report the outcome of exploratory analyses as though those analyses had been planned a priori (John et al., 2012). *Preregistration* involves posting a time-stamped record of method and analysis plans online prior to data collection. Its aim is to make analytic flexibility transparent, helping reviewers better evaluate the research. A common misconception is that a preregistration is meant to restrict the carrying out of analyses; actually, preregistrations do allow additional post-hoc analyses, but their purpose is to make sure post-hoc analyses are clearly labeled as such (e.g., van ‘t Veer & Giner-Sorolla, 2016).

If a preregistered plan for the research is available, it is important to assess the level of completeness and detail in that plan compared to the procedures reported in the article. Some “preregistrations” are so brief and vague that they do little to identify when post-hoc liberties have been taken, providing only the illusion of transparency. Norms for assessing the quality of preregistrations are still in development. For one protocol, see Veldkamp (2017, https://psyarxiv.com/g8cjq/). If researchers do deviate substantially from their preregistered analyses, even for good reasons (e.g., the data failed to meet assumptions of the proposed test), you can ask them to also report the outcome of the preregistered analyses for full transparency (e.g., as an appendix).

If the research under review was not preregistered, it may be difficult to tell which analyses were planned in advance and which were data-dependent, but some clues may lead you to suspect post-hoc analysis. For example, data exclusion rules or transformations might be reported only in the Results sections and without any explicit rationale, or may vary from one study to the next without justification. The concern here is that the researchers may have (not necessarily intentionally) made analytic decisions to produce a significant result that would not replicate when using alternative reasonable analytic specifications, or in a new dataset. That doesn't mean that those results have no value, but they should be viewed with skepticism pending direct replication.

You can ask researchers to address concerns about post-hoc flexibility in your review. The strongest reassurance would come from a direct, preregistered replication. However, you can also ask the authors to indicate which analyses, if any, were exploratory, or to adopt a more stringent standard for statistical significance (e.g., $p < .005$; Benjamin et al., 2018). Finally, you can ask the researchers to demonstrate that their findings are robust under reasonable alternative specifications (e.g., with and without covariates, different exclusion criteria, model specifications; see Simonsohn, Simmons, & Nelson, 2015; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

Data and materials. If the authors submitted data, materials, and/or analysis code as part of the review process, or provided a link to a preregistration document detailing their data collection and analysis plans, you should determine whether these resources are in a usable form. If the materials are not available or usable, let the editor know, and ask if there is a way to obtain them. When they are available, we encourage you to examine such materials for completeness and accuracy. Data variables should clearly correspond to the variables reported in the text. Materials should allow a third party to re-run the study, mapping clearly onto the conditions, variables, and reporting. Running analyses with

available data is usually beyond the call of a reviewer's duty, but might be worth doing if it helps check apparent errors or strong alternative possibilities for the authors' conclusions.

Go beyond "*p* < .05 per study." For a long time, in many areas of psychology, reviewers judged whether a study supported a hypothesis by whether its key test was significant at $p < .05$. A multi-study paper was judged to support its hypotheses only when each study's key result was significant. To meet these standards, authors often omitted (or were asked to omit) non-significant studies from reports, even though statistically they were consistent with evidence favoring the hypothesis. Another part of playing this game was "*p*-hacking": selectively stopping data collection, excluding observations or conditions, applying data transformations, exploring covariates, or reporting one analysis out of many, all to achieve $p < .05$ (Simmons et al., 2011).

The distribution of *p*-values from all tests of a true hypothesis should lead us to expect relatively few results between $p = .01$ and $p = .05$ (Simonsohn, Nelson, & Simmons, 2014). The higher the statistical power of the test, the larger the proportion of results with $p < .01$, and the fewer nonsignificant results (assuming a true effect). For example, if power is 80%, then about 59% of confirmatory tests should yield $p < .01$, whereas only about 21% should yield *p* between .01 and .05 (Lakens, 2014; see also the interactive calculator at http://rpsychologist.com/d3/pdist/). But some literatures in psychology report too many significant results, relative to the power of the studies (Francis, 2014; Schimmack, 2012). And, although authors disagree on the evidence for a "bump" in reported *p*-values just under .05 (Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016; Masicampo & Lalande, 2012), there is a growing awareness that .05 is not a hard cut-off, and that single values close to it on either side are weak evidence (see Benjamin et al., 2018; and Lakens et al., in press, for contrasting views on whether or not psychology should set alpha at .005.)

So, be wary of multiple studies, each with key *p*-values just under .05. Values in this range are infrequent enough, and it should be even more rare to see them across multiple studies. The pattern might have arisen by chance, but you should seek assurance that it is not from selective reporting or *p*-hacking. A detailed and accurate preregistered analysis plan provides the greatest confidence (Lindsay et al., 2016; van't Veer & Giner-Sorolla, 2016). Without such evidence of constraints on researcher degrees of freedom, you might look for or request a disclosure statement that all measures, manipulations and exclusions were reported (Simmons et al., 2012; see https://osf.io/hadz3/ for a Standard Reviewer Statement).

Inzlicht (2015) gave an account of a lab that was encouraged to report all studies it had run to test a hypothesis, instead of just the significant ones, precisely because a paper it had submitted showed a pattern of *p*-values unusually close to significance. Including the lab's "file drawer" of nonsignificant findings, the overall picture still supported the hypothesis, albeit at a more modest effect size. Reporting all relevant studies, excluding only ones that fail methodological checks independently of the hypothesis, is a practice in line with both common-sense reporting ethics and the standards of professional bodies (American Psychological Association, 2010, p. 12). Although it is sometimes difficult to know when an unpublished study is part of the same or a different line of research, reviewers should encourage full reporting of studies that would have reasonably been included to support the argument of the paper at hand, had they come out with significant results.

