Ledgerwood, A., Soderberg, C., & Sparks, J. (2017). Designing a study to maximize informational value. In J. Plucker & M. Makel (Eds.), Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research (pp. 33-58). Washington, DC: American Psychological Association.

2

DESIGNING A STUDY TO MAXIMIZE INFORMATIONAL VALUE

ALISON LEDGERWOOD, COURTNEY K. SODERBERG, AND JEHAN SPARKS

KEY POINTS

Use this chapter to guide your methodological decisions before you start collecting or analyzing your data, in order to maximize what you can learn from your results. The toolbox of cutting-edge strategies provided here will enable you to

- understand the importance of statistical power, boost it when needed, and consider strategies for confronting real-world challenges to running highly powered studies;
- consider both the benefits and drawbacks of using online samples;
- distinguish between exploratory and confirmatory analyses so that you can learn as much as possible from your data; and
- plan programs of research that include direct, systematic, and/or conceptual replications.

http://dx.doi.org/10.1037/0000033-003
Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research, M. C. Makel and J. A. Plucker (Editors)

Recent years have witnessed a broad movement to improve methods and practices across scientific disciplines (e.g., Begley & Ellis, 2012; Button et al., 2013; Ledgerwood, 2014, 2016; McNutt, 2014; Nosek, Spies, & Motyl, 2012; Nyhan, 2015). In response to a renewed focus on how to maximize the knowledge we get from the work that we do, researchers have developed an impressive toolkit of new and newly rediscovered methodological and statistical practices. In this chapter, we draw on the cutting-edge literature on this topic to highlight a number of crucial decisions many social scientists face when designing a study that influence how informative the results can be. In other words, this chapter is the one to consult before you begin to conduct a study or analyze a preexisting data set. The decisions you make at such critical phases of the research process can have a dramatic impact on how much you learn from your eventual results.

THE CRITICAL IMPORTANCE OF POWER

One key set of decisions you will need to make at this early stage of the research process centers on the issue of statistical power. Adequately powering your study is crucial for maximizing the informational value of your eventual results, for reasons relating to both Type I error (the likelihood of erroneously detecting an effect in your study when no true effect exists) and Type II error (the likelihood of failing to detect a true effect).

Researchers often think about the issue of power as an issue of avoiding Type II errors: You want high power because it increases the likelihood that you will detect an effect if the effect is there. This way of thinking about power leads to the idea that it is desirable to have high power but that low power is only a problem if you do not see an effect. Researchers who think about power in this way might (understandably but erroneously) conclude that if they run an underpowered study and detect an effect, it constitutes especially trustworthy and impressive evidence for an effect ("I found it even with low power working against me!").

However, the fact is that low power also undermines our ability to trust effects when we do see them, in that reducing power reduces the *positive predictive value* (PPV) of a significant finding (see Button et al., 2013). PPV is the probability that a statistically significant result reflects a true positive (a real effect in the population). The PPV of your own findings would be the proportion of all of your significant results that are true positives—in other words, the likelihood that any given significant effect you detect is real. As the power of your study decreases, the number of true positives in your personal pool of significant results decreases. Meanwhile, though, if your Type

I error rate is constant, the number of false positives in your personal pool stays constant. The dwindling number of true positives means that the probability of any one of your significant results being true goes down. In addition, when power drops below 50%, effect sizes start to become dramatically overestimated, and when power drops below 10%, they can be in the wrong direction (leading you to conclude, e.g., that your manipulation increases your dependent variable when in fact the opposite is true; see Gelman & Carlin, 2014). Thus, low power reduces your ability to trust your results not only when you fail to see a significant effect but also when you do see one.

Another way to think about the issue of low power is that underpowered studies tend to produce "bouncier" effect size estimates (see Figure 2.1 for an illustration). In other words, the estimates produced by underpowered research will tend to fluctuate more wildly from one study to the next or from one subjective researcher decision to the next (e.g., decisions about whether to exclude outliers or drop an item from a scale), compared with the

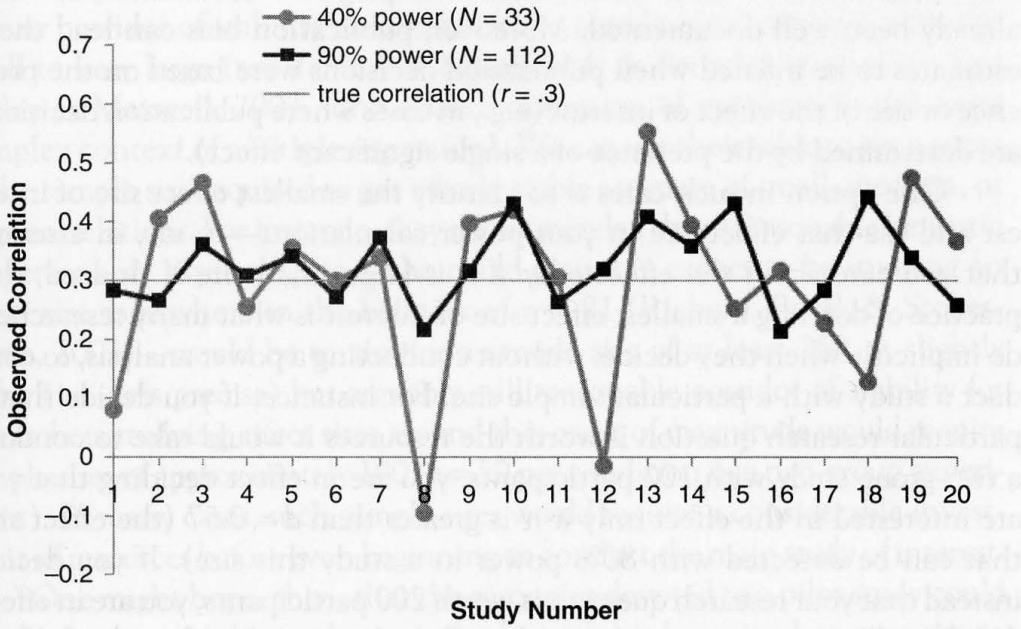

Figure 2.1. Underpowered studies tend to produce estimates that bounce more wildly from one study to the next, compared with more highly powered studies. Here we see illustrative results from a simple simulation in which a researcher runs 20 underpowered studies (powered at 40% to detect a medium-sized effect of $r = .30$) or 20 highly powered studies (powered at 90% to detect the medium-sized effect). Notice that the estimates produced by the underpowered studies (gray line) fluctuate widely around the true population correlation of .30, whereas the estimates produced by the highly powered studies (black line) cluster more tightly near the true population parameter.

more precise and stable estimates provided by highly powered studies (see Cumming, 2012; Ioannidis, 2008; and Schönbrodt & Perugini, 2013). This makes it harder to glean useful information from your results and can lead to problems later on when you or other researchers try to replicate your findings (Maxwell, 2004). Taken together, these issues point to the critical importance of estimating power when planning a study, so that you can not only boost power when needed but also acknowledge the uncertainty inherent in underpowered studies when high power cannot be achieved.

