New developments in research methods

Alison Ledgerwood

UC Davis

Chapter in press. Please cite as:

Ledgerwood, A. (in press). New developments in research methods. In R. F. Baumeister & E. J. Finkel (Eds.), *Advanced Social Psychology, Second Edition*. Oxford University Press.

Abstract

Recent events have placed psychological science at the forefront of a broad movement across scientific disciplines to improve research methods and practices. Many of the methodological tools emerging from this context have focused on two core aims: (1) maximizing what researchers can learn from each individual study that they conduct, and (2) maximizing what researchers can learn from synthesizing across multiple studies on a given topic. This chapter provides readers with the concrete tools they need to pursue each of these aims in their own research, with a focus on discussing when, why, and how to implement each tool. Readers will learn how to think about and distinguish between exploratory research questions versus confirmatory hypotheses and between exploratory versus confirmatory analyses, how to plan for unexpected results, how to boost statistical power, and the importance of conducting different kinds of replications. Readers will also learn about cutting-edge methods for conducting within-paper meta-analyses and adjusting for publication bias in meta-analysis.

Keywords: Best practices, preregistration, pre-analysis plan, open science, false positive, replication, HARKing, exploratory research, publication bias, power analysis

New developments in research methods

Our field has witnessed a recent surge of attention to questions about how researchers can maximize the knowledge they get from the research they do (Reis, this volume). This new emphasis has placed psychological science at the forefront of a broad movement to improve methods and practices across scientific disciplines (Begley & Ellis, 2012; Button et al., 2013; Ledgerwood, 2016; McNutt, 2014; Nosek, Spies, & Motyl, 2012; Nyhan, 2015; Vazire, 2017).

In many ways, it makes sense that psychology would spearhead these new meta-scientific and quantitative developments. After all, many of the recently voiced concerns about methods and practices in psychology have been raised before, and repeatedly: Our field has a long and rich history of noting the challenges posed by publication bias, low statistical power, and the failure to distinguish clearly between exploratory and confirmatory analyses (Cohen 1962; de Groot, 2014; Greenwald 1975; Kerr, 1998; Maxwell, 2004; Rosenthal 1979). Until recently, however, although many scholars tended to nod along when such concerns were raised, attempts to address them were somewhat isolated and sporadic. Meanwhile, entrenched reviewer expectations and traditional publishing models created an inertia that limited change.

Yet something new happened over the last decade: The audience interested in listening to and acting on methodological concerns suddenly mushroomed (Simons, in press). This surge of interest arose partly in response to a confluence of specific papers and events within the field (e.g., Bem, 2011; Tilburg University, 2011; Simmons, Nelson, & Simohnson, 2011; Vul, Harris, Winkielman, & Pashler, 2009). These developments within the field were situated within the broader context of replicability concerns emerging across disciplines ranging from cancer research to neuroscience, and they were galvanized by new communication technologies that kept conversations about methods and practices front and center (Fanelli, 2018; Ledgerwood,

2014; Spellman, 2015). The resulting widespread interest in understanding and leveraging ideas for improving methods and practices has created a new and exciting opportunity for advancing the rigor and quality of our science.

Many of the methodological and statistical tools and recommendations that have emerged within this new context have focused on two core aims. The first aim is to maximize what researchers can learn from each individual study they conduct. If each study is a brick, then the goal here is to ensure that each brick is solid; a useful and reliable building block (Poincaré, 1902; Forscher, 1963). This chapter will survey several tools that promote this first aim, including distinguishing between exploratory and confirmatory hypotheses and analyses, planning for unexpected results, boosting statistical power, and conducting direct, systematic, and conceptual replications.

The second aim is to maximize what researchers can learn from synthesizing across multiple studies. The goal in this case is to improve the soundness of the structures built from the individual bricks. This chapter will survey new advancements in meta-analytic techniques that promote this second aim, including within-paper meta-analysis and methods for adjusting for publication bias.

Making Good Bricks: How to Maximize What We Learn from Individual Studies

Numerous considerations go into designing an informative study, from thinking broadly about how to balance different scientific goals (Brewer & Crano, 2014; Finkel, Eastwick, & Reis, 2017) to making careful decisions about how to write effective questionnaire items (Bandalos, 2018; Schwarz, 1996) or whether to use an online sample (Ledgerwood, Soderberg, & Sparks, 2017). This chapter focuses on reviewing a subset of these tools that have recently received considerable attention in the field as part of the push to improve methods and practices

in psychological science.