Reviewers should also place less emphasis on the *p*-values of single studies. Better evidence can be gained from measures of precision (e.g., confidence intervals, credible intervals), or on Bayes factors, which provide a symmetrical measure of evidence for the null and alternative hypotheses (Cumming, 2014; Wagenmakers et al., 2017). Often, when presenting a series of studies and commenting on their individual significance, better

understanding can be had by aggregating comparable results over that series (Goh, Hall, & Rosenthal, 2016),

Aggregate evidence, however, becomes unreliable if only significant studies are reported. To mitigate publication bias, you can ask for an internal meta-analysis of all relevant studies conducted by the research team, which may include studies that were not included in the original report. But, by the same token, you should have realistic expectations about what a fully reported set of tests of a true hypothesis looks like (Lakens & Etz, 2017). Even a strongly supported proposition can sometimes include nonsignificant results here and there. Also, these considerations should not stop you from to recommending publication of methodologically strong single-study papers. One high-powered study can be more informative than several underpowered studies (Schimmack, 2012).

Evaluate measurement and manipulation validity. Reviewers should make sure that the constructs discussed in an article were indeed the constructs that were measured in the project. Ideally, assessments should be sensitive to differences in what the researchers intended to measure (across individuals or manipulations; Borsboom & Mellenbergh, 2004). The interpretation of findings based on improperly validated measures can be meaningless at worst, and suspect at best. Accessible discussions of these issues can be found in Flake, Pek, and Hehman (2017) and Fried and Flake (2018). Questions relevant to the validity of measures include:

- Have the authors reported scale reliabilities computed from their data? Indicators of internal consistency such as Cronbach's alpha are important to include but are commonly misreported as indicators of validity (Flake et al., 2017). In particular, a high alpha does not speak clearly to whether constituent items represent a single dimension or multiple dimensions. Factor analysis is needed to assess whether item intercorrelations match the intended structure, one aspect of valid measurement.

- Did the authors use previously validated measures? Check for reporting of, or references to, validation studies of the measures, including tests for construct, convergent, and divergent validity.
- Did the authors use measures as originally developed and validated, or have they modified the original scale? Have any modifications been well justified and fully reported? Modifying scales without reporting the full details can complicate replication studies, and modifying scales without assessing their validity can lead to uncertainty in measurement.
- Did the authors report findings based on single-item measures? Single-item measures may not adequately capture the intended construct. They require special consideration and validation (see Flake et al., 2017).

If reviewers find that answers to any of the above questions are unclear, it is important to request the missing information in your review. Authors should be encouraged to address weaknesses with measurement validity in the Discussion section of the manuscript, describing specifically how uncertainty in the measures used may affect the interpretation of the results and the generalizability of the study.

Evaluate sensitivity as well as validity. Measurement concerns are part of a larger issue that is becoming more important with increased understanding of methodology: sensitivity. Traditionally, psychology reviewers are keen to point out alternative explanations for a significant or *positive* result. Confounded manipulations, conceptually ambiguous measures, and statistical artifacts are just a few things that can threaten the interpretation of apparently positive results. Certainly, reviewers should stay on the lookout for all such issues.

In contrast, psychology reviewers are often less attuned to problems that might compromise the interpretation of non-significant findings, such as small sample size, weak manipulations, poor measurement reliability, restricted range, and ceiling or floor effects.

Such flaws can reduce a method's *sensitivity* (ability to detect a positive result). Low sensitivity may obscure a phenomenon that exists in the population but is missed or underestimated in the sample. This clouds the interpretation of non-significant results and casts doubt on the replicability of significant (*positive*) results. A common misconception is that a positive result is all the more impressive for having "survived" a study with low sensitivity (as criticized by Loken & Gelman, 2017). Reviewers should reject this view, and look out for flaws in the sensitivity as well as validity of methods.

Low sensitivity renders significant results relatively more likely to be false positives rather than true positives, especially when the finding is unlikely (Ioannidis, 2005, 2008; Zöllner & Pritchard, 2007). For example, if a finding is only 10% likely to be true and statistical power is low (50%), then 47% of $p < .05$ results will reflect a false effect. The false-positive problem, then, is likely to be particularly pernicious for surprising, counterintuitive findings not well-supported by theory.

Low-sensitivity methodology also sets a bad example. A lab that uses it is more likely to waste their effort on a false-negative finding, and their findings are less likely to be replicated. And, in a climate of low-sensitivity methodology, selective reporting can be justified more readily. If a study did not work, it is easy to say that the methods must have been bad, rather than taking it as evidence against the hypothesis (LeBel & Peters, 2011). Finally, many inferential statistical tests lose their robustness to violations of data assumptions under low sample size or other conditions of low sensitivity.

In experimental research, another sensitivity issue parallel to measurement validity is manipulation validity. It is common for researchers to take a short cut and assume that an effect of an independent variable (IV) on a dependent variable (DV) is sufficient proof that a manipulation is valid. But this assumption conflates the effect being tested (change in IV relates to change in DV) with the validity of the manipulation (manipulation effectively

changes IV). Especially when results are null, either in original research or a subsequent replication, showing that the manipulation is valid in the sampled population can help rule out manipulation failure as a prosaic explanation.

