Journal of Pacific Rim Psychology

www.cambridge.org/prp

Calibrate your confidence in research findings: A tutorial on improving research methods and practices

Aline da Silva Frost and Alison Ledgerwood

Department of Psychology, University of California, Davis, California, USA

Original Article

Cite this article: da Silva Frost A. and Ledgerwood A. (2020) Calibrate your confidence in research findings: A tutorial on improving research methods and practices. *Journal of Pacific Rim Psychology*, Volume 14, e14. https://doi.org/10.1017/prp.2020.7

Received: 25 November 2019
Revised: 1 March 2020
Accepted: 3 March 2020

Keywords:
statistical power; preregistration; preanalysis plan; open science; replicability; positive predictive value

Author for correspondence:
Aline da Silva Frost,
Email: asfrost@ucdavis.edu

Abstract

This article provides an accessible tutorial with concrete guidance for how to start improving research methods and practices in your lab. Following recent calls to improve research methods and practices within and beyond the borders of psychological science, resources have proliferated across book chapters, journal articles, and online media. Many researchers are interested in learning more about cutting-edge methods and practices but are unsure where to begin. In this tutorial, we describe specific tools that help researchers calibrate their confidence in a given set of findings. In Part I, we describe strategies for assessing the likely statistical power of a study, including when and how to conduct different types of power calculations, how to estimate effect sizes, and how to think about power for detecting interactions. In Part II, we provide strategies for assessing the likely type I error rate of a study, including distinguishing clearly between data-independent ("confirmatory") and data-dependent ("exploratory") analyses and thinking carefully about different forms and functions of preregistration.

In recent years, psychology has been at the forefront of a broad movement across scientific disciplines to improve the quality and rigor of research methods and practices (Begley & Ellis, 2012; Button et al., 2013; Ledgerwood, 2016; McNutt, 2014; Nosek, Spies, & Motyl, 2012; Nyhan, 2015; see Spellman, 2015, for a helpful synopsis). The field as a whole is changing: Conversations about improving research practices have become mainstream, journals and societies are adopting new standards, and resources for improving methods and practices have proliferated across journal articles, book chapters, and online resources such as blogs and social media (Simons, 2018). As attention to methodological issues has surged, researchers have become increasingly interested in understanding and implementing methodological tools that can maximize the knowledge they get from the work that they do.

At the same time, for the average researcher, it can be daunting to approach this new wealth of resources for the first time. You know that you want to understand the contours of recent developments and to learn as much as possible from the research you do, but where do you even begin? We think that one of the most important methodological skills to develop is how to calibrate your confidence in a finding to the actual strength of that finding.

In this tutorial, we seek to provide a toolbox of strategies that can help you do just that. If a finding is strong, you want to have a relatively high level of confidence in it. In contrast, if a finding is weak, you want to be more skeptical or tentative in your conclusions. Having too much confidence in a finding can lead you to waste resources chasing and trying to build on an effect that turns out to have been a false positive, thereby missing opportunities to discover other true effects. Likewise, having too little confidence in a finding can lead you to miss opportunities to build on solid and potentially important effects. Thus, in order to maximize what we learn from the work that we do as scientists, we want to have a good sense of how much we learn from a given finding.

We divide this tutorial into two main parts. The first part will focus on how to estimate statistical power, which refers to the likelihood that a statistical test will correctly detect a true effect if it exists (i.e., the likelihood that if you are testing a real effect, your test statistic will be significant). The second part will focus on type I error, which refers to the likelihood that a statistical test will incorrectly detect a null effect (i.e., the likelihood that if you are testing a null effect, your test statistic will be significant). Arguably, one of the central problems giving rise to the field's so-called "replicability crisis" is that researchers have not been especially skilled at assessing either the statistical power or the Type I error rate of a given study – leading them to be overly confident in the evidential value and replicability of significant results (see Anderson, Kelley, & Maxwell, 2017; Lakens & Evers, 2014; Ledgerwood, 2018; Nosek, Ebersole, DeHaven, & Mellor, 2018a; Open Science Collaboration, 2015; Spellman, Gilbert, & Corker, 2017; Pashler & Wagenmakers, 2012). For example, Bakker, van Dijk, and Wicherts (2012) estimated the average statistical power in psychological experiments to be only 35%, and even large studies may have lower statistical power than researchers intuitively expect when measures are

© The Author(s) 2020. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.

CAMBRIDGE UNIVERSITY PRESS

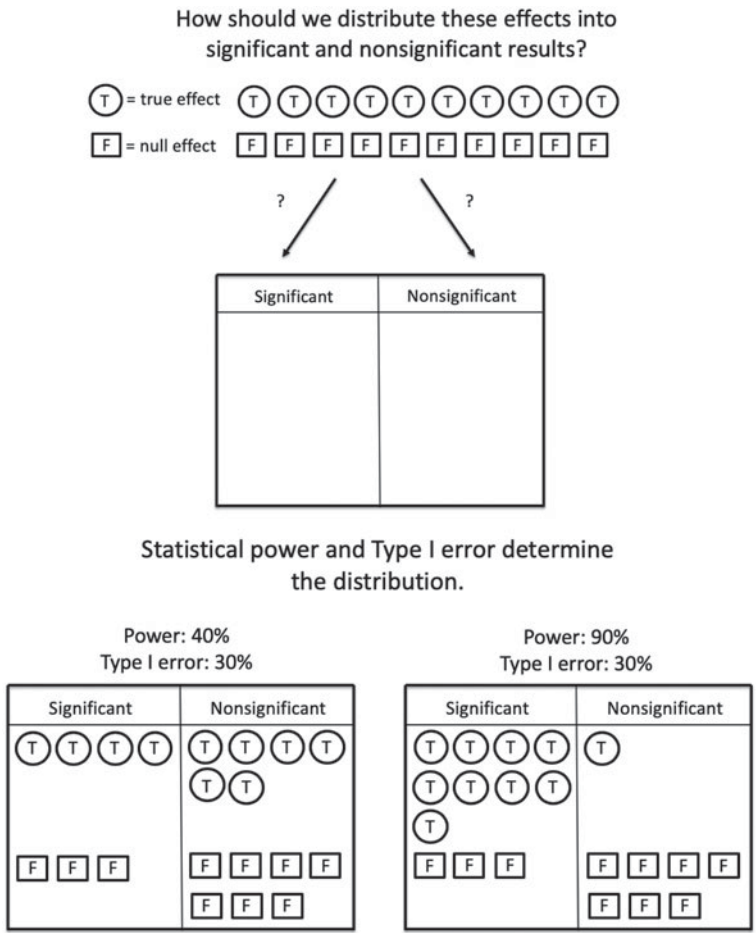


Figure 1. Consider the case of a researcher testing 10 true effects and 10 false effects. Perhaps they will follow up or publish significant results but leave nonsignificant results in a file drawer. The statistical power and type I error rate of the studies will determine how the effects are sorted into a set of significant results (follow up!) and a set of nonsignificant results (file drawer). Notice that because power is higher in the scenario on the right (vs. left), the likelihood that any one of the significant findings reflects a true effect in the population is also higher.