Effect Size Estimation

The potentially tricky part of any power calculation is estimating the effect size of interest. There are multiple ways to construct this estimate. Perhaps the most obvious is to look to the prior literature for similar studies or meta-analyses that provide an estimate of the expected effect size (or that enable you to put limits on the range of plausible effect sizes; see Gelman & Carlin, 2014). However, such estimates are not always available; for instance, some researchers are often interested in studying new effects that have not already been well documented. Moreover, publication bias can lead these estimates to be inflated when publication decisions were based on the presence or size of the effect of interest (e.g., in cases where publication decisions are determined by the presence of a single significant effect).[1]

One option in such cases is to identify the smallest effect size of interest and use that effect size in your power calculations—to say, in essence, that you care about the effect only if it is larger than size X. Indeed, this practice of defining a smallest effect size of interest is what many researchers do implicitly when they decide, without conducting a power analysis, to conduct a study with a particular sample size. For instance, if you decide that a particular research question is worth the resources it would take to conduct a two-group study with 100 participants, you are in effect deciding that you are interested in the effect only if it is greater than $d = 0.57$ (the effect size that can be detected with 80% power in a study this size). If you decide instead that your research question is worth 200 participants, you are in effect deciding that you are interested in the effect if it is greater than $d = 0.40$. A power analysis in this context simply makes the implicit decision an explicit one, allowing you to consider whether it is worth conducting the study (and perhaps it is not, if you realize you only have enough power to detect a giant

[1]Note that it is extremely difficult to adjust for publication bias accurately in meta-analysis. If it is likely that a given literature has been influenced by publication bias, meta-analysis may be more helpful for providing a range of plausible estimates that result from assuming different kinds of publication bias, rather than a single estimate that purports to provide the "real" effect size (see McShane, Böckenholt, & Hansen, 2016).

effect), whether you want to spend more resources, and how confident you can be in your conclusions.

Another option is to conduct a pilot study to provide an initial estimate of the effect size, which you could then use to inform the power calculation for your main study. This approach seems promising in theory, but it is worthwhile to consider several issues that may limit its usefulness in practice. First, the size of the pilot study must be large enough to provide a relatively precise estimate of the effect size—after all, a wildly inaccurate effect size estimate based on a small sample is not much help for planning your study. How large is large enough? Schönbrodt and Perugini (2013) suggested that researchers consider when a sample size will be large enough that effect size estimates reach a "corridor of stability" around the true population effect size. In other words, as the sample size increases, effect sizes go from bouncing wildly around the true population effect size to moving within a narrower and narrower corridor; as the corridor narrows, one can be more and more confident that the effect size estimate from any particular study is fairly accurate.

On the basis of their simulation, Schönbrodt and Perugini (2013) provided sample sizes typically required to reach a very narrow or moderately narrow corridor of stability around bivariate correlations ranging from quite small to very large (see Table 1 of Schönbrodt & Perugini, p. 611; see also Kelley & Maxwell, 2003, for a related discussion of precision in the more complex context of multiple regression). We can use their table to get a sense of the sample size required to get a fairly stable estimate of small, medium, or large correlations. For instance, they recommended that a reasonable heuristic for personality psychologists—who could plausibly expect to be studying an effect size somewhere in the ballpark of $r = .21$ (Richard, Bond, & Stokes-Zoota, 2003)—would be to aim for a sample size of at least 250. A slightly wider (i.e., less precise) but arguably still reasonable corridor of stability for researchers studying effect sizes around this order of magnitude would require sample sizes of approximately 100 ($n = 50$ per condition in a two-group experiment). In many cases, such sample sizes would require a considerable investment of resources before even beginning to conduct the main study of interest.

One might hope, then, that the resources devoted to a pilot study could also be incorporated into the main study, and this hope is what gave rise to the idea of an *internal pilot study*, or the idea that researchers could start collecting data to estimate an unknown effect size and then use this estimate to decide on an ultimate total sample size for that same study, in what has been termed an *adaptive design* (see, e.g., Lakens & Evers, 2014, and Wittes & Brittain, 1990). However, research has shown that reestimating a final sample size on the basis of the size of a treatment effect (e.g., a mean difference between conditions) can substantially inflate Type I error rates, and there is controversy about the best way to correct for this problem (e.g., Gordon Lan,

Soo, Siu, & Wang, 2005; Proschan & Hunsberger, 1995). For now, then, we recommend using one the strategies suggested above, or sequential analyses (described in more detail below).

Calculating Power

Once you have an effect size estimate in hand, you can conduct a classic power analysis to estimate the required sample size to achieve your desired level of power (often 80%) with your desired Type I error level (usually .05; Cohen, 1988; Ellis, 2010). One popular (and free) program that calculates power for many types of frequently used designs is G*Power (see http://www.gpower.hhu.de/en.html). However, it is important to note that such power analyses can be overly optimistic, for at least two reasons. First, when an effect size estimate comes from the prior literature on a topic, publication bias can cause that estimate to be too high (and the resulting sample size calculation to be too low to provide adequate power). To safeguard against the bias introduced by inflated effect size estimates, Perugini, Gallucci, and Costantini (2014) recommended that researchers conduct a *safeguard power analysis*, which constructs a confidence interval around the effect size estimate taken from previous research and uses the lower bound of this confidence interval in the power calculation.[2]

A second reason why classic power analyses can be overly optimistic is that they fail to take into account effect size heterogeneity; that is, the possibility that the size of an effect can vary across settings, samples, and operationalization of variables (McShane & Böckenholt, 2014). For instance, the effect of an SAT preparation course on SAT scores might be larger when the course is taught in a quiet room where students can concentrate versus a loud setting with many distractions. Classic power formulas ignore this possibility (they assume that heterogeneity is 0) and can therefore lead researchers to run underpowered studies, especially when effect sizes are small to medium. To address this problem and help researchers account for heterogeneity when planning studies, McShane and Böckenholt (2014) provided a new tool for calculating power that accounts for effect size heterogeneity and allows researchers to explore the potential consequences of heterogeneity when planning their sample sizes.[3]

Of course, these different strategies for calculating power will produce different estimates of the sample size necessary for your study. Which is right? Given that both publication bias and effect size heterogeneity characterize many areas of research, it is likely that the sample size suggested by a classic

[2]See p. 3 of their supplementary materials for the R code to run a safeguard power analysis (http://journals.sagepub.com/doi/suppl/10.1177/1745691614528519/suppl_file/10.1177_1745691614528519_SuppData.pdf).