Ideally, a manipulation will be validated on a criterion variable that directly measures the independent variable. For example, if the accessibility of thoughts about power is being manipulated, then power words should be responded to more quickly in a decision task. This testing might be done in the same study that tests the main hypothesis, as a "manipulation check." If there are concerns about participant awareness, though, the testing can be done on a separate sample (Kidd, 1976). Although manipulation checks have previously been criticized as unnecessary (Sigall & Mills, 1998), such critiques were based on their inability to further enlighten positive results. With an increased emphasis on publishing and evaluating null results, testing manipulations has become more important.

Know how to evaluate null claims. Nonsignificant *p* values do not, by themselves, provide evidence for the null hypothesis. Evaluate a conclusion that an effect is nonexistent as carefully as you would evaluate a claim that it exists. Values of *p* greater than .10 are often obtained when the null hypothesis is false but sensitivity is low. If a manipulation causes a half-standard-deviation change in the population mean of a dependent variable (that is, effect size $d = .5$), then about half of experiments comparing two independent groups of 23 subjects will fail to reject that false null hypothesis at the .05 level (that is, statistical power is only .50). Bayesian approaches provide a more useful way to assess how much data favor the null hypothesis (Wagenmakers et al., 2017). Alternatively, equivalence tests based on NHST have also been developed (Lakens, 2017). Both procedures depend on assumptions about what range of effect sizes are functionally null, which should be described before reporting them. One does not need to be an expert in Bayesian or equivalence statistics to request that authors do more to justify or qualify conclusions that an effect is nonexistent.

The general problem of drawing misguided inferences from nonsignificant *p* values can creep up in many other forms. For example, if a report of model-fitting analyses interprets a non-significant chi-square statistic (or change in chi-square) to conclude that the model fits (or that two models fit equally well), you should consider whether the study was sufficiently powered to detect misspecifications (Hu & Bentler, 1998). Also, if researchers claim to find "full mediation" based on a non-significant direct effect (setting aside more general issues with statistical mediation; Bullock, Green, & Ha, 2010), you should consider how much power they had to detect small direct effects. In both cases, reviewers can ask researchers to provide power analyses or qualify their conclusions.

Moreover, the difference between significant and nonsignificant is often itself not statistically significant (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Be especially wary if authors interpret a significant effect in one condition or experiment, versus a non-significant effect in another, as informative, without reporting a test of the interaction between condition/experiment and effect. Similarly, when one correlation or regression coefficient is significant, another is not, and the authors claim that the first coefficient is significantly larger than the second, you can ask for appropriate statistical comparisons to support this claim (Clogg, Petkova, & Haritou, 1995; Steiger, 1980). These non-exhaustive examples illustrate the need for reviewers to be vigilant about appropriate interpretations of nonsignificant results.

Assess constraints on generality. Researchers have always been expected to describe limitations of their research in the Discussion section, but such statements are often pallid, incomplete, and drowned out by louder claims of the importance of the findings. Simons, Shoda, and Lindsay (2017) proposed a stronger and more structured "constraints on generality" (COG) statement, which identifies the aspects of a study (e.g., participants, materials, procedures, historical/temporal context) that the authors believe are essential to

observing the effect. This information is important in evaluating the contribution of the manuscript, and for facilitating replications and tests of boundary conditions. Just as important, the COG statement tends to foster intellectual humility about the generalizability and importance of the finding beyond the limited samples and materials in the research. Some journals already require a COG statement. As a reviewer, you can ask for one as well if the conclusions seem broader than can be justified by the studies.

Writing the Review

Address replicability. An important question to ask yourself when reviewing is: "How confident am I that a direct replication of this study would yield a similar pattern of findings?" Replicability is not the only characteristic of good science—the best work is also interesting, informative, and relevant—but it is a fundamental starting point. In your reviews, we recommend you cite specific reasons why you have (or lack) confidence in the replicability of the work, such as the statistical robustness, open reporting, and methodological sensitivity of the reported findings.

If replicability is in question, you might suggest in your review that the authors be invited to conduct a preregistered direct replication, perhaps with increased statistical power and/or other improvements, but designed to replicate the same study as exactly as possible. This may include a "no-fault" clause that makes clear that the new study will be evaluated independently of what the results show, as long as the overall case for the hypothesis is presented reasonably. This approach assumes that similar data can be obtained without tremendous burden (e.g., intensive methods, non-convenience samples). If not, a reviewer can insist that conclusions be calibrated to the strength of the data. Similarly, openly exploratory work may still be worth publishing if the discussion of results and limitations is appropriate, if the findings are theoretically informed and have potential to generate new hypotheses, and if the data and materials are publicly available (e.g., McIntosh, 2017).

Communicate your own limits. When you are not familiar with a methodology or statistical test used in a manuscript, it is important to communicate this to the editor, at the same time recognizing that your perspective on other issues may still be valuable. Acknowledging your limits is part of the practice of intellectual humility, and it helps editors become aware when they don't have the expertise on board they need. This may lead them to seek out the opinion of an expert in the topic.

Take the right tone. When we asked 22 editors what they would say to reviewers, the most frequent advice was to keep a constructive, respectful tone (see https://osf.io/hbyu2/). When reviewing with attention to transparency and replicability, it can be tempting to frame departures from best practices as dishonesty or cheating. Indeed, making accusations can be psychologically rewarding (Hoffman, 2014). Not surprisingly, researchers tend to respond defensively when terms like "questionable research practices," "p-hacking," etc. are aimed at them. However, many errors happen unintentionally, and many research practices now seen as inappropriate have long been standard in some areas of psychology, entrained by mentors and the gatekeepers of publication. In our view, a polite and reasoned tone is more likely to succeed. Explain the reasons for your recommendations; not all authors or editors are well educated in the new standards. Avoid inflammatory labels in favor of more neutral phrases, such as "low robustness." Always maintain a degree of humility, keeping in mind that your perceptions of flaws may be mistaken.