not highly reliable (see Kanyongo, Brook, Kyei-Blankson, & Gocmen, 2007; Wang & Rhemtulla, in press). Meanwhile, common research practices can inflate the type I error rate of a statistical test far above the nominal alpha (typically $p < .05$) selected by a researcher (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Wang & Eastwick, in press).

Importantly, you need a good estimate of both quantities – statistical power and type I error – in order to successfully calibrate your confidence in a given finding. That is because both quantities influence the *positive predictive value* of a finding, or how likely it is that a significant result reflects a true effect in the population.[1]

For example, imagine that in the course of a typical year, a researcher has ten ideas that happen to be correct and ten ideas that happen to be incorrect (that is, she tests ten effects that are in fact true effects in the population and ten that are not). Let us focus on what happens to the correct ideas first. As illustrated in Figure 1, the statistical power of the researcher's studies determines how many of these true effects will be detected as significant. If the studies are powered at 40% (left side of Figure 1), four out of ten studies will correctly detect a significant effect, and six out of ten studies will fail to detect the effect that is in fact present in the population. If the studies are powered at 90% (right side of Figure 1), nine out of ten studies will correctly detect the significant effect, and only one will miss it and be placed in the file drawer.

However, statistical power is only part of the story. Not all ideas are correct, and so let us focus now on the ten ideas that happen to be incorrect (that is, she tests ten effects that are in fact null effects in the population). As illustrated in Figure 1, the type I error rate of the researcher's studies determines how many of these null effects will be erroneously detected. If the studies have a type I error rate of 30%, three of the ten null effects will be erroneously detected as significant, and the other seven will be correctly identified as nonsignificant.[2]

The researcher, of course, does not know whether the effects are real or null in the population; she only sees the results of her statistical tests. Thus, what she really cares about is how likely she is to be right when she reaches into her pile of significant results and declares: "This is a real effect!" In other words, if she publishes or devotes resources to following up on one of her significant effects, how likely is it to be a correctly detected true effect, rather than a false positive? Notice that the answer to this question about the positive predictive value of a study depends on both statistical power and type I error rate. In Figure 1, the positive predictive value of a study is relatively low when power is low (on the left side): The likelihood that a significant result in this pool of significant results reflects a true effect is 4 out of 7 or 57%. In contrast, the positive predictive value of a study is higher when power is higher (on the right side of Figure 1): The likelihood that a significant result in this pool of significant results reflects a true effect is 9 out of 12, or 75%. Thus, if we are to understand how much to trust a significant result, we want to be able to gauge *both* the likely statistical power and the likely type I error rate of the study in question.

At this point, readers may wonder about the trade-off between statistical power and type I error, given that the two are related. For example, one way to increase power is to set a higher alpha threshold for significance testing (e.g., $p < .10$ instead $< .05$), but this

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press

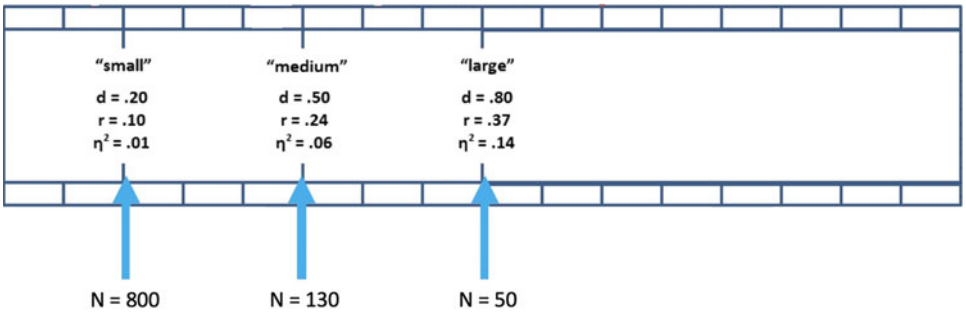


Figure 2. Sample sizes needed to achieve 80% power in a two-condition, between-subjects study. This figure helps you organize visually the effect size intuitions.

practice will also increase type I error. However, there are other possible ways to increase power that do not affect the type I error rate (e.g., increasing sample size, improving the reliability of a measure) – and it is these strategies that we discuss below. More broadly, in this tutorial, we focus on providing tools to assess: (1) the likely statistical power; and (2) the likely type I error rate of a study result, and offer guidance for how to increase statistical power or constrain type I error for researchers who want to be able to have more confidence in a given result.

Part I: How to Assess Statistical Power

Develop Good Intuitions About Effect Sizes and Sample Sizes

One simple but useful tool for gauging the likely statistical power of a study is a well- developed sense of the approximate sample size required to detect various effects. Think of this as building your own internal power calculator that provides rough, approximate estimations.

You want to be able to glance at a study and think to yourself: "Hmm, that's a very small sample size for studying this type of effect with this sort of design – I will be cautious about placing too much confidence in this significant result" or "This study is likely to be very highly powered – I will be relatively confident in this significant result." In other words, it is useful to develop your intuitions for assessing whether you are more likely to be in a world that looks more like the left side of Figure 1 or in a world that looks more like the right side.