[3]See https://blakemcshane.shinyapps.io/hetsampsize for a tutorial and instructions.

power analysis will lead you to run an underpowered study. Getting a sense of the sample sizes recommended by these updated techniques can give you useful information about how well powered your study is likely to be, allowing you to make informed choices about where to devote your resources and how much to trust your eventual findings.

Confronting Real-World Challenges to Running Highly Powered Studies

It is one thing to know you need a sample size of 250 to shed light on your research question and another to actually get that sample. Depending on your institutional resources (e.g., whether you have a large subject pool and/or funds to pay participants), the type of research you conduct (e.g., survey studies vs. intensive laboratory procedures), and the type of participants you need (e.g., adults vs. children, individuals vs. couples), obtaining large samples can be very challenging. The solution is not to ignore power considerations or conduct only easy studies but instead to confront the power challenge head on, get creative whenever possible, and—when necessary—acknowledge the limitations that arise when ideals are constrained by reality.

Large samples are one route to high power, but they are not the only one (see Asendorpf et al., 2013, and Ellis, 2010). Understanding some of the other factors that affect power can provide you with a toolbox of different strategies for conducting well-powered research. For instance, when feasible, within-subject (vs. between-subjects) designs can dramatically boost the power of an experiment (see Greenwald, 1976, for a deeper consideration of the benefits and drawbacks of within-subject designs). Likewise, researchers would do well to invest in reliable measures of their constructs. Power drops as measurement error increases—indeed, although a scale reliability of $\alpha = .70$ is often described as "adequate," such low levels of reliability can lead to substantially underpowered studies, especially when one is examining small effects (Ledgerwood & Shrout, 2011; see Stanley & Spence, 2014, for a vivid illustration of how measurement error can produce results that fluctuate wildly). Conversely, identifying or constructing and validating highly reliable measures can give your study a much-needed power boost.[4]

[4]Researchers often attempt to address measurement error at the analysis phase; for instance, latent variables are often used because they protect against the bias produced by measurement error (in that they help ensure that the estimates produced across studies will accurately center on the true population parameter). However, latent variables are even more adversely affected by measurement error than are observed variables when it comes to power. To address this issue, Ledgerwood and Shrout (2011) offered a two-step approach to testing mediation models using both latent and observed variables that maximizes both accuracy and power when unreliability is unavoidable. Still, careful planning at the design phase of a study to minimize measurement error can allow you to avoid accuracy–power trade-offs altogether and provides the best route to making your later analyses as informative as possible.

In experimental designs, the careful choice of a covariate can boost power by soaking up some of the noise in your dependent variable. For instance, a researcher interested in whether stressful (vs. relaxing) situations make people less likely to behave cooperatively could reduce some of the unexplained variance in her dependent variable of cooperative behavior by measuring individual differences in the general predisposition to be cooperative and using this variable as a covariate in her analyses. Note, however, that using covariates can lead you astray in experimental designs[5] if the covariate changes the pattern of condition means rather than simply reducing error variance or you attempt analyses with and without the covariate and report only those that reach significance (e.g., Simmons, Nelson, & Simonsohn, 2011). You can check the first by running your analysis with and without the covariate and comparing the means and error terms; you can avoid the second by selecting and recording your intended analysis plan ahead of time (see the section Distinguishing Between Exploratory and Confirmatory Research later in this chapter).

Yet another power-boosting strategy worth considering is the option of aggregating several small, underpowered studies in a small-scale meta-analysis that can provide reliable results. Researchers limited by the number of participants they can recruit in a particular time frame (e.g., an academic semester or year) or in a particular setting (e.g., a political rally) might choose to run two or more separate, small studies testing the same effect that, when aggregated, achieve adequate power. Researchers who face resource constraints in terms of their access to participants or research funds can initiate a multi-laboratory collaboration in which two or more research teams conduct a study using identical protocols. The results of such study sets can then be pooled across settings or laboratories using meta-analytic techniques to provide greater power than any one study alone (see Braver, Thoemmes, & Rosenthal, 2014, for more on small-scale meta-analyses and relevant R code[6]). Alternatively, you can start a community-augmented meta-analysis (see Tsuji, Bergmann, & Cristia, 2014) that provides a simple way for any researcher who conducts a similar study to add their data to a continually updating online meta-analysis.

Whereas some research contexts make it challenging to attain adequate power, others make it easy—for instance, researchers who work with very large data sets can often run highly powered analyses with ease. Note that when power is very high, effect size estimation becomes much more informative than significance testing because even tiny correlations can reach significance in very large samples; it is important in such cases to think carefully about

[5]In correlational designs, on the other hand, failing to include an important covariate can lead to omitted-variable bias (Kennedy, 2003).

[6]This is also available at http://www.human.cornell.edu/hd/qml/software.cfm.

effect sizes (and what effect sizes are meaningful) rather than focusing solely on whether an effect can be detected (see, e.g., Gignac & Szodorai, 2016; Hill, Bloom, Black, & Lipsey, 2008; Valentine & Cooper, 2003).

Sequential Analyses

Ensuring adequate power can also be challenging when you are conducting initial studies in new lines of research for which you have very little information about the likely size of an effect. If you guess too high when estimating your effect size, your study could be woefully underpowered; if you guess too low, you could waste substantial resources when a smaller sample would have been sufficient to detect significance. In such cases, sequential analyses can provide a valuable tool that allows you to adequately power your study to detect a potentially small effect size but to stop early and conserve resources if the effect turns out to be larger than anticipated (Lakens & Evers, 2014; Proschan, Lan, & Wittes, 2006).