Promote transparency. If the manuscript does not include open science practices that give reviewers access to materials, analysis code, and/or the data, you may include in your review arguments for making such materials available in subsequent revisions. Your arguments may be directed to the editor as much as to the author. For example, if the journal endorses APA ethical standards for publishing, you could ask for a statement of full disclosure of measures, manipulations and exclusions, because those standards prohibit "[…]

omitting troublesome observations from reports to present a more convincing story […]" (American Psychological Association, 2010, p. 12). To support full disclosure, you could also invoke the American Statistical Association's guideline that *p*-values can only be interpreted correctly with full knowledge of the hypotheses tested (Wasserstein & Lazar, 2016) and note that with exploratory analyses, the focus should be on confidence intervals and effect sizes, rather than *p*-values. The strongest commitment to openness goals is represented by the PRO initiative (Morey et al., 2016), which involves an overt commitment to only complete a review if all data and materials are made available. No matter what form it takes, even if your request for more openness is denied, it will make the editor and author more aware of changing norms.

If the authors did provide data, materials, analysis code, or preregistrations, report in your review what depth of scrutiny you gave to these additional materials. Note any obstacles or limitations you encountered, for example, if you were unable to check the analysis code because you are not familiar with that programming language. It is not necessarily your job to make sure those resources are usable and correct. However, reporting the depth of your own efforts will help the editor fulfill his or her obligation.

Some journals offer special recognition in the form of "badges" granted to articles that meet criteria for transparent processes (e.g., an open-data badge, a pre-registration badge, and an open-materials badge; see https://osf.io/tvyxz/). If the journal for which you are reviewing offers such badges, consider mentioning that fact, with the aim of encouraging authors to share more information and improve the review process. If the article is already applying for badges, keep in mind that most journals rely on authors' declarations that the data, materials and/or preregistration are adequate. Authors and readers might benefit from your input if you check badge-supporting material for usefulness and completeness.

Think about signing reviews. Finally, you may also consider breaking the usual anonymity of peer review, signing your reviews to promote transparency and openness on your side of the process. There are good arguments for either signing or not signing all reviews (e.g., Peters & Ceci, 1982, and accompanying peer commentary; Tennant et al., 2017). We recommend adopting a general policy about whether you will or will not sign all reviews, at any given career stage. Without a general policy, you may be tempted to associate yourself with only the reviews that make a favorable impression (e.g., positive feedback) while avoiding accountability by not signing reviews that make a less favorable impression (e.g., critical feedback). If you do sign, we recommend you state explicitly that this is a general policy for you, after giving your name.

Signed reviews can have tangible benefits for authors, providing context for suggestions and a sense of fairness in critique, and for reviewers, giving exposure, credit and accountability. But signing also carries risk, especially if you are not yet in a permanent employment situation. Some authors may seek retribution if they feel their submissions have been inappropriately criticized. Reviewers with more job security and seniority, however defined, have less to lose by signing. These concerns are also relevant when deciding whether to accept requests to review for journals that have adopted open review practices such as unblinded review, publication of reviews alongside the final article, or direct interaction between authors and reviewers during the review process (see Ross-Hellauer, 2017; Walker & Rocha da Silva, 2015).

Special Cases: Replication Studies

The new approach to methods includes a growing willingness to publish close replications of previous research, which previously might have been rejected because they lacked novelty. Main concerns in a replication study are somewhat different from a primary research manuscript. You do not need to evaluate the theoretical rationale, and your analysis

of methods will focus on how closely the replication follows the original, and whether any changes in method are necessary or justified. Brandt et al. (2014) provide detailed guidance on what makes a replication strong. In brief, just as with original studies, reviewers should give more credence to replications that were preregistered, had adequate power, used methods shown to be sensitive (e.g., validating manipulations and measures in the new context), and provided detailed methods sections, open data, and analysis scripts. Given that most journals will publish replication results even if null, it becomes especially important to reduce the risk of failing to replicate due to insensitive methods.

If the authors bill their study as a close (or "direct") replication, their manuscript should report discrepancies between their study and the original study (Brandt et al., 2014). The importance of these discrepancies depends on the scope of the claims made in the original paper. For example, if the samples used in the original and replication studies differ in gender, age, ethnicity, or nationality, you should refer to the original paper to assess if the authors of the original paper generalized their claims across these demographics. If they did, and the replication finds weaker or opposite results when compared to the original study, it is fair for the replication authors to conclude that their findings reduce confidence in the original claim. However, if the original authors' claims were specific to a population, and the replication sampled a different population, it is not a close replication and does not directly address the original effect. Discrepancies may also need to be introduced, in order to reproduce the psychological effect in a new context. For example, when replicating a North American study on perceptions of baseball players, cricket would be a more appropriate sport to command participants' knowledge and interest in India.

In reviewing replications, you may have to assess claims about the new state of evidence, taking into account original and replication studies. Gelman (2016) suggested using a time-reversal heuristic to assess the evidence in a replication and the original study: if the

replication had been published first, would it seem more compelling? Just as no single study can determine whether an effect exists, neither can any replication. So, don't be too concerned with judging replications as "successful" or "failed." Instead, think meta-analytically, across the individual studies. Does the replication reinforce or change your beliefs about the effect (or does it do neither)? In any event, it is important to treat positive and negative results in a replication evenhandedly. Although failing to replicate a well-known effect may be more newsworthy than successfully replicating it, both types of evidence need to be reported for science to progress.