How do you build this internal calculator? You can start by memorizing some simple benchmarks. For a simple two-condition, between-subjects study, the sample size required to detect a medium effect size of $d = .50$ with 80% power is about $N = 130$ (65 participants per condition; see Figure 2. Notice that $d = .50$ is equivalent to $r = .24$ and $h^2 = .06$). To detect a large effect of $d = .80$ with 80% power requires about $N = 50$ (25 per condition). And to detect a small effect size of $d = .20$ with 80% requires about $N = 800$ (400 per condition; Faul, Erdfelder, Lang, & Buchner, 2007).

Next, start developing your sense of how big such effects really are. Cohen (1992) set a medium effect size at $d = .50$ to "represent an effect likely to be visible to the naked eye of a careful observer" (p. 156), so a medium-sized effect is one that we might observe simply by watching people closely. A large effect size of $d = .80$ is typically an effect that even a casual observer would notice (e.g., the correlation between relationships satisfaction and breakup is approximately this magnitude; Le, Dove, Agnew, Korn, & Mutso, 2010). And a small effect size of $d = .20$ is typically too small to be seen with the naked eye (e.g., if you're interested in testing a counterintuitive prediction that would be surprising to most people, it is likely to be a small effect if it is true). Pay attention to effect size estimates from meta-analyses and very large studies in your area of research to hone your intuitions within your particular research area.

Finally, keep in mind that interactions can require much larger sample sizes, depending on their shape. For example, imagine that you are interested in powering a two-group study to detect a medium-sized effect. You conduct Study 1 with a total sample size of $N = 130$ and find that indeed, your manipulation (let's call it Factor A) significantly influences your dependent measure. Next, you want to know whether Factor B moderates this effect. How many participants do you need to have 80% power to detect an interaction in Study 2, where you manipulate both Factor A and Factor B in a 2×2 design?

As Table 1 illustrates, the answer depends on the shape of the interaction you expect (see Giner-Sorolla, 2018; Ledgerwood, 2019). If you expect a knockout interaction (i.e., you think the effect you saw in Study 1 will appear in one condition of Factor B and disappear in the other), it turns out you need to quadruple the sample size you had in Study 1 to have 80% power to detect the interaction in Study 2. If you expect a perfect cross-over interaction (i.e., you think the effect you saw in Study 1 will appear in one condition of Factor B and reverse completely in the other), you need the *same* sample size you had in Study 1 to have 80% power to detect the interaction in Study 2 (although you will probably want to double the sample size to provide 80% power to detect each of the simple main effects). And if you expect a 50% attenuation (i.e., you think the effect you saw in Study 1 will appear in one condition of Factor B and be reduced to half its original size in the other), you need about *14 times* the sample size you used in Study 1.

Strategies for a Planned Study

Conduct an a priori power analysis

The rules of thumb described above can be useful, but when you are planning your own study, you can conduct a formal, a priori power analysis to decide how many participants you need in order to achieve your desired level of power. In an a priori power analysis, you input your desired level of power (e.g., 80%), your planned statistical test (e.g., a t test comparing two between-subjects conditions), and your estimated effect size (e.g., $d = .40$), and the program tells you the necessary sample size (e.g., $N = 200$). The central challenge in this kind of power analysis is to identify a good estimate of the expected effect size.

Getting a good estimate of the expected effect size can be tricky for multiple reasons. First, effect size estimates (like any estimate) will fluctuate from one study to the next, especially when sample sizes are smaller (see Ledgerwood, Soderberg, & Sparks, 2017, Figure 1, for an illustration). In other words, an effect size estimate from any given study can underestimate or overestimate the true size of an effect. Second, publication bias tends to inflate the effect sizes reported in a given literature. Historically, significant results

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press

Table 1. Rules of thumb for powering a 2 × 2 between-subjects Study 2 that seeks to moderate a main effect observed in Study 1

Expected shape of interaction	Required total sample size for Study 2	Example: *N* needed if Study 1 tested an effect of $d = .50$
Cross-over	2 × *N* in Study 1[a]	$N \approx 260$
Knockout	4 × *N* in Study 1	$N \approx 520$
50% attenuation	14 × *N* in Study 1	$N \approx 1820$

Note: In this illustration, Study 1 is powered to provide 80% power to detect a main effect. In Study 2, a researcher wants to test whether the effect observed in Study 1 is moderated by a second variable in a 2 × 2 between-subjects factorial design.

were (and continue to be) more likely to be published, and null effects are more likely to be shuttled to a file drawer rather than shared with the scientific community (Anderson et al., 2017; Ledgerwood, 2019; Rosenthal, 1979). Because any given study can underestimate or overestimate an effect size, and because overestimates are more likely to hit significance, publication bias effectively erases a sizable portion of underestimates from the literature. Therefore, a published effect size estimate is more likely to be an overestimate than an underestimate. And, when one averages the effect size estimates that do make it into the published literature, that average is usually too high (Anderson et al., 2017).

To get around these issues, we have two options: a large study or a meta-analysis. In both cases, it is important to consider publication bias. The first option is to find an estimate of a similar effect from a large study (e.g., a total sample size of approximately $N = 250$ or larger for estimating a correlation between two variables or a mean difference between two groups; Schönbrodt & Perugini, 2013). If you find such a study, ask yourself whether the paper would have been published if it had different results (e.g., a paper that describes its goal as estimating the *size* of an effect may be less affected by publication bias than a paper that describes its goal as demonstrating the *existence* of an effect). The second option is to find an estimate of a similar effect from a meta-analysis. Meta-analyses aggregate results of multiple studies, so their estimates ought to be more accurate than an estimate from a single study. However, because they also sample studies from the literature, they can overestimate the size of an effect. For this reason, look for meta-analysis that carefully model publication bias (e.g., by using sensitivity analyses that employ various models of publication bias to produce a range of possible effect size estimates; McShane, Bockenholt, & Hansen, 2016; see Ledgerwood, 2019, for a fuller discussion).

By identifying a good estimate of the expected effect size, such as those from large studies and meta-analyses that account for publication bias, we can conduct a priori power analyses that will be reasonably accurate. You can conduct an a priori power calculation using many simple-to-use programs, such as G*Power (Faul et al., 2007); PANGEA (Westfall, 2016) for general analysis of variance (ANOVA) designs and pwrSEM (Wang & Rhemtulla, in press) for structural equation models.