In a sequential design, you choose ahead of time both a planned total sample size as well as the number of points throughout data collection at which you will conduct interim analyses on your data. At each interim analysis point that you choose you will have the option of stopping data collection early if the *p* value for the planned analysis falls below a planned criterion point. Whereas unplanned optional stopping inflates Type I error (Sagarin, Ambler, & Lee, 2014), planned optional stopping in a sequential analysis holds Type I error constant by portioning out the total desired alpha level (often .05) across the interim and final analyses.

To calculate the criterion for each interim analysis, you can use the GroupSeq package in R, which includes a graphical user interface for those who are unfamiliar with the R programming language (a step-by-step guide to using GroupSeq can be found at https://osf.io/qtufw/). This package will also calculate all the adjustments to *p* values, effect sizes, and confidence intervals necessary to account for the fact that sequential analysis was used (Lakens, 2014).

There are a few different options for setting the criterion at each interim analysis point (DeMets & Lan, 1995). Some, such as the O'Brien–Fleming method, require researchers to choose the number of interim analyses they will conduct ahead of time and to make them equally spaced. For instance, if you plan a study with a target final sample size of 300 participants and want to conduct two interim analyses, you would have to conduct those analyses after collecting data from 100 participants and then after collecting data from 200 participants. Other methods, which include different types of spending functions, allow more flexibility: You must decide a priori the upper bound on the sample size and the type of spending function you will use, but you do not

have to choose the number of interim analyses ahead of time or keep them equally spaced.[7] The R package can compute the appropriate statistics for a few different types of spending functions and for both equally and nonequally spaced interim analyses, allowing you flexibility in choosing which approach works best for you.

Although sequential designs can provide a valuable tool for balancing between the goals of boosting power and conserving resources, it is important to also acknowledge their downsides. In particular, the effect sizes obtained from sequential analyses will tend to be inflated because early interim analyses are conducted on relatively small samples and, as we saw above, small samples produce widely fluctuating estimates. Early interim analyses are therefore more likely to hit significance when a fluctuating estimate is too large, and so sequential analyses tend to overestimate effect sizes; moreover, the earlier the study is stopped, the greater the inflation will be (Zhang et al., 2012). Thus, sequential analyses are best suited for studies in which researchers are mainly interested in testing whether an effect exists rather than determining a stable estimate of the effect size itself (Lakens, 2014). Researchers interested in estimating effect sizes should use larger samples and/or use meta-analytic techniques to gain more stable, precise estimates.

ONLINE SAMPLES

As researchers have begun to pay more attention to power and the importance of adequate sample sizes, recruiting online samples has become increasingly popular. Multiple platforms now enable data collection from online participants, including Project Implicit, Amazon's Mechanical Turk (MTurk), Prolific Academic, CrowdFlower, Microworkers, and others (see, e.g., Chandler, Mueller, & Paolacci, 2014; Peer, Samat, Brandimarte, & Acquisti, 2015).

There are both benefits and drawbacks to online participant pools. One obvious benefit is that many of the platforms mentioned above allow social scientists to recruit large convenience samples quickly and at a relatively low cost. One obvious drawback is that researchers are limited to study procedures that can be effectively implemented online, and many aspects of the testing environment (e.g., distractions, multitasking) are not under the researcher's control.

[7]You may need an initial guess of the planned total sample size and the spacing of interim analyses to run the power calculations for the design, but the actual number and spacing of the analyses in the final study can deviate from the initial specifications without significantly influencing the Type I error rate, as long as the spacing of the analyses is independent of the results at any given interim analyses (see DeMets & Lan, 1995).

Other costs and benefits are perhaps less obvious. For instance, MTurk's large participant pool may allow for a more diverse sample than the typical college student sample (Buhrmester, Kwang, & Gosling, 2011; DeSoto, 2016), which could help researchers address the potential generalizability concerns that come with an exclusive reliance on undergraduate samples (Sears, 1986). Moreover, evidence suggests that MTurk data quality is high (Buhrmester et al., 2011; Crump, McDonnell, & Gureckis, 2013; Peer et al., 2015) but also potentially variable (Paolacci & Chandler, 2014). MTurk participants appear to be both extrinsically and intrinsically motivated to take surveys (e.g., Paolacci, Chandler, & Ipeirotis, 2010), and research on attention checks suggests that MTurk substantially outperforms other crowdsourcing websites and performs equally well as community samples (Peer et al., 2015). In comparison with student samples, the evidence has been more mixed, with different studies showing that MTurkers are less attentive (Goodman, Cryder, & Cheema, 2013), equally attentive (Paolacci et al., 2010), and more attentive (Hauser & Schwarz, 2015). The latest consensus appears to be that MTurkers tend to pay better attention than college students (Hauser & Schwarz, 2015; Klein et al., 2014). This newer evidence suggests that student samples often involve participants with low intrinsic motivation to pay attention to instructions (i.e., they participate only to fulfill a course requirement) and little time to learn the norms of survey participation. In contrast, the high intrinsic motivation and nonnaïveté of the MTurk population may provide the right mix of incentives to pay attention, and the attention gap between student samples and MTurk samples may widen as MTurkers become increasingly familiar with survey participation (Hauser & Schwarz, 2015). In addition, MTurk data appear to be at least as reliable (in terms of both measurement reliability and replicability) as data from laboratory experiments; however, that reliability varies across different platforms (e.g., CrowdFlower and MicroWorkers may be less reliable than MTurk; see Buhrmester et al., 2011; Peer et al., 2015).

On the other hand, online participant pools may have hidden drawbacks. Participant attrition can differ dramatically across conditions, leading to erroneous conclusions if researchers do not take steps to minimize, check, and transparently report attrition rates (see Zhou & Fishbach, 2016, for concrete recommendations). Recent research has also highlighted the small size of the total population of MTurk participants: A typical laboratory can access a pool of approximately 7,300 MTurkers at a given point in time, and this pool is largely shared with the many other researchers in the world running studies on MTurk at the same time (DeSoto, 2016; Stewart et al., 2015). Moreover, turnover is fairly slow: It takes about 7 months for half of this participant pool to leave and be replaced by new participants (Stewart et al., 2015). Whereas familiarity with survey participation in general may have

benefits, as noted above, nonnaïveté with respect to particular paradigms and hypotheses may be costly. Many MTurkers are "professional survey-takers" who complete multiple related studies (Chandler et al., 2014; Peer et al., 2015), and they may use the Internet to find answers to survey questions (Goodman et al., 2013). Although there are some methods to mitigate these problems,[8] they may still produce inaccurate effect size estimates and could limit both internal and external validity (Berinsky et al., 2012; Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015).