Some editors may ask you to judge how important it was to replicate the effect in the first place, as one would judge the importance of any novel research. In this case, weigh the strength of existing evidence, and the original research's impact on scholarship and society. If the effect has been closely replicated numerous times, has little theoretical or societal value, or has been largely ignored in the academic literature and press, then the replication may be judged as relatively less important (Brandt et al., 2014).

Special Case: Registered Reports.

More and more journals are inviting Registered Reports (RRs; see https://cos.io/rrr) as a special form of preregistration. In an RR, researchers submit a detailed proposal of a study to a journal for peer review before collecting the data. When data are collected, they then submit the complete manuscript reporting results, which will be accepted in principle regardless of results if the approved proposal has been followed faithfully. RRs are quite new, but their adoption appears to be increasing rapidly (see Nosek & Lindsay, 2018). Anecdotal reports indicate that reviewers find being involved with RRs gratifying. They can help researchers avoid mistakes in the first place, rather than just pointing out mistakes after they are made.

Peer review of RRs will involve you at two stages. In Stage 1, you will be asked to evaluate the importance and quality of the proposed study prior to data collection. At this stage, evaluate the proposal as you would a normal Introduction and Method section, and consider whether the analysis plan makes sense as the complete basis for a Results section. As with replications, the possibility of null results means that sensitivity of the methodology is especially important.

After data are collected and analyzed according to the plan, the editor may ask you to assess the report at Stage 2, now with Results and Discussion sections based on the data. At this stage, evaluate whether the research conformed to the plan, whether any changes from the proposal were well-justified, and whether other conditions for validity were met (e.g., avoiding floor and ceiling effects; passing manipulation checks, being accurately and clearly reported). If the answer is yes, then the manuscript should ultimately be accepted, although revisions might be required to improve readability or to modify the conclusions.

Conclusion

Serving as a peer reviewer provides opportunities to learn about your academic field, to become known and respected (at least to editors), and, most importantly, to shift norms and shape the future of the field. As best practices in research evolve, so too will best practices in peer review. To contribute to psychology's renaissance (Nelson et al., 2018) and credibility revolution (Vazire, 2018), peer reviewers should promote the good practices of transparency, validity, robustness, and intellectual humility. We hope that these concrete guidelines can help peer reviewers at all career stages provide more effective reviews, improving the trustworthiness of the published literature and scientific progress as a whole.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. Washington, DC: American Psychological Association. Retrieved from https://www.apa.org/ethics/code/ethics-code-2017.pdf

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*(11), 1547-1562. doi:10.1177/0956797617723724

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Borsboom, D., & Mellenbergh, G. J. (2004). The concept of validity. Psychological Review, 111, 1061–1071.

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. https://doi.org/10.1016/j.jesp.2013.10.005

Brown, N. J. L. & Heathers, J. A. J. (2016). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science, 8*, 363-369.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, *14*(1), 119–135. https://doi.org/10.1017/S0140525X00065675

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal Of Abnormal And Social Psychology*, *65*(3), 145-153. doi:10.1037/h0045186

COPE Council. (2017, September). COPE ethical guidelines for peer reviewers. Retrieved January 31, 2018, from https://publicationethics.org/files/Ethical_Guidelines_For_Peer_Reviewers_2.pdf

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Cumming, G., Calin-Jageman, R. (2017). *Introduction to the New Statistics*. New York: Routledge.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Behavioral and Brain Sciences*, *38*, e130. https://doi.org/10.1017/S0140525X14000430

Epskamp, S. & Nuijten, M. B. (2016). statcheck: Extract statistics from articles and recompute *p* values. Retrieved from http://CRAN.R-project.org/package=statcheck. (R package version 1.2.2)

Flake, J., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370-378.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review, 21*(5), 1180–1187. https://doi.org/10.3758/s13423-014-0601-x

Gelman, A. (2016, January 26). The time-reversal heuristic - a new way to think about a published finding that is followed up by a large, preregistered replication (in context of Amy Cuddy's claims about power pose). Retrieved January 31, 2018, from http://andrewgelman.com/2016/01/26/more-power-posing/

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Giner-Sorolla, R., van Kleef, G., & Amodio, D. (in press). Three strong moves to improve research and replications alike: Commentary on Zwaan et al. *Behavioral and Brain Sciences*.

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*(10), 535–549. https://doi.org/10.1111/spc3.12267

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121,* 200–206.

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, e1935. https://doi.org/10.7717/peerj.1935

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hoffman, M. B. (2014). *The Punisher's brain: The evolution of judge and jury*. Cambridge, UK: Cambridge University Press.

Howell, D. C. (2010). *Statistical methods for psychology* (7th edition). Belmontl, CA: Wadsworth.

Inzlicht, M. (2015, November). Guest post: A tale of two papers. Retrieved January 31, 2018, from http://sometimesimwrong.typepad.com/wrong/2015/11/guest-post-a-tale-of-two-papers.html

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

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5). https://doi.org/10.1097/EDE.0b013e31818131e7

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

Kidd, R. F. (1976). Manipulation checks: Advantage or disadvantage? *Representative Research in Social Psychology*, *7*(2), 160-165.

Krueger & Heck (in press). Putting the p-value in its place. *The American Statistician.*

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*(7), 701-710.