Of course, sometimes it is not possible to identify a good effect size estimate from a large study or meta-analysis that accounts for publication bias. If your only effect size estimate is likely to be inaccurate and/or biased (e.g., an effect size estimate from a smaller study), you can use a program that accounts for uncertainty and bias in effect size estimates (available online at https://designingexperiments.com/shiny-r-web-apps under the penultimate heading, "Bias and Uncertainty Corrected Sample Size for Power" or as an R package; see Anderson et al., 2017).

Identify the biggest sample size worth collecting

In other situations, you are simply not sure what effect size to expect – perhaps you are starting a brand-new line of research, or perhaps the previous studies in the literature are simply too small to provide useful information about effect sizes. A useful option in such cases is to identify the smallest effect size of interest (often abbreviated to SESOI) and use that effect size in your power calculations (Lakens, 2014). In other words, if you only care about the effect if it is at least medium in size, you can power your study to detect an effect of $d = .50$.

Basic researchers often feel reluctant to identify a SESOI, because they often care about the direction of an effect regardless of its size (e.g., competing theories might predict that a given manipulation will increase or decrease levels on a given dependent measure, and a basic researcher might be interested in either result regardless of the effect size). However, with a minor tweak, the basic concept becomes useful to everyone. If identifying a SESOI feels difficult, identify instead the largest sample you would be willing to collect to study this effect.

Let's call this the "Biggest Sample Size Worth Collecting", or BSSWC. For example, if you decide that a given research question is worth the resources it would take to conduct a two-group experiment with a total of $N = 100$ participants, you are effectively deciding that you are only interested in the effect if it is at least $d = .56$ (the effect size that $N = 100$ would provide about 80% power to detect). Notice, then, that a BSSWC of 100 participants is equivalent to a SESOI of $d = .56$ – they are simply two different ways of thinking about the same basic idea. Notice, too, that it is worth making the connection between these two concepts explicit. For example, a social psychologist might consider whether the effect they are interested in studying is larger or smaller than the average effect studied in social-personality psychology ($r = .21$ or $d = .43$; see Richard, Bond, & Stokes-Zoota, 2003). Unless they have a reason to suspect their effect is much larger than average, they may not want to study it with a sample size of only $N = 100$ (because they will be under-powered; see Figure 1).

Conserve resources when possible: Sequential analyses

Once you have determined the maximum sample size you are willing to collect (either through an a priori power analysis or by determining the maximum resources you are willing to spend on a given study), you can conserve resources by using a technique called sequential analysis. Sequential analyses allow you to select a priori the largest sample size you are willing to collect if necessary as well as middle points where you would stop data collection earlier if you could (Lakens & Evers, 2014; Proschan, Lan, & Wittes, 2006). For example, if you have a wide range of plausible effect size estimates and are unsure about how many participants to run, but you know you are willing to collect a total sample size of $N = 600$ to detect this particular effect, a sequential analysis may be the best option. Sequential analyses are planned ahead of time, before looking at your results, and preserve a maximum type I error rate of 5%. In contrast, if you check your data multiple times *without* using a formal sequential analysis, type I error rates inflate (see Sagarin, Ambler, & Lee, 2014; Simmons et al., 2011).

To conduct a sequential analysis, you would first decide how many times you will want to check your data before reaching your final sample size – in our example, $N = 600$. Each additional check will reduce power by a small amount in exchange for the possibility of stopping early and conserving resources. Imagine that you decide to divide your planned sample into three equal parts, so that you conduct your analysis at $n = 200$, $n = 400$ and $N = 600$

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press

Table 2. Alpha thresholds for sequential analyses

Divide your sample size into equal parts	Stop at percent of total *N*	Example (total *N* = 600)	Alpha threshold	Decision guide	
2	50%	300	.025	$p < .025$?	if yes, significant.
					if no, continue collection
	100%	600	.034	$p < .034$?	if yes, significant
					if no, it is not significant
3	33%	200	.017	$p < .017$?	if yes, significant.
					if no, continue collection
	66%	400	.022	$p < .022$?	if yes, significant
					if no, continue collection
	100%	600	.028	$p < .028$?	if yes, significant
					if no, it is not significant
4	25%	150	.013	$p < .013$?	if yes, significant
					if no, continue collection
	50%	300	.016	$p < .016$?	if yes, significant
					if no, continue collection
	75%	450	.020	$p < .020$?	if yes, significant
					if no, continue collection
	100%	600	.025	$p < .025$?	if yes, significant
					if no, it is not significant

Note: Once you have planned your total sample size and how many times you will want to stop and check the results, use this table to determine the alpha cut-off thresholds you will use to determine significance at each planned analysis time point.

participants (you can follow this example in Table 2). You would then pause data collection at each of these points and check the results. If the *p* value of the analysis is less than a predetermined alpha threshold (see Table 2), you would determine that the test is statistically significant. If it is greater than the threshold, you would continue collecting data up until the final sample size of $N = 600$.

For instance, imagine that you collect the first planned set of 200 participants, or 33% of the total *N*. You pause data collection and analyze the data; if the *p* value of the focal analysis is below the first alpha threshold of .017, the result is significant and you can stop data collection early (saving 400 participants). If the *p* value of the focal analysis is not below .017, you continue to collect data from 200 more participants. When you check the results again, if the *p* value is below the second alphas threshold of .022, the result is significant and you can stop data collection early (saving 200 participants). If the *p* value is not below .022, you continue to collect data from 200 more participants and then conduct the focal analysis one final time on the total sample of 600 participants. If the *p* value is below .028, the result is significant. If it is not below .028, the result is not significant. In either case, you stop collecting data because you have reached your final planned sample size.

Importantly, by computing specific, adjusted alpha thresholds depending on the planned number of stopping points, sequential analysis enables researchers to check the data multiple times during data collection while holding the final type I error rate at a maximum of .05. This allows you to balance the goals of maximizing power and conserving resources. Table 2 provides the alpha thresholds for common sequential analyses where a researcher wants to divide their total planned sample (of any size) into 2–4 equal parts and hold their type I error rate at .05 or below (for an example of how to write up a sequential analysis, see Sparks & Ledgerwood, 2017). If you want to stop more frequently or at unevenly spaced points, you can use the GroupSeq R package and step-by-step guide provided by Lakens (2014; resources available at https://osf.io/qtufw/) to compute other alpha thresholds.