Exploratory and Confirmatory Predictions and Analyses

One set of issues that has received considerable attention in recent years revolves around the importance of clearly and transparently distinguishing between exploratory and confirmatory aspects of a study. One of these issues stems from the goal of properly testing theories: If you want to *test* a theoretical prediction (e.g., "downward social comparison improves affect"), the prediction must be constructed independently from data that will be used to test that prediction (Mayo, 1991). When researchers portray a post-hoc *explanation* for a study's results as a theoretical *prediction* that they made before seeing those results (often called "HARKing," or Hypothesizing After Results are Known), it leads readers to mistakenly infer that data used to inform a theory were actually used to test the theory, when no such test has yet occurred (Kerr, 1998; Rubin, 2007).[1] Blurring the distinction between exploration and confirmation with respect to theory testing hinders scientific progress because it undermines our ability to accurately assess the extent to which various theories have been tested.

Another of these issues stems from the goal of controlling Type I error (i.e., the likelihood that you see a significant result in your data when in fact it is just chance fluctuation). If you want to conduct an analysis with a specific Type I error rate (most commonly, .05), decisions about how to construct the dataset and analyze the data must be made independently from the data themselves (Gelman & Loken, 2014). When researchers do not clearly and transparently label which of these decisions were made before versus after knowing something about the data in question, it can lead to an unknown (and sometimes very high) degree of Type I error inflation (Mills, 1993; Simmons, Nelson, & Simonsohn, 2011). Blurring the distinction

[1] Think of it this way: To provide a fair test of a theoretical prediction, a study must be able to either corroborate or falsify the prediction. Once you use a study's results to inform a theoretical prediction, you cannot test that prediction using those same results.

between exploration and confirmation with respect to data analytic decisions hinders scientific progress because it undermines our ability to accurately calibrate our confidence in a particular study's results.

Thus, when planning a new study, it is important to make a conscious choice about whether you want it to be exploratory or confirmatory in terms of its (1) predictions and/or (2) analyses (see Figure 1). First, you must ask yourself: Is my goal in conducting this study to test a theoretical prediction? If so, then the data cannot influence the theoretical prediction that you choose to make. Second, you must ask yourself: Do I want to know my precise Type I error rate (which can enable greater confidence in a study's results)? If so, then the data cannot influence the decisions you make about dataset construction and analysis. Note that there are good reasons to answer "no" as well as "yes" to either or both of these questions, and many studies have a combination of exploratory and confirmatory elements.

Figure 1: A study can be exploratory or confirmatory in terms of its predictions as well as in terms of its analyses.

	Exploratory Research Question (Information gathering, theory building)	**Confirmatory Hypothesis (Theory testing; involves a directional, *a priori* prediction)**
Exploratory Analyses (Data-dependent researcher decisions)	Question: Does X affect Y? Approach: Explore results using data-dependent analyses.	Prediction(s): Theory 1 predicts that X will increase Y, whereas Theory 2 predicts that X will decrease Y. Approach: Explore results using data-dependent analyses.
Confirmatory Analyses (Data-independent researcher decisions)	Question: Does X affect Y? Approach: Pre-analysis plan that specifies the exact manipulation, measure construction, and analysis.	Prediction(s): Theory 1 predicts that X will increase Y, whereas Theory 2 predicts that X will decrease Y. Approach: Pre-analysis plan that specifies the exact manipulation, measure construction, and analysis.

For example, imagine that a researcher wants to test a theory-derived prediction that a particular framing manipulation will influence participants' attitudes. If she is able to foresee the decisions she will have to make about data analysis, she may want to plan these analytic decisions ahead of time (so that the results of these data-independent analyses can provide stronger evidence testing the prediction). This part of her study would be captured by the bottom right quadrant of Figure 1 (directional predictions, data-independent analyses). Meanwhile, the researcher might have an additional, exploratory question she wants to ask in this study that isn't based on theory or existing literature—perhaps she wonders whether her framing effect would generalize to a low-SES sample. She could plan ahead of time her analytic decisions for testing this research question—a "confirmatory" analysis of an "exploratory" research question (bottom left quadrant of Figure 1). Alternatively, she might prefer to approach analyses of the low-SES sample in a data-dependent way, given that this new sample might respond to the paradigm or manipulation in unexpected ways (top left quadrant of Figure 1).