Lakens, D. (2014, May 29). The probability of *p* values as a function of the statistical power of a test. Retrieved January 31, 2018, from http://daniellakens.blogspot.ca/2014/05/the-probability-of-p-values-as-function.html

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. European Journal of Social Psychology, 44(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. (In press). Justify your alpha. *Nature Human Behaviour*. https://doi.org/10.17605/OSF.IO/9S3Y6

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*(4), 371–379. https://doi.org/10.1037/a0025172

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016, November 30). Research Preregistration 101. *APS Observer, 29*(10). Retrieved from https://www.psychologicalscience.org/observer/research-preregistration-101

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584-585.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology, 65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537-563. DOI: 10.1146/annurev.psych.59.103006.093735

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., … Zwaan, R. A. (2016). The peer reviewers' openness initiative: Incentivizing open research practices through peer review. *Royal Society Open Science, 3*(1), 150547. https://doi.org/10.1098/rsos.150547

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, *14*(9), 1105 -1107.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1422-1425.

Nosek, B. A., & Lindsay, D. S. (2018, March). Preregistration becoming the norm in psychological science. *APS Observer*.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods, 48(4),* 1205-1226. DOI: 10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Peer reviewers' openness initiative. (2014, September 13). Retrieved January 31, 2018, from https://opennessinitiative.org/the-initiative/

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, *5*(2), 187–195. https://doi.org/10.1017/S0140525X00011183

Psychological Science (2015). Retraction of "Sadness impairs color perception." *Psychological Science, 26(11),* doi: 10.1177/0956797615597672

Roediger, H. L. (2007, April 1). Twelve tips for reviewers. *APS Observer*, *20*(4). Retrieved from https://www.psychologicalscience.org/observer/twelve-tips-for-reviewers

Ross-Hellauer, T. (2017). What is open peer review? A systematic review [version 2; referees: 4 approved]. *F1000Research*, *6*(588). https://doi.org/10.12688/f1000research.11369.2

Samuelson, P. L., Jarvinen, M. J., Paulus, T. B., Church, I. M., Hardy, S. A., & Barrett, J. L. (2015). Implicit theories of intellectual virtues and vices: A focus on intellectual humility. *The Journal of Positive Psychology*, *10*(5), 389–406. https://doi.org/10.1080/17439760.2014.967802

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. https://doi.org/10.1037/a0029487

Schönbrodt, F. D., Wagenmakers, E-J., Zehetleitner, M., Perugini, M. (2017) Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322-339. https://doi.org/10.1037/met0000061

Schönbrodt, F. D., Maier, M., Heene, M., & Zehetleitner, M. (2015). Commitment to research transparency. Retrieved January 31, 2018, from http://www.researchtransparency.org

Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary?. *Personality and Social Psychology Review*, *2*(3), 218-226.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach. *Psychometrika, 74*, 107–120. doi:10.1007/s11336-008-9101-0

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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012, October 14). A 21-word solution. *Dialogue*, *26*(2). Retrieved from https://papers.ssrn.com/abstract=2160588

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*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. http://dx.doi.org/10.2139/ssrn.2694998

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.

Tennant, J., Dugan, J., Graziotin, D., Jacques, D., Waldner, F., Mietchen, D., … Colomb, J. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review [version 3; referees: 2 approved]. *F1000Research*, *6*(1151). https://doi.org/10.12688/f1000research.12037.3

Tesser, A., & Martin, L. (2006). Reviewing empirical submissions to journals. In R. J. Sternberg (Ed.), *Reviewing scientific works in psychology.* (pp. 3–29). Washington, DC: American Psychological Association. https://doi.org/10.1037/11288-001

van't Veer, A., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*(1). https://doi.org/10.1525/collabra.74

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13*(4), 411-417.

Veldkamp, C. (2017). The human fallibility of scientists: dealing with error and bias in academic research. Retrieved from: https://psyarxiv.com/g8cjq/

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2017). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-017-1323-7

Walker, R., & Rocha da Silva, P. (2015). Emerging trends in peer review—a survey. *Frontiers in Neuroscience, 9*, 169. https://doi.org/10.3389/fnins.2015.00169

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Westfall, J. (2016). PANGEA: Power ANalysis for GEneral ANOVA designs. Unpublished manuscript retrieved from https://pdfs.semanticscholar.org/ca52/e5d4976713ecdd62fa10a501d0bf094a30a2.pdf

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

Zöllner, S., & Pritchard, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *The American Journal of Human Genetics, 80*(4), 605–615. https://doi.org/10.1086/512821

Appendix A

Outline of Advice for Promoting Robustness and Transparency When Reviewing Quantitative Empirical Research Manuscripts in Psychology

- **Preparing to review**
  - Understand *p*-values and power.
  - Know the importance of specifying predictions ahead of time.
  - Know assumptions underlying frequently used statistical tests.
  - If you don't know much about some of the techniques used by the authors, acknowledge it to the editor.
  - Consult the journal's statistical and reporting standards before you review.

- **Statistical reporting elements to look for or request**
  - A priori or sensitivity power analysis (post-hoc analyses are not of much use).
  - Whether decisions such as analyses, exclusions, and transformations were determined a priori or post hoc. For post hoc analyses, more evidence (e.g., replication) may be needed.
  - Whether optional stopping in data collection was used, and if yes, how it was corrected for.
  - Descriptive statistics such as means, standard deviations, and correlations.
  - A methodological disclosure statement verifying that the article reports the existence of all measures, manipulations and exclusions in the study.
  - Keep an eye out for errors in reporting, such as wrong degrees of freedom, or incorrect inferential statistics (e.g., using Statcheck).