Consider multiple approaches to boosting power

After conducting an a priori power analysis, you may find that to have your desired level of power, you need a larger sample size than you initially imagined. However, it is not always possible or practical to collect large sample sizes. You may be limited by the number of participants available to you (especially when studying hard-to-reach populations) or by the finite money and personnel hours that you have to spend on collecting data. Whatever the situation, all researchers face trade-offs and constraints based on resources.

Given such constraints, it is often useful to consider multiple approaches to boosting the power of a planned study. When possible and appropriate, making a manipulation within-subjects instead of between-subjects can dramatically boost the power of an experiment (see Greenwald, 1976; Rivers & Sherman, 2018). Likewise, you can increase power by strengthening a manipulation and by improving the reliability of your measures (see e.g., Ledgerwood & Shrout, 2011). In addition, it is sometimes possible to select ahead of time a planned covariate that correlates strongly with the dependent measure of interest (e.g., measuring extraversion as a covariate for a study that examines self-esteem as the focal dependent variable; see Wang, Sparks, Gonzales, Hess, & Ledgerwood, 2017). Finally, one of the most exciting developments

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press

in the "cooperative revolution" created by the open science movement is the proliferation of opportunities for large-scale collaborations (e.g., the Psychological Science Accelerator, ManyLabs, ManyBabies, and StudySwap; see Chartier, Kline, McCarthy, Nuijten, Dunleavy, & Ledgerwood, 2018). When it simply is not feasible to study the research question you want to study with high statistical power, consider collaborating across multiple labs and aggregating the results.

Strategies for an Existing Study

Conduct a sensitivity analysis

When you want to assess the statistical power of an existing study (e.g., a study published in the literature or a dataset you have already collected), you can conduct a type of power analysis called a *sensitivity analysis* (Cohen, 1988; Erdfelder, Faul & Buchner, 2005). In a sensitivity analysis, you input the actual sample size used in the study of interest (e.g., $N = 60$), the statistical test (e.g., a t test comparing two between-subjects conditions), and a given level of power (e.g., 80%), and the program tells you the effect size the study could detect with this level of power (e.g., $d = .74$). The central goal for this kind of power analysis is to provide a good sense of the range of effect sizes that an existing study was adequately powered to detect.

For example, perhaps you have already conducted a study in which you simply collected as many participants as resources permitted, and you ended up with a total sample of $N = 164$ participants in a two-group experimental design. You could conduct sensitivity analyses to determine that your study had 60% power to detect an effect size of $d = .35$ and 90% power to detect an effect of $d = .51$. Armed with the effect size intuitions we discussed in an earlier section, you could then ask yourself whether the effect size you are studying is likely to be on the smaller side or on the larger side (e.g., is it an effect that a careful observer could detect with the naked eye?). By thinking carefully about this information, you can gauge the likely statistical power of your study (e.g., is it more like the left side or the right side of Figure 1?) and calibrate your confidence in the statistical result accordingly.

Don't calculate "post-hoc" or "observed" power

Many types of power analysis software also provide an option for computing power called *post hoc power* or *observed power*. This type of power analysis is highly misleading and should be avoided (Gelman & Carlin, 2014; Hoenig & Heisey, 2001). In a post hoc power analysis, you input the effect size estimate from a study as if it is the true effect size in the population. However, as discussed earlier, a single study provides only one estimate of the true population effect size, and this estimate tends to be highly noisy: It can easily be far too high or far too low (see Figure 1 in Ledgerwood et al., 2017; Schönbrodt & Perugini, 2013). Furthermore, because researchers tend to be more interested in following up on and publishing significant results, and because a study is more likely to hit significance when it overestimates (vs. underestimates) an effect size, researchers are especially likely to conduct post hoc power analyses with overestimated effect size estimates.

The result of using post-hoc power is an illusion of a precise power estimate that in fact is (1) highly imprecise and (2) redundant with the p value of the study in question. In other words, post-hoc power or observed power appears to provide a new piece of very precise information about a study, when in fact it provides an already known piece of imprecise information. It is loosely akin to attempting to gauge the likelihood that a coin flip will result in "heads" rather than "tails" based on flipping the coin, observing that the result is "heads," and then deciding based on this observation that the coin flip must have been very likely to result in the outcome you saw. Thus, post-hoc power or observed power ultimately worsens your ability to calibrate your confidence to the strength of a result.

Part II: How to Assess Type I Error Rates

As Figure 1 illustrates, if we want to correctly calibrate our confidence in a significant result, we want to be able to gauge not only the likely statistical power of the test in question, but also its likely type I error rate. Researchers often assume that their type I error rate is simply set by the alpha cut-off against which a p value is compared (traditionally, $p < .05$). In reality, however, the likelihood of mistakenly detecting a significant effect when none exists in the population can be inflated beyond the nominal alpha rate (.05) by a number of factors.

Understand How Data-Dependent Decisions Inflate the Type I Error Rate

Perhaps most importantly, the type I error rate can inflate – often by an unknown amount – when the various decisions that a researcher makes about how to construct their dataset and analyze their results are informed in some way by the data themselves. Such decisions are called *data-dependent* (or often "exploratory," although this term can have multiple meanings and so we avoid it here for the sake of clarity). For example, if a researcher decides whether or not to continue collecting data based on whether their primary analysis hits significance, the type I error rate of that test will inflate a little (if they engage in multiple rounds of such "optional stopping," type I error can increase substantially; see Sagarin et al., 2014). Likewise, running an analysis with or without a variety of possible covariates until one hits significance can inflate type I error (see Wang et al., 2017), as can testing an effect with three slightly different dependent measures and reporting only the ones that hit significance. In fact, even the common practice of conducting a 2×2 factorial ANOVA and reporting all effects (two main effects and an interaction) has an associated type I error rate of about 14% rather than the 5% researchers typically assume (see Cramer et al., 2016). In all of these cases, the problem arises because there are multiple possible tests that a researcher could or does run to test their research question (e.g., a test on a subsample of 100 and a test on a subsample of 200; a test on one dependent measure versus a test on a different dependent measure). When the decision about which test to run and report is informed by knowledge of the dataset in question, the type I error rate starts to inflate (see Gelman & Loken, 2014; Simmons, Nelson, & Simonsohn, 2011).