In addition, although online convenience samples can improve your ability to generalize your findings when they offer greater diversity than a typical college student sample, they can be far from representative; for instance, an MTurk sample is unlikely to enable researchers to generalize across cultures (e.g., Henrich, Heine, & Norenzayan, 2010). Moreover, using online samples can constrain generalizability to the extent that they impose methodological constraints on your research (e.g., requiring the use of hypothetical scenarios rather than in-person interactions in the laboratory; see Eastwick, Hunt, & Neff, 2013).

Overall, then, online participant pools offer a promising tool for enabling researchers to improve power by collecting larger samples, and yet they are not a panacea. The choice to use them should be a considered and careful one, and the push for greater power should not cause us to lose sight of their very real limitations. Online platforms such as MTurk may be particularly well suited for simple and infrequently used paradigms (rather than complex or commonly used ones). In the meantime, we should ask questions about the generalizability of conclusions that rely exclusively on data collected from online participant pools and/or hypothetical rather than live scenarios in the same way that we ask questions about the reliability of underpowered research.

Our ability to learn useful information from research conducted with online participant pools also depends on the extent to which researchers can work together to maintain these limited and shared resources. For instance, a single researcher's choice to use deception in an MTurk study may have ramifications for how those participants respond in future studies (see Hertwig & Ortmann, 2008; Jamison, Karlan, & Schechter, 2008): An MTurk worker deceived about an ostensible interaction with another participant in one study may be suspicious of the existence of a real interaction partner in a subsequent study. Researchers may want to consider strategies for minimizing deception in at least some of these shared participant pools (Bardsley, 2000).

[8]For example, researchers can ask questions about prior experience with tasks and set a priori exclusion criteria to eliminate participants with certain levels of experience.

Likewise, researchers interested in promoting the importance of highly powered studies should consider the practical tools that will help push research practices in this direction, including strategies to increase both the size of online participant pools and the capabilities of online platforms to support different types of methods and paradigms. More broadly, scholars in the social sciences would do well to design platforms for online data collection that align researcher incentives to maximize individual self-interest (i.e., to recruit a large sample as quickly as possible for the lowest cost) with the goal of preserving a high-quality shared participant pool for future use.

DISTINGUISHING BETWEEN EXPLORATORY AND CONFIRMATORY RESEARCH

When planning a study, it is important to think about whether you would like any of your eventual analyses to be confirmatory—that is, set and recorded ahead of time—rather than more exploratory and flexible in nature. Sometimes you may want to conduct purely exploratory research. The goal in such cases is to use a bottom-up approach to learn about the patterns suggested by a particular data set, in order to generate new hypotheses and/or inform future studies. Exploratory analyses can be data dependent (i.e., researchers can tailor their analytic approach to the particular nuances of the data to help capture potentially interesting patterns). For example, an exploratory approach to a correlation table might reveal a suggestive pattern in which one variable positively (and perhaps nonsignificantly) predicts several items that could plausibly tap a common construct; a researcher might then fruitfully collapse those related items into a single measure and discover a stronger relation between the predictor and the new aggregate measure (Ghiselli, Campbell, & Zedeck, 1981). In exploratory analyses, then, the point is to learn from suggestive patterns in the data rather than to use inferential statistics for the purpose of testing particular a priori hypotheses and drawing strong conclusions.

Purely exploratory research can be enormously generative, especially in the first phases of a research program when venturing into new scientific territory. Often, however, researchers are interested in conducting research that has a confirmatory component (frequently in addition to an exploratory component). If you are interested in being able to attach a high level of confidence to a particular finding (e.g., you want to be able to conclude that your experimental manipulation influenced your key dependent measure of interest or that two groups differ in their level of a particular attribute), it is important to set and record the analysis plan that you will use to test this particular finding ahead of time.

Recording Your Analysis Before Examining a Data Set

There are two simple reasons why it is important for researchers to record their analysis plan before looking at a given data set. First, the number of different ways that you look for a result changes your Type I error rate (i.e., the likelihood that you see a result in your data that is actually just chance fluctuation). For instance, if you test a single correlation with an alpha set at .05, you have a 5% chance of erroneously concluding that there is a relation between those two variables when there is none. Of course, if you test 10 different correlations, your chance of erroneously detecting a relation between at least one pair of variables increases substantially. Perhaps less intuitive is that testing an effect in multiple ways (e.g., before and after excluding a subset of participants from the analysis, using any one of several potential outcome measures) increases your Type I error rate as well (see Gelman & Loken, 2014; Kaplan & Irvin, 2015; MacCallum, Roznowski, & Necowitz, 1992; Sagarin et al., 2014; Simmons et al., 2011). Thus, to be able to interpret a small p value (e.g., $p < .05$) as strong evidence for your effect you need to know that you have not unintentionally inflated your Type I error rate by testing your effect in multiple ways. Alternatively, you can in some cases account for data-dependent flexibility by adjusting your p value (as in the case of optional stopping [see Sagarin et al., 2014] or post hoc adjustments for multiple comparisons [see Welkowitz, Cohen, & Lea, 2012]). Either way, the goal is to be able to take a statistical result at face value in terms of the strength of evidence it provides for a particular finding: If you do not know what your Type I error rate is, you cannot get a good sense of how strong the evidence is for a given conclusion (de Groot, 2014).

Second, because scientists are human, and because the human mind tends to be biased in how it processes and remembers information—especially when we are motivated to reach a particular conclusion—we cannot rely on our own minds to accurately remember what our original analysis plan was (Chaiken & Ledgerwood, 2012; Kunda, 1990; Nosek, Spies, & Motyl, 2012). In other words, once you see a significant correlation in your table of correlations, or once you notice that your effect is significant when you analyze the data one way but not the other, your human mind is quite capable of convincing you that this was the one test you intended to run all along. Recording your plan ahead of time enables you to circumvent human bias—you can know for sure which analyses you planned and which were data dependent, so that you can accurately distinguish between confirmatory and exploratory findings.