Exploratory research questions vs. confirmatory hypotheses. The first way in which a study can be exploratory or confirmatory is in terms of its predictions. Many research questions—especially those asked early on in a topic area, during the theory generation phase—are exploratory vis-à-vis theory. You might have an intuition or a question about how two variables are related, but there is no strong theory clearly articulating that they should be related in a particular way. In these cases, the results of your study might give you good ideas for building toward a future theory, but they will not provide clear evidence for or against an existing theory or claim.

Once theories are developed and refined, they make specific, testable hypotheses that can be supported or refuted. You might be interested in testing a theory that makes a clear prediction about the relationship between two or more variables. In these cases, the results of your study may support the hypothesis, thereby providing a data point that corroborates the theory, or refute the hypothesis, thereby providing a data point that falsifies the theory. Preregistering a hypothesis is crucial for theory testing because it circumvents the natural human tendency to misremember past events in line with current knowledge and goals (Ross & Wilson, 2000)—for example, a researcher who finds an initially unexpected result may easily convince herself that the result actually fits the theory she was trying to test. Such motivated remembering hampers theory testing because it can lead theories to be unfalsifiable to the extent that any result can be reinterpreted as providing evidence in favor of the theory.

How to preregister a theoretical prediction. If you want your study to provide a confirmatory test of a theoretical prediction, it is useful to specify and record that hypothesis ahead of time, before conducting the study (or at least before looking at the data). A *test* of a prediction should be a fair test—that is, it should be possible to specify a set of results that would support the claim being tested as well as a set of results that would refute that claim. For example, the prediction that downward social comparison will enhance perceived ability (derived from social comparison theory; Gerber, Wheeler, & Suls, 2017) would be supported by results showing that perceived ability is higher following a downward (vs. upward) social comparison manipulation, and refuted by results showing the opposite. Preregistration provides a useful tool for ensuring that a study provides a fair test of a theoretical claim, enabling researchers to assess whether the theory can predict something new as well as explaining something already known.

Directional hypotheses can be recorded privately (e.g., stored on a shared lab drive) or publicly (e.g., a preregistration uploaded to an online repository or stated in a published theoretical article). Public preregistrations of theoretical predictions may be particularly useful and compelling insofar as accountability to a public audience can (under the right circumstances) help researchers think more carefully and evenhandedly (Lerner & Tetlock, 1999). When *preregistering a theoretical prediction*, it is important to specify clearly the theory or model you are using to derive the prediction. You should also describe the prediction not only in terms of the relevant conceptual variables (e.g., "psychological distance will increase abstraction") but also in terms of the specific manipulations and measures you will use in this particular study (e.g., "asking participants to imagine their life in one year vs. one week will increase their scores on the Behavioral Identification Form, a common measure of abstraction"). Note that you can preregister multiple competing predictions to test two or more theories against each other (e.g., "whereas construal level theory would predict that distance will increase the value placed on abstract features of a product, models of temporal discounting would predict the reverse").

When writing up the results of a study whose results bear on a theory of interest, you should transparently state whether or not you set and recorded your predictions ahead of time. For example, you might clarify that the results of your study helped you refine a theory (i.e., that you were theory building, not theory testing), or that you preregistered your hypotheses (i.e., that you were theory testing), or that you did not predict a result but think it seems consistent with a particular theory (i.e., that you are connecting to a theory without seeking to refine or test it). Such clarity and transparency is critical for enabling readers to understand whether to interpret your results as *informing* versus *testing* versus *connecting to* the theory or theories that you discuss in your paper.