Clearly Distinguish Between Data-Dependent and Data-Independent Analyses with a Preanalysis Plan

Of course, the fact that data-dependent analyses inflate type I error does not mean that you should never let knowledge of your data guide your decisions about how to analyze your data. Data-dependent analyses are important to get to know your data and to help generate new hypotheses and theories. Moreover, in some research areas, it is difficult or impossible to analyze a dataset without already knowing something about the data (e.g., political science studies of election outcomes; Gelman & Loken, 2014). Data-dependent analyses are often extremely useful, but we want

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press

to *know* that an analysis is data-dependent so that we can calibrate our confidence in the result accordingly.

Thus, an important tool to have in your toolkit is the ability to distinguish clearly between data-dependent and data-independent analyses. A well-crafted *preanalysis plan* allows you to do just that. A preanalysis plan involves selecting and writing down ahead of time the various researcher decisions that will need to be made about how to construct and analyze a dataset, before looking at the data. Writing down the decisions ahead of time is important to circumvent human biases in thinking and memory (Chaiken & Ledgerwood, 2011; Nosek et al., 2012) – after looking at the data, it is very easy to convince ourselves we actually intended to do these particular tests and make these particular decisions all along. By creating a record of which decisions were in fact data-independent, preanalysis plans allow researchers to distinguish between data-dependent and data-independent analyses. For example, if you write down ahead of time that you will include a carefully chosen covariate in your analysis and you follow that plan, you can rest assured that you have not unintentionally inflated your type I error rate (Wang et al., 2017). On the other hand, if you decide after looking at the data to include a different covariate or none at all, you can calibrate your confidence in that data-dependent analysis accordingly (e.g., being more tentative about that result until someone can test if it replicates).

Preanalysis plans thus enable you to plan your analyses with a constrained type I error rate, allowing you to know what this rate is. However, the plan is not a guarantee for keeping an alpha level below the desired rate (usually .05). If you plan multiple comparisons (e.g., you plan to test all effects in a two-way ANOVA; Cramer et al., 2016) or inappropriate statistical tests (e.g., you use multiple regression rather than latent variables to test the incremental validity of a psychological variable, which can produce spurious results due to measurement error; see Westfall & Yarkoni, 2016; Wang & Eastwick, in press), your type I error rate may be higher than you imagine. Also, it is important not to follow the plan blindly, and always check whether the assumptions of a statistical test are met given the data.

When constructing a preanalysis plan for the first time, it is often useful to start with a template designed for your type of research. For example, psychological researchers conducting experiments often find the AsPredicted.org template useful because it clearly identifies the most common researcher decisions that an experimental psychologist will need to make and provides clear examples of how much detail to include about each one. Other templates are available on OSF (see https://osf.io/zab38/), or you can create your own tailored to your own particular research context (see Table 3 for an example). Do not be surprised to find that you forget to record some researcher decisions the first time you create a preanalysis plan for a given line of research. It can be hard to anticipate all the decisions ahead of time. But even when a preanalysis plan is incomplete, it can help you clearly identify those analyses that were planned ahead of time and those that were informed in some way by the data.

Table 3. Common decisions to specify in a preanalysis plan

Consider describing:	Example:
Planned sample size and stopping rule	Target total $N = 200$ We will collect data until Qualtrics indicates that there are completed surveys from 200 participants.
Inclusion criteria	University students 18 and older who have not participated in a previous study in this line of work will be allowed to participate.
Exclusion criteria	Participants will be excluded if they respond incorrectly to the attention check at the end of the study, as coded by a researcher blind to the rest of the data. UPDATED 10/20/2019 after opening the data file but before running any analyses: We noticed that two participants spent less than 2 seconds reading the screen that displayed the manipulation, whereas everyone else spent at least 30 seconds, so we decided to exclude these two participants.
Manipulation(s) and conditions	Consensus information (2 between-subjects conditions): Participants read that 70% of students at their university support vs. oppose a new bike law.
Predictor(s) and how they will be constructed	N/A
Dependent measure(s) and how they will be constructed	Primary/Focal DV: Participants' own attitudes toward the bike law (average of the five-item scale) Additional DV: Attitude strength (average of the four-item scale)
Any planned covariates	N/A
Planned statistical tests involving specific operational variables	Primary/Focal analysis: Independent *t* test (two-tailed) examining the effect of condition on participants' attitudes toward the bike law. Additional analysis: Independent *t* test (two-tailed) examining the effect of condition on participants' attitude strength.
Any planned follow-up or subgroup analyses	No
Any plan for type I error control (e.g., for multiple comparisons)	No

Note: Notice that although some templates ask you to identify a research question or prediction as a simple way to help readers understand the focus of your study, you can create a preanalysis plan even when you have no prediction about how your results will turn out (see Ledgerwood, 2018). Notice too that preanalysis plans must be specific to be useful (e.g., if the dependent measure does not specify how many items will be averaged, it is not clear whether the decision about which items to include was made before or after seeing the data).

Distinguish Between Different Varieties of Preregistration and their Respective Functions

As described above, preanalysis plans can be very useful for clearly distinguishing between data-independent and data-dependent analyses. *Preregistering* a preanalysis plan simply means recording it in a public repository (e.g., OSF, AsPredicted.org, or socialscienceregistry.org). However, it is important to recognize that the term *preregistration* is used in different ways by different researchers both within and beyond psychology, and that these different definitions often map onto different goals or functions (Ledgerwood & Sakaluk, 2018; see also Navarro, 2019). Table 4 outlines the most common varieties of preregistration and their intended functions.