- **Dealing with data, materials, and preregistrations**
  - Check the availability of any preregistrations, data, materials, and analysis code.
  - Optionally, examine the completeness and accuracy of data, materials, and analysis code.
  - Optionally, examine the specificity and completeness of any preregistrations.
  - Tell the editor how far you went in checking these materials.

- **Evaluating statistical outcomes**
  - Assess the quality of evidence without relying on $p < .05$ per study as either necessary or sufficient for drawing a positive conclusion.
  - Under complete and transparent reporting, multiple studies all showing *p*-values close to .05 are uncommon; assess accordingly.
  - If you aren't sure about replicability of results, consider a request for a pre-registered additional study, or more transparent reporting of existing studies.
  - Claims of null effects should be evaluated as carefully as positive effects, e.g., with Bayesian or NHST equivalence testing.

- **Assessing constraints on generality**
  - Consider asking for a statement on what aspects of the study authors believe are essential to observing the effect.

- **Promoting transparency**

  - If the journal does not require sharing data, materials, or analysis code or does not require a statement for why they aren't shared, consider requesting them.
  - Decide whether you will sign or not sign all of your reviews.

- **Reviewing replications**
  - Use the same level of scrutiny for replications as original studies.
  - If a direct replication, do authors demonstrate similarity/discrepancy between the studies? Is discrepancy needed to reproduce the psychological variables in a new context?
  - Factors to consider if examining "need" for a replication: strength of existing evidence, the effect's theoretical importance or potential value to society, and the original research's prior impacts on other research and society.
  - Don't be too concerned with assessing success or failure of the replication; think meta-analytically about what the sum of all results says about an effect.

- **Registered Reports**
  - Evaluate proposal's Introduction and Method sections as usual.
  - Assess whether the analysis plan covers all of the important details and can serve as the complete basis for a Results section.
  - In final report, assess whether method and analysis plans were reported, and assess the rationale for any deviations.

Appendix B

Resources on Robustness and Transparency in Psychological Research

This list is intended to be a useful starting point for reviewers seeking to improve their understanding of the methodological and statistical underlying psychology's credibility revolution. We recognize that there are many more references and resources out there; we do not claim that this list is comprehensive nor that the pieces included represent the "gold standard" among all possible resources.

Open Science

Center for Open Science. https://cos.io/

> The Center for Open Science provides tools, training, support, and advocacy for encouraging open scientific practices. Their website contains more background on the goals of open science, as well as various services and training opportunities that reviewers can take advantage of to stay up to date with the latest developments.

Open Science Framework. https://osf.io/

> The Open Science Framework (OSF) provides a public repository for researchers to share their materials, data, and analysis scripts. As a reviewer, you can ask authors to consider making the basis of their scientific claims available through the OSF or another public repository.

Transparency and Openness Promotion (TOP) guidelines. https://cos.io/our-services/top-guidelines/

> Eight guidelines (e.g., regarding data transparency) crafted by a group led by Brian Nosek of the Center for Open Science and initially described in an article published in *Science* in 2015). To date the guidelines have been implemented (at varying levels of stringency) by 850 journals. Find out if the journal for which you are reviewing has endorsed the TOP guidelines.

Statistical Power

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

> This is a classic paper that provides background education on the rationale for power analysis, and the sample sizes required to detect "small", "medium", and "large" effect sizes with 80% power for the simplest analyses.

Magnusson, K. (2018). Understanding statistical power and significance testing: An interactive visualization [Web app]. Retrieved from http://rpsychologist.com/d3/NHST/

> Reviewers can use this brief primer (with an interactive visualization) to refine their understanding of how power, Type I and Type II errors, effect size, sample size, and alpha are related to each other.

Champely, S. (2018). pwr: Basic Functions for Power Analysis [R package]. Retrieved from https://CRAN.R-project.org/package=pwr

This R package provides power analysis functions that reviewers may want to use to assess the statistical power of the reported analyses, and to encourage authors to comment on these issues. The quick-start guide is available at: https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*, 75-191. https://doi.org/10.3758/bf03193146

For reviewers who are not familiar with R, G*Power 3 is another free program with a point-and-click interface that can be used to conduct a range of power analyses during peer review.

Anderson, S.F., Kelley, K., & Maxwell, S.E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, 28, 1547-1562 doi: 10.1177/0956797617723724

This article provides a readable summary of basic concepts of statistical power (similar to other treatments) but it goes beyond them by offering a way to take both publication bias and estimate uncertainty into account when planning sample size. Useful for evaluating sample size justifications, especially for replications. There is an associated shiny app at https://designingexperiments.com/shiny-r-web-apps/.

Westfall, J. (2018). Power Analysis for GEneral Anova designs (PANGEA) [Web app]. Retrieved from https://jakewestfall.org/pangea/

This power analysis program provides power calculations for general ANOVA designs, and can flexibly handle designs with any number of fixed or random factors, each with any number of levels, and with any valid pattern of nesting or crossing of the factors. You might suggest this for authors in need of power analysis resources.

Cumming, G. (2014). The New Statistics: *Why and How*. *Psychological Science, 25*, 7-29. https://doi.org/10.1177/0956797613504966

Cummings avoids the term “power” because he believes that psychologists should abandon NHST in favour of on emphasis on precision of effect size estimates. But his book is very engaging and compelling in explaining why *p* values are themselves very unreliable. Background reading for reviewers.