Confirmatory findings are useful because they allow you to have a high level of confidence in a particular observed relation between operational variables in your study. For instance, if you plan to test the effect of being in a high (vs. low) stress situation on a measure of creativity in a (well-powered)

study and you find a significant result, you can conclude with a reasonable level of confidence that your manipulation affected your measure. In other words, confirmatory research allows you to place a high degree of trust in the relations you observe between the particular manipulations and/or measures in your study; you can trust that the result you see is likely to be truly there, instead of an artifact of chance.

Confirmatory research is therefore an important complement to exploratory research because it allows researchers to infer with confidence the presence of a specific relation between operational variables. On the other hand, exploratory research can help bolster confidence in the meaning of that specific relation (Finkel, Eastwick, & Reis, 2015). For instance, a significant effect of a stress manipulation on a creativity measure does not guarantee that these operational variables are accurately tapping their intended constructs. If exploratory analyses were to reveal that stress also influences a host of other cognitive outcomes that (like creativity) require cognitive resources, you might begin to suspect that the initial result you observed was part of a broader story about stress and cognitive resources, not creativity per se. Such exploratory analyses can be especially important when working with large data sets or when conducting independent conceptual replications is difficult or costly (see Finkel et al., 2015). A preanalysis plan should never prevent researchers from conducting additional exploratory analyses—the point is simply to clearly and transparently label such additional analyses as exploratory instead of as specified ahead of time (Casey, Glennerster, & Miguel, 2012; Chambers, Feredoes, Muthukumraswamy, & Etchells, 2014; de Groot, 2014; Humphreys, Sanchez de la Sierra, & van der Windt, 2013).

Setting Plans for Confirmatory Research

There are a variety of ways to set and record a preanalysis plan for confirmatory research, ranging from very basic to very detailed and from private to public. For instance, a research team might develop a set of core features (e.g., planned total sample size, planned exclusion criteria, any planned confirmatory statistical tests) that they always record for themselves before conducting a study, so that they can easily distinguish between exploratory and confirmatory findings later when conducting their analyses. Another research team might prefer to publicly preregister a detailed preanalysis plan for each study, using an independent registry (e.g., http://www.socialscience registry.org, http://www.opensciencefframework.org, http://www.egap.org/design-registration).[9] The most useful format and content of a preanalysis

[9]Note that public preregistration has the added benefit of helping to address the file drawer problem (Rosenthal, 1979; Simes, 1986).

TABLE 2.1
Common Content for a Preanalysis Plan

Consider specifying	Example
Planned sample size	Target total: $N = 200$
Inclusion and exclusion criteria	Participants must respond correctly to attention-check item
Variable construction	Predictor: Group identification (average of 10-item measure) DV: Willingness to pay for identity-related products (average of dollar amounts indicated for each of the five products presented)
Primary versus secondary outcome measures	Primary outcome measure: Willingness to pay DV Secondary measure: Liking for products (average of liking ratings for each of the five products)
Any planned covariates	Annual household income measure
Planned statistical tests involving specific operational variables	Linear regression (regressive willingness to pay on group identification with income as a covariate)
Any planned follow-up or subgroup analyses	N/A

Note. DV = dependent variable; N/A = not applicable.

plan is likely to vary across research teams and projects, depending on the type of research, the complexity of the analyses, and the norms of a given field (see Casey et al., 2012, for an excellent example of how to grapple with the nuances and trade-offs involved in choosing the timing and level of detail for various elements in a preanalysis plan). If you are new to confirmatory plans, consider starting with something basic that you feel comfortable with (something is better than nothing) and building from there. Your main goal is to ensure that you will be able to accurately distinguish between exploratory and confirmatory analyses and conclusions and that any decisions you make for your confirmatory analyses are independent of the data themselves. Common examples of content that a researcher might specify in a preanalysis plan are listed in Table 2.1 (see also Glennerster & Takavarasha, 2013).

Of course, it is difficult to anticipate every possible complication that can arise in the research process, and there will be times when you need to alter a preanalysis plan after recording it.[10] For instance, you might plan to conduct a linear regression only to realize on seeing the data that the pattern

[10]There are also research contexts in which preanalysis plans are simply not feasible. In such cases, a more complete analysis of the data that explicitly takes into account all possible comparisons may be the best way forward (Gelman & Loken, 2014; see Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016, for concrete recommendations).

is curvilinear. In such cases, the preanalysis plan should never prevent you from performing the more statistically appropriate test; instead, you should transparently record the change to the preanalysis plan and note the rationale. More broadly, it is always important, regardless of whether you are conducting exploratory or confirmatory analyses, to test your statistical assumptions and to actually look at your data. Are your measures skewed? Could your results be misleading because of an extreme outlier, a failed manipulation, the presence of an unexpected moderator, an unanticipated ceiling effect, or a measure with limited variability? The point of a preanalysis plan is not to constrain your data analysis to the rote and unconsidered implementation of a fixed analysis script—the point is to clearly distinguish between what you planned ahead of time and what you chose to do after looking at your data.

PLANNING PROGRAMMATIC RESEARCH: DIRECT, SYSTEMATIC, AND CONCEPTUAL REPLICATION

No matter how carefully you plan your study to maximize its informational value, at the end of the day it is still a single study—a data point that can usefully contribute to a cumulative understanding of a phenomenon rather than providing a definitive, stand-alone conclusion (see Braver et al., 2014; Cumming, 2012; Ledgerwood & Sherman, 2012). You want that data point to be as informative as possible, but you may also want to accumulate multiple data points that can together provide a more substantial contribution to a given topic area. When considering how best to assemble a package of studies, it is useful to consider how direct, systematic, and conceptual replication could each contribute to your cumulative understanding of a research question (see also Chapter 14, this volume).

Direct replications (also called *close* or *exact* replications) aim to repeat the procedures used in a prior study as closely as possible (Fabrigar & Wegener, 2016; Hendrick, 1991; Schmidt, 2009). Direct replications serve to increase confidence in an observed relationship between particular operational variables (i.e., the specific manipulations and/or measures used in a previous study). For instance, if exploratory analyses in an initial study provide evidence suggestive of a particular pattern of results, a direct replication would provide an opportunity to confirm that pattern in an independent data set. If you wish to increase your confidence in a particular finding (e.g., you observed an interesting effect of your manipulation on your primary outcome measure, but only after an unanticipated change to your preanalysis plan), direct replication is often a useful next step.