Researchers in psychology often use the term "preregistration" to mean a preanalysis plan, and advocate using this type of preregistration to reduce unintended type I error inflation and help researchers correctly calibrate their level of confidence or uncertainty about a given set of results (e.g., Nosek, Ebersole, DeHaven, & Mellor, 2018a; Simmons, Nelson, & Simonsohn,

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press

Table 4. Different definitions of preregistration and their intended purpose

Definition	Goal
Preanalysis plan	Distinguish data-independent vs. data-dependent analyses; constraining unintended type I error inflation
Write down as much information as you can about your study before you conduct it	Transparency: Someone else can check what you said you planned ahead of time against what you actually wrote down
Record your theoretical predictions	Theory falsification
Record your intuitive predictions	Figure out how good you are personally at guessing a study's outcome
Record the existence of your study in a centralized, searchable repository	Combat publication bias
Registered report	All of the above plus reviewer objectivity (reviewers can evaluate the methods and planned analyses without being biased by the whether the results fit their intuitions or theories)

Table 5. Summary of recommendations

When planning a study:
1. Assess and maximize statistical power – Conduct an a priori power analysis. – Identify the biggest sample size worth collecting. – Use sequential analysis to conserve resources when possible. – Consider multiple approaches to boosting power.
2. Avoid unintended or invisible type I error inflation – Clearly distinguish between data-dependent and data-independent analyses with a preanalysis plan. – Distinguish between different varieties of preregistration and their respective functions. If your goal is to avoid type I error inflation, preregister a preanalysis plan that clearly constrains researcher degrees of freedom for any planned, data-independent analyses.
When evaluating a study that has already been conducted:
1. Assess the likely power of a given statistical test – Conduct a sensitivity analysis to assess power to detect a range of effect sizes. – *Do not* calculate "post-hoc" or "observed" power.
2. Assess the likely type I error rate – Understand how data-dependent decisions inflate the type I error rate. – Look for a preanalysis plan. Is it clear which analyses were data-independent? Are researcher decisions well constrained or flexible? Does the preanalysis plan account for multiple comparisons and are the analyses appropriate?

2017). But researchers in psychology and other disciplines also use the term "preregistration" to mean other practices that do not influence type I error (although they serve other important functions). Distinguishing between different varieties or elements of preregistration and thinking carefully about their intended purpose (and whether a given preregistration successfully achieves that purpose) is crucial if we want to correctly calibrate our confidence in a given set of results.

For example, researchers across disciplines sometimes use the term "preregistration" to mean a peer-reviewed registered report, where a study's methods and planned analyses are peer reviewed before the study is conducted; in such cases, the decision about whether to publish the study is made independently from the study's results (e.g., Chambers & Munafo, 2013). This type of preregistration can help constrain type I error inflation (insofar as the analyses are specified ahead of time and account for multiple comparisons), while also achieving other goals like combatting publication bias (because the decision about whether to publish the study does not depend on the direction of the results). However, the reverse is not true: A preanalysis plan by itself does not typically combat publication bias (primarily because in psychology such plans are not posted in a public, centralized, easily searchable repository).

Similarly, researchers often talk about preregistration as involving recording a directional prediction before conducting a study (e.g., Nosek et al., 2018a), which can be useful for theory falsification. However, writing down one's predictions ahead of time does not influence type I error: The probability of a given result occurring by chance does not change depending on whether a researcher correctly predicted it ahead of time (Ledgerwood, 2018). Thus, a researcher who records their predictions ahead of time without also specifying a careful preanalysis plan runs the risk of unintended type I error inflation (Nosek, Ebersole, DeHaven, & Mellor, 2018b).

Thus, in order to correctly calibrate our confidence in a given study's results, we need to know more than whether or not a study was "preregistered" – we need to ask *how* a study was preregistered. Did the preregistration contain a careful and complete preanalysis plan that fully constrained flexibility in dataset construction and analysis decisions? Did the plan successfully account for multiple comparisons? And did the researcher exactly follow the plan for all analyses described as data-independent? Thinking carefully and critically about preregistration will help you identify which of the goals (if any) listed in Table 4 have been achieved by a given study, and whether you should be more or less confident in that study's conclusions.

Conclusion

In this tutorial, we have discussed a number of strategies that you can use to calibrate the confidence you have in the results of your own studies as well as studies from other researchers (see Table 5). These strategies address typical issues researchers face when they try to assess the likely statistical power and type I error rate of a given study. By improving our ability to gauge statistical power and type I error, we can distinguish between study results that provide relatively strong building blocks for our research programs (those with high positive predictive value, as illustrated on the right side of Figure 1) and study results that provide more tentative evidence that needs to be replicated before we build on it (those with low positive predictive value, as illustrated on the left side of Figure 1). To help us get a better sense of the power of a study, we can develop good effect size intuitions; conduct a priori power analyses when we are in the planning phase of a project; and conduct sensitivity analyses when data has already been collected. To help us get a better sense of the type I error rate of a study, we can clearly distinguish between data-dependent (exploratory) and data-independent (confirmatory) analyses using a preanalysis plan; think critically about different varieties of preregistration; and evaluate whether a given preregistration successfully achieves

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press

its desired function(s). Together, these strategies can help improve our research methods as scientists, allowing us to maximize what we learn from the work that we do.

Financial support. None.

Notes

1. The positive predictive value of a finding also depends on the base rate of true effects being tested in the population. For the examples we provide in this section, we assume this base rate of true effects being tested is 50%. When it is higher (e.g., when a researcher tests incremental questions in a very well-established research literature where a hypothesis is quite likely to be true), the positive predictive value of a study will be higher. When it is lower (e.g., when a researcher tests a bold new idea in a new research literature, or postulates a counterintuitive effect), the positive predictive value of a study will be lower. The formula for computing positive predictive value is PPV = (Power*R)/(Power*R + α), where R is the odds of a found effect indeed being non-null among the effects being tested, depending on the base rate of true effects (see Button et. al., 2013; Pashler & Harris, 2012; Ioannidis, 2005 for fuller discussions of positive predictive value).

2. Note that although a type I error rate of .30 is quite far from most researchers' desired type I error rate of .05, it is not unrealistic to suggest that the actual type I error rate of a study can be considerably higher than the nominal alpha rate (typically $p < .05$). We return to this issue in Part II of this tutorial (see also Simmons, Nelson, & Simonsohn, 2011, for a vivid demonstration, and Gelman & Loken, 2014, for an in-depth discussion).

References

Anderson S.F., Kelley K. and 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.

Bakker M., van Dijk A. and Wicherts J.M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, **7**, 543–554.

Begley C.G. and Ellis L.M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Button K.S., Ioannidis J.P., Mokrysz C., Nosek B.A., Flint J., Robinson E.S. and Munafò, M.R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, **14**, 365–376.