Exploratory vs. confirmatory analyses. The second way in which a study can be exploratory or confirmatory is in terms of its dataset construction and analysis decisions. Exploratory in this sense means that your decisions about your data (e.g., when to stop collecting data, how to construct your measures, and how to analyze your results) are to some degree data-dependent—tailored to the particular nuances of your data to help capture potentially interesting patterns. For example, you might run a study, find an ambiguous result (e.g., two condition means differ in a potentially interesting way but $p = .11$), and decide to collect another 100 participants to see if the difference disappears (suggesting that it was noise) or becomes clearer (suggesting that it's probably worth following up; Sagarin, Ambler, & Lee, 2014). Or, you might explore a correlation table and notice that one variable positively (and perhaps nonsignificantly) predicts several items that could potentially tap a common construct; you might then collapse those related items together in a single measure and find a stronger relation between the predictor and the new aggregate measure, suggesting a fruitful direction for further research (Ghiselli, Campbell, & Zedeck, 1981). Or, you might explore the effect of a failed manipulation on a number of auxiliary measures to try to gain insight into what went wrong and how you could design a better manipulation in the future (e.g., if telling student participants that a policy will affect their classmates failed to influence the perceived relevance of the policy, you might explore whether participants reported liking and caring about their classmates).

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 drawing strong conclusions based on *p*-values. Purely exploratory research can be enormously generative, especially in the first phases of a research program when venturing into new scientific territory.

In contrast, confirmatory analyses are data-independent—planned ahead of time, before knowing anything about how the variables in your dataset are related to each other. If you are interested in being able to attach a high level of confidence to a particular statistical result (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 individual difference measure), it is important to set and record the analysis plan that you will use to test this particular finding ahead of time. There are two simple reasons for this.

First, your Type I error rate increases to the extent that you look for a given result in a variety of different ways. For instance, if you test one correlation and set your alpha at .05, you have a 5% chance of incorrectly concluding that there is an association between those two variables in the population when in fact none exists. But of course, if you test ten different correlations, your chance of erroneously detecting an association when none exists between at least one pair of variables increases substantially. There are many other forms of flexible testing that can increase Type I error as well—for example, testing an effect before and after excluding a subset of participants from the analysis, testing an effect with and without a variety of covariates included in the model, or testing an effect on several different outcome measures (Gelman & Loken, 2014; Kaplan & Irvin, 2015; MacCallum, Roznowski, & Necowitz, 1992; Sagarin et al., 2014; Simmons et al., 2011). In reality, knowing anything about your data can produce flexible testing by subtly influencing the kinds of tests you think to run (e.g., noticing that one condition mean is higher than three others might lead you to think to run a complex contrast; knowing that two variables are correlated in a large dataset might lead you to think of testing the correlation between a pair of related variables). In order to interpret a small p-value (like $p < .05$) as relatively 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 or by tailoring the test you choose to run to what the data happen to look like (de Groot, 2014). In some cases, you can account for data-dependent flexibility (often called "researcher degrees of freedom;" Simmons et al., 2011) by adjusting your *p*-value (as in the case of optional stopping; Sagarin et al., 2014, or post-hoc adjustments for multiple comparisons; Welkowitz, Cohen, & Lea, 2012); of course, such adjustments require that you know exactly which analyses were data-dependent. In many contexts, then, setting an analysis plan can be a useful tool: It enables researchers to take a statistical result at face value in terms of the strength of evidence it provides for a particular finding (Nosek, Ebersole, DeHaven, & Mellor, 2018).

Second, analysis plans enable scientists to circumvent human biases that can otherwise creep into the data analysis and inference process. As noted earlier, the human mind tends to be biased in how it processes and remembers information, especially when a person is motivated to reach a particular conclusion (Chaiken & Ledgerwood, 2011; Kunda, 1990; Nosek, Spies, & Motyl, 2012). Thus, once you notice that an effect is significant when you analyze the data one way but not another way, your own mind can easily convince you that whichever test "worked" was the most appropriate test—perhaps even the test you intended to run all along. Recording a plan ahead of time allows you to clearly demarcate for yourself which analyses you planned and which were data-dependent, thus enabling you to accurately distinguish between findings that came from confirmatory versus exploratory analyses.

How to preregister a pre-analysis plan. There is no single correct way to record a pre-analysis plan, but a pre-analysis plan can only help you meet the goals described above if you specify your data-analytic choices (a) in enough detail that they effectively constrain any foreseeable researcher degrees of freedom available in your study and (b) with enough clarity

that a reviewer or reader can easily compare what you planned to do (as recorded in your pre-analysis plan) with what you actually did (as described in your paper). It is therefore crucial to think carefully about your options and to select one that works well for your research context.