Effect Sizes

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*, 7-29. https://doi.org/10.1177/0956797613504966

Encourages researchers to move beyond a focus on statistical significance, to an emphasis on effect sizes and confidence intervals. Reviewers may find this article useful for enhancing their understanding of these issues, and can ask authors to provide confidence intervals and discuss effect sizes.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

A how-to for navigating the large number of power statistics applicable to designs that compare distinct groups. Can inform reviewer recommendations about power analysis.

Understanding *p* values

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997. https://doi.org/10.1037//0003-066x.49.12.997

Reviews the problems with NHST and common misunderstandings of *p* values. Reviewers can read this to refine their understanding of *p* values.

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

Despite an arguably overstated title, this paper makes a compelling case for the limitations of p-values alone, and the need to evaluate truth claims referring also to statistical power and prior probability. Background reading for reviewers to understand the logic of NHST and evidence.

Schönbrodt, F. (2014). When does a significant *p*-value indicate a true effect? Understanding the Positive Predictive Value (PPV) of a *p*-value [Web app]. Retrieved from http://shinyapps.org/apps/PPV/

Interactive demonstration of p-values' predictive value based on Ioannidis (2005)

Wasserstein, R. L., & Lazar, N. A. (2016). The American Statistical Association's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133. https://doi.org/10.1080/00031305.2016.1154108

A broad consortium of frequentists and Bayesian statisticians approved this message about the limitations of p-values, including the need for exact values, additional information, and reporting the full context of analyses. Very useful authority for reviewers to cite in support of full disclosure and a nuanced approach to significance.

Magnusson, K. (2018). Distribution of *p*-values when comparing two groups: An interactive visualization [Web app]. Retrieved from http://rpsychologist.com/d3/pdist/

Reviewers can use this interactive app to hone their intuitions about what distributions of *p* values look like under different assumptions.

Sequential Analyses

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

This how-to article argues persuasively that if appropriately reported and controlled, collecting participants in successive groups until a stopping point is reached is not cheating, but an efficient method of collecting data in the face of uncertainty about effect sizes. Reviewers can suggest this and the following two papers if it emerges authors have been sampling sequentially without error correction.

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

Similar argument to Lakens (2014), above, presenting a simple method of adjusting p-values for sequential collection.

Schönbrodt, F. D., Wagenmakers, E-J., Zehetleitner, M., Perugini, M. (2017) Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22,* 322-339. https://doi.org/10.1037/met0000061

A Bayesian approach to sequential testing, which the previous two articles approach using NHST.

Interpreting Null Results

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60,* 328–331. https://doi.org/10.1198/000313006X152649

Explains why changes in statistical significance are often not themselves statistically significant. Reviewers can read this article to become more aware of this issue, and cite it as support in reviews.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8,* 355–362. https://doi.org/10.1177/1948550617697177

Describes one way to demonstrate evidence for the null within a null hypothesis significance testing framework. Reviewers may ask authors to use equivalence tests (or Bayesian methods; see below) to provide further context for null findings.

Bayesian Approaches

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review, 25,* 5-34. https://doi.org/10.31234/osf.io/q46q3

An explanation of the probability theory underlying Bayesian analysis and some use-cases with Harry Potter-themed examples. Good preparation for evaluating Bayesian analyses, which are becoming more common in submitted manuscripts.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35-57

Ten arguments for using Bayesian analysis and rebuttals to the most commonly heard objections. A somewhat more general approach to the goal of understanding the utility and necessary parameters for Bayesian analysis.

Detecting Statistical Discrepancies

Nuijten, M., & Rife, S. (2018). statcheck [Web app]. Retrieved from http://statcheck.io/

Automatically analyzes documents for discrepancies between reported inferential statistics in text and p-values. Reviewers may wish to run papers through statcheck, either using R or using the online interface.

Pre-registration

Brief overview: Lindsay, Simons, & Lilienfeld (2017) Research preregistration 101 (with FAQs) https://www.psychologicalscience.org/observer/research-preregistration-101

Provides an accessible and brief overview (with FAQs) about preregistration. This can help introduce reviewers who are unfamiliar with preregistration to this practice.

More extensive discussion: van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology — A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2–12. doi:10.1016/j.jesp.2016.03.004

This article provides a more extensive discussion about the elements of preregistration, with a proposed standard template.

Open Science Framework resources: http://help.osf.io/m/registrations

The OSF provides resources and templates for preregistration.

AsPredicted.org

AsPredicted provides a simple framework for preregistration. Reviewers who are new to preregistration might want to consult this template to better understand the key ways in which preregistration can constrain research degrees of freedom.

Methodological Disclosure and Generalizability Statements

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012, October 14). A 21-word solution. *Dialogue*, *26*(2). Retrieved from https://papers.ssrn.com/abstract=2160588

A simple implementation of a methodological disclosure statement that allows authors to confirm, in 21 words, that they have reported how they determined their sample size, all manipulations, and all measures. Something for reviewers to ask for if the journal does not require it.

Nosek, B. A., Simonsohn, U., Moore, D. A., Nelson, L. D., Simmons, J. P., Sallans, A., & LeBel, E. P. (2018, August 13). Standard reviewer statement for disclosure of sample, conditions, measures, and exclusions. Retrieved from https://osf.io/hadz3/

A standard statement, endorsed by the Center for Open Science, that reviewers can use to request a methodological disclosure statement along the lines of the 21-word solution.

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. https://doi.org/10.1177/1745691617708630

A proposal for authors to explicitly define the scope of the conclusions that are justified by the data, but specifying the target populations (of people, situations, and stimuli) that they expect their findings to be able to replicate in. Reviewers can ask for such a statement if it seems like the authors are drawing conclusions that are broader than appear to be justified by the samples used in their paper.