Chaiken S. and Ledgerwood A. (2011). A theory of heuristic and systematic information processing. *Handbook of theories of social psychology* (Vol. **1**, pp. 246–166). Los Angeles, CA: SAGE.

Chambers C. and Munafo M. (2013, June 5). Trust in science would be improved by study pre-registration. *The Guardian*. https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

Chartier C., Kline M., McCarthy R., Nuijten M., Dunleavy D.J. and Ledgerwood A. (2018, November 30). The cooperative revolution is making psychological science better. *APS Observer*. Retrieved from https://www.psychologicalscience.org/observer/the-cooperative-revolution-is-making-psychological-science-better

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

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

Cramer A.O., van Ravenzwaaij D., Matzke D., Steingroever H., Wetzels R., Grasman R.P. and Wagenmakers E.J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, **23**, 640–647.

Erdfelder E., Faul F. and Buchner A. (2005). Power analysis for categorical methods. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1565–1570). CHicester, U.K.: Wiley.

Faul F., Erdfelder E., Lang A.G. and Buchner A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, **39**, 175–191.

Giner-Sorolla R. (2018, January 24). Powering your interaction [blog post]. https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Gelman A. and Carlin J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, **9**, 641–651.

Gelman A. and Loken E. (2014). The statistical crisis in science. *American Scientist*, **102**, 460–465.

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

Hoenig J.M. and Heisey D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, **55**, 19–24.

Ioannidis J.P.A. (2005). Why most published research findings are false. *PLoS Med*, **2**, e124.

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

Kanyongo G.Y., Brook G.P., Kyei-Blankson L. and Gocmen G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, **6**, 9.

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

Lakens D. and Evers E.R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, **9**, 278–292.

Le B., Dove N.L., Agnew C.R., Korn M.S. and Mutso A.A. (2010). Predicting nonmarital romantic relationship dissolution: A meta-analytic synthesis. *Personal Relationships*, **17**, 377–390.

Ledgerwood A. (2016). Introduction to the special section on improving research practices: Thinking deeply across the research cycle. *Perspectives on Psychological Science*, **11**, 661–663.

Ledgerwood A. (2018). The preregistration revolution needs to distinguish between predictions and analyses. *Proceedings of the National Academy of Sciences*, **115**, E10516–E10517.

Ledgerwood A. (2019). New developments in research methods. In E.J. Finkel & R.F. Baumeister (Eds.), *Advanced social psychology* (pp. 39–61). Oxford University Press.

Ledgerwood A. and Sakaluk J. (2018, February). *Preregistration, actually: How can (and should) researchers use preregistration pluralistically?* Paper presented at the Society for Improving Psychological Science preconference, held before the annual conference of the Society for Personality and Social Psychology, Portland, OR.

Ledgerwood A. and Shrout P.E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, **101**, 1174–1188.

Ledgerwood A., Soderberg C.K. and Sparks J. (2017). Designing a study to maximize informational value. In M.C. Makel & J.A. Plucker (Eds.), *Toward a more perfect psychology: Improving trust, accuracy, and transparency in research* (pp. 33–58). Washington, DC: American Psychological Association.

McNutt M. (2014). Journals unite for reproducibility. *Science*, **346**, 679.

McShane B.B., Böckenholt U. and Hansen K.T. (2016). Adjusting for publication bias in meta- analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, **11**, 730–749.

Navarro D. (2019, January 17). Prediction, pre-specification and transparency [blog post]. https://featuredcontent.psychonomic.org/prediction-pre-specification-and-transparency/

Nosek B.A., Ebersole C.R., DeHaven A.C. and Mellor D.T. (2018a). The preregistration revolution. *Proceedings of the National Academy of Sciences*, **115**, 2600–2606.

Nosek B.A., Ebersole C.R., DeHaven A.C. and Mellor D.T. (2018b). Reply to Ledgerwood: Predictions without analysis plans are inert. *Proceedings of the National Academy of Sciences*, **115**, E10518.

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

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press

Nyhan B. (2015). Increasing the credibility of political science research: A proposal for journal reforms. *PS: Political Science & Politics*, **48**, 78–83.

Open Science Collaboration. (2015). *Science*, 28.

Pashler H. and Harris C.R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, **7**, 531–536.

Pashler H. and Wagenmakers E-J. (2012). Editors' introduction to the Special Section on Replicability in Psychological Science: A crisis of confidence? *Perspectives on Psychological Science*, **7**, 528–530.

Proschan M.A., Lan K.G. and Wittes J.T. (2006). *Statistical monitoring of clinical trials: A unified approach*. Springer.

Richard F.D., Bond Jr C.F. and Stokes-Zoota J.J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, **7**, 331–363.

Rivers A.M. and Sherman J. (2018, January 19). Experimental design and the reliability of priming effects: Reconsidering the "Train Wreck". *PsyArXiv*.

Rosenthal R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, **86**, 638–641.

Sagarin B.J., Ambler J.K. and Lee E.M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, **9**, 293–304.

Schönbrodt F.D. and Perugini M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, **47**, 609–612.

Simons D.J. (2018). Introducing advances in methods and practices in psychological science. *Advances in Methods and Practices in Psychological Science*, **1**, 3–6.

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

Simmons J., Nelson L. and Simonsohn U. (2017, November 6). How to properly preregister a study [blog post]. http://datacolada.org/64

Sparks J. and Ledgerwood A. (2017). When good is stickier than bad: Understanding gain/loss asymmetries in sequential framing effects. *Journal of Experimental Psychology: General*, **146**, 1086–1105.

Spellman B.A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, **10**, 886–899.

Spellman B., Gilbert E.A. and Corker K.S. (2017). Open Science: What, why, and how. *PsyArXiv*.

Wang Y.A. and Eastwick P.W. (in press). Solutions to the problems of incremental validity testing in relationship science. *Personal Relationships*.

Wang Y.A. and Rhemtulla M. (in press). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial.

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

Westfall J. (2016). PANGEA [computer program]. https://jakewestfall.shinyapps.io/pangea/

Westfall J. and Yarkoni T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One*, **11**, e0152719.

https://doi.org/10.1017/prp.2020.7 Published online by Cambridge University Press