Systematic replications aim to vary presumably incidental aspects of the context in which a finding was initially obtained, in order to test the critical

assumption that those details are in fact irrelevant to the finding (Kantowitz, Roediger, & Elmes, 2014). Systematic replications help increase confidence in the generalizability of an observed relation between particular operational variables. For instance, upon noting an interesting correlation in an initial study, a survey researcher might want to systematically replicate it in a second study that varies the order of the survey questions, to rule out the possibility that the initial results might be specific to a particular question order (Schwarz, 1999). An experimental researcher might want to systematically replicate the effect of a manipulation on a particular measure in a second study using different stimuli, to test whether the initial results were specific to a particular stimulus set (Roediger, 2012; see also Westfall, Judd, & Kenny, 2015). Systematic replications help scientists combat confirmation bias in their research process by pushing them to explicitly consider and test whether variables presumed irrelevant for producing an effect might be relevant. Systematic replications encourage the question "What *shouldn't* be important for producing this effect?" rather than only "What *should* be important?" Systematic replication is therefore often a useful intermediate step between direct and conceptual replication.

Conceptual replications aim to vary the particular operationalizations of a given theoretical construct (i.e., the manipulations and/or measures used in a particular study), in order to test whether different operationalizations of the same theoretical construct will produce the same effect. Conceptual replications serve to increase confidence in the meaning of a particular result. If multiple possible operationalizations of the same theoretical variable produce the same pattern of findings, you can be more confident that the results reflect something about the theoretical construct rather than the particular operationalization used to assess it (Brewer & Crano, 2014; Cook & Campbell, 1979). Conceptual replications are therefore useful when you are confident about the presence of a particular pattern of results between operational variables but you want to know if they are really tapping the intended theoretical concepts. If your theoretical predictions hold up across a range of operationalizations, then you can be more confident that you are learning about the underlying concepts and theory rather than a specific instance of an effect (Crandall & Sherman, 2016; Fabrigar & Wegener, 2016).

In general, when conducting replications it is important to appreciate how widely results can fluctuate from one study to the next because of chance, especially with small samples and imperfect measures, and to adopt a cumulative approach that aggregates across studies rather than counting each one in isolation as a "success" or "failure" (see Braver et al., 2014; Eastwick, Neff, Finkel, Luchies, & Hunt, 2014; Fabrigar & Wegener, 2016; Stanley & Spence, 2014). For instance, suppose you conduct a confirmatory analysis of a particular effect in three independent data sets and find a significant result

in one case and a nonsignificant result in the other two. The best approach to understanding these data would be a meta-analytic one that aggregates across the three findings to provide a cumulative understanding of the effect (rather than concluding that one study "worked" and the other two "failed to replicate"; see also Gelman & Stern, 2006).

It is also important to recognize that the goals served by conducting an independent direct, systematic, or conceptual replication can be served in other ways as well, and the best tool for pursuing a given goal is likely to vary across different research contexts. For instance, the goal of attaining high confidence in a given relation between operational variables can be served by conducting a series of smaller, tightly controlled experiments or by conducting one very large and well-powered study in the first place. The goal to increase confidence in the generalizability of a given relation between operational variables can be served by conducting a series of systematic replication studies that vary in the stimulus set used, or by including a larger set of stimuli in the original study and treating stimuli as a random factor in the design (Judd, Westfall, & Kenny, 2012). And the goal to increase confidence in the meaning of a particular result can be served by conducting conceptual replications, or by conducting additional analyses in a large data set that help provide converging evidence for an effect across a range of measures, boosting confidence in convergent and divergent validity (Finkel et al., 2015). Choose the tools that work best for addressing your particular goals in your own particular research context.

CONCLUSION

The decisions you make when planning a study or a series of studies have important implications for how much you learn from your results. How can you ensure adequate power? Who will comprise your sample? How will you distinguish between exploratory and confirmatory findings? What tools will best enable you to have a high level of confidence in your results, and what kind of confidence is most important to you at this stage of the research process? A careful consideration of these questions will help maximize the information you learn from the work that you do.

RECOMMENDED READING

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, 9, 333–342.

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

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

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

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 4*, 609–612.

REFERENCES

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119. http://dx.doi.org/10.1002/per.1919

Bardsley, N. (2000). Control without deception: Individual behaviour in free-riding experiments revisited. *Experimental Economics, 3*, 215–240. http://dx.doi.org/10.1023/A:1011420500828

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

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*, 351–368.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, 9, 333–342. http://dx.doi.org/10.1177/1745691614529796

Brewer, M. B., & Crano, W. D. (2014). Research design and issues of validity. In H. T. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 11–26). New York, NY: Cambridge University Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3–5. http://dx.doi.org/10.1177/1745691610393980

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

Casey, K., Glennerster, R., & Miguel, E. (2012). Reshaping institutions: Evidence on aid impacts using a pre-analysis plan. Report No. W17012, National Bureau of Economic Research, Cambridge, MA. http://dx.doi.org/10.3386/w17012

Chaiken, S., & Ledgerwood, A. (2012). A theory of heuristic and systematic information processing. In P. A. M. van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 246–266). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781446249215.n13

Chambers, C. D., Feredoes, E., Muthukumraswamy, S. D., & Etchells, P. J. (2014). Instead of "playing the game" it is time to change the rules: Registered reports at *AIMS Neuroscience* and beyond. *AIMS Neuroscience*, *1*, 4–17.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112–130. http://dx.doi.org/10.3758/s13428-013-0365-7

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, *26*, 1131–1139. http://dx.doi.org/10.1177/0956797615585115

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. New York, NY: Rand McNally.