Pre-analysis plans can be private or public and range from very basic to very detailed. For instance, one research team might decide to use an independent registry (e.g., AsPredicted.org, opensciencefframework.org, socialscienceregistry.org) to publicly preregister a detailed pre-analysis plan for each study they conduct.[2] Another team might develop an internal lab workflow in which they always record certain core elements of a study ahead of time (e.g., planned total sample size, planned exclusion criteria, and any planned confirmatory statistical tests) so that they can easily distinguish for themselves between exploratory and confirmatory findings. The most useful format and content of a pre-analysis plan will vary across research teams and projects, depending on the type of research, the complexity of the analyses, and the norms of a given field (Casey et al., 2012). Note that if you want to maximize transparency, it will be useful to (1) post your pre-analysis plan in a public repository (as opposed to keeping it private) and (2) make sure that it is easy for others to compare what you planned to do with what you actually did. Note too that a pre-analysis plan does not automatically prevent Type I error inflation: For example, one could record ahead of time a plan to examine all main effects and interactions in a three-way ANOVA, but without a plan to adjust for multiple comparisons, the Type I error rate for this pre-registered ANOVA would be quite high (Cramer et al., 2016).

The first time you create a pre-analysis plan, consider starting with something basic that you feel comfortable with, and build from there. Your main goals are to ensure that you will be

[2] Note that public pre-registration can have the added benefit of helping address publication bias, or the "file-drawer problem" in which nonsignificant results are never shared with the scientific community (Rosenthal, 1979; Simes, 1986), but only to the extent that (a) the preregistration can be easily found and understood by other researchers and (b) your results and/or data can be easily found as understood as well.

able to accurately distinguish between exploratory and confirmatory analyses, and that any decisions you make for your confirmatory analyses are independent of the data themselves. Figure 2 lists common examples of content that a researcher might specify in a pre-analysis plan. Labs may also find it helpful to register their "Standard Operating Procedures" for common decisions that are relevant across many of the studies that the lab tends to run (e.g., standard procedures for using attention checks on MTurk or for handling outliers in reaction time data; see Lin & Green, 2016).

Figure 2. Common Content for a Pre-Analysis Plan

Consider Specifying:	**Example:**
Planned sample size and stopping rule	Target total N = 100 We will collect data until MTurk indicates that we have completed surveys from 100 participants.
Inclusion criteria	MTurkers aged 18 and up will be allowed to participate.
Exclusion criteria	Participants will be excluded if they respond incorrectly to the attention check at the beginning of the study (i.e., if they do not select "Blue" when asked what color appears in the blue square). UPDATED 10/20/2017 after downloading the data but before running any analyses: We noticed that 3 participants completed the study in under 3 minutes whereas the average participant took 18 mins, so we decided to exclude these 3 participants.
Manipulation(s) and conditions	Group identity symbol (2 within-subjects conditions): Control (pictures of alien creatures grouped around a palm tree) vs. symbol (pictures of alien creatures grouped around a flag)
Predictor(s) and how they will be constructed	N/A
Dependent measure(s) and how they will be constructed	Primary/Focal DV: Perceived group entitativity (average of the six-item scale for each picture of alien creatures) Additional DV: Perceived threat (average of the two-item scale for each picture of alien creatures)
Any planned covariates	N/A
Planned statistical tests involving specific operational variables	Primary/Focal analysis: Paired t-test (two-tailed) examining the effect of condition on average perceived entitativity for control pictures vs. symbol pictures. Additional analysis: Paired t-test (two-tailed) examining the effect of condition on average perceived threat for control pictures vs. symbol pictures.

Any planned follow-up or subgroup analyses	No
Any plan for Type I error control (e.g., for multiple comparisons)	No

Note. Notice that pre-analysis plans must be specific to be useful: They must clearly constrain potential researcher degrees of freedom. For example, writing "participants have to be paying attention" does not clearly constrain flexibility in data analysis because there are multiple ways to decide whether participants were paying attention (e.g., passing a particular attention check vs. how long participants spent completing the survey vs. whether a participant clicked the same number for every item on a survey).

Gray areas in distinguishing between exploratory and confirmatory analyses. The distinction between exploratory and confirmatory analyses is