Crandall, C., & Sherman, J. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. http://dx.doi.org/10.1016/j.jesp.2015.10.002

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*, e57410. http://dx.doi.org/10.1371/journal.pone.0057410

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

de Groot, A. D. (2014). The meaning of "significance" for different types of research [translated and annotated by E.-J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, . . . H. L. J. van der Maas]. *Acta Psychologica*, *148*, 188–194. (Original work published 1956) http://dx.doi.org/10.1016/j.actpsy.2014.02.001

DeMets, D. L., & Lan, K. K. G. (1995). The alpha spending function approach to interim data analyses. In P. Thall (Ed.), *Recent advances in clinical trial design and analysis* (pp. 1–27). Boston, MA: Kluwer Academic. http://dx.doi.org/10.1007/978-1-4615-2009-2_1

DeSoto, A. (2016, March). Under the hood of Mechanical Turk. *APS Observer*, *29*(3). Retrieved from https://www.psychologicalscience.org/publications/observer/2016/march-16/under-the-hood-of-mechanical-turk.html

Eastwick, P. W., Hunt, L. L., & Neff, L. A. (2013). External validity, why art thou externally valid? Recent studies of attraction provide three theoretical answers. *Social & Personality Psychology Compass*, *7*, 275–288. http://dx.doi.org/10.1111/spc3.12026

Eastwick, P. W., Neff, L. A., Finkel, E. J., Luchies, L. B., & Hunt, L. L. (2014). Is a meta-analysis a foundation, or just another brick? Comment on Meltzer, McNulty, Jackson, and Karney (2014). *Journal of Personality and Social Psychology, 106*, 429–434. http://dx.doi.org/10.1037/a0034767

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511761676

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

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108*, 275–297. http://dx.doi.org/10.1037/pspi0000007

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641–651. http://dx.doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460–465. http://dx.doi.org/10.1511/2014.111.460

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

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences: Origin and evolution*. New York, NY: W. H. Freeman.

Gignac, G. E., & Szodorai, E. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78.

Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*. Princeton, NJ: Princeton University Press.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213–224. http://dx.doi.org/10.1002/bdm.1753

Gordon Lan, K. K., Soo, Y., Siu, C., & Wang, M. (2005). The use of weighted Z-tests in medical research. *Journal of Biopharmaceutical Statistics, 15*, 625–639. http://dx.doi.org/10.1081/BIP-200062284

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83*, 314–320. http://dx.doi.org/10.1037/0033-2909.83.2.314

Hauser, D. J., & Schwarz, N. (2015). Elaborative thinking increases the impact of physical weight on importance judgments. *Social Cognition, 33*, 120–132. http://dx.doi.org/10.1521/soco.2015.33.2.120

Hendrick, C. (1991). Replications, strict replications, and conceptual replications: Are they important? In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 41–49). Newbury Park, CA: Sage.

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

Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior, 18*, 59–92. http://dx.doi.org/10.1080/10508420701712990

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172–177. http://dx.doi.org/10.1111/j.1750-8606.2008.00061.x

Humphreys, M., Sanchez de la Sierra, R., & van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis, 21*, 1–20. http://dx.doi.org/10.1093/pan/mps021

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640–648. http://dx.doi.org/10.1097/EDE.0b013e31818131e7

Jamison, J., Karlan, D., & Schechter, L. (2008). To deceive or not to deceive: The effect of deception on behavior in future laboratory experiments. *Journal of Economic Behavior & Organization, 68*, 477–488. http://dx.doi.org/10.1016/j.jebo.2008.09.002

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54–69. http://dx.doi.org/10.1037/a0028347

Kantowitz, B., Roediger, H., III, & Elmes, D. (2014). *Experimental psychology*. Stamford, CT: Cengage Learning.

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLOS ONE, 10*, e0132382. http://dx.doi.org/10.1371/journal.pone.0132382

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8*, 305–321. http://dx.doi.org/10.1037/1082-989X.8.3.305

Kennedy, P. (2003). *A guide to econometrics*. Cambridge, MA: MIT Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology, 45*, 142–152. http://dx.doi.org/10.1027/1864-9335/a000178

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

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

Lakens, D., & 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. http://dx.doi.org/10.1177/1745691614528520

Ledgerwood, A. (2014). Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, 9, 275–277. http://dx.doi.org/10.1177/1745691614529448

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. http://dx.doi.org/10.1177/1745691616662441

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, *7*, 60–66. http://dx.doi.org/10.1177/1745691611427304

Ledgerwood, A., & 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. http://dx.doi.org/10.1037/a0024776

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. http://dx.doi.org/10.1037/0033-2909.111.3.490

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

McNutt, M. (2014, January 17). Reproducibility. *Science*, *343*, 229. http://dx.doi.org/10.1126/science.1250475

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, 9, 612–625. http://dx.doi.org/10.1177/1745691614548513

McShane, B. B., Böckenholt, U., & 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. http://dx.doi.org/10.1177/1745691616662243

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. http://dx.doi.org/10.1177/1745691612459058

Nyhan, B. (2015). Increasing the credibility of political science research: A proposal for journal reforms. *Political Science & Politics*, 48, 78–83. http://dx.doi.org/10.1017/S1049096515000463

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188. http://dx.doi.org/10.1177/0963721414531598

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419.

Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2015). *Beyond the Turk: An empirical comparison of alternative platforms for crowdsourcing online behavioral research.*

SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2594183. http://dx.doi.org/10.2139/ssrn.2594183

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, 9, 319–332. http://dx.doi.org/10.1177/1745691614528519

Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, *51*, 1315–1324. http://dx.doi.org/10.2307/2533262

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

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331

Roediger, H. L., III (2012, February). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*(9). Retrieved from http://www.psychologicalscience.org/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html

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

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

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. http://dx.doi.org/10.1037/a0015108

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. http://dx.doi.org/10.1016/j.jrp.2013.05.009

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105. http://dx.doi.org/10.1037/0003-066X.54.2.93

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515–530. http://dx.doi.org/10.1037/0022-3514.51.3.515

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*, 751–754. http://dx.doi.org/10.1093/biomet/73.3.751

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

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, 9, 305–318. http://dx.doi.org/10.1177/1745691614528518

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702–712. http://dx.doi.org/10.1177/1745691616658637

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making, 10*, 479–491.

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science, 9*, 661–665. http://dx.doi.org/10.1177/1745691614552498

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

Welkowitz, J., Cohen, B. H., & Lea, R. B. (2012). *Introductory statistics for the behavioral sciences*. New York, NY: Wiley.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10*, 390–399. http://dx.doi.org/10.1177/1745691614564879

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine, 9*, 65–71. http://dx.doi.org/10.1002/sim.4780090113

Zhang, J. J., Blumenthal, G. M., He, K., Tang, S., Cortazar, P., & Sridhara, R. (2012). Overestimation of the effect size in group sequential trials. *Clinical Cancer Research, 18*, 4872–4876. http://dx.doi.org/10.1158/1078-0432.CCR-11-3118

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/pspa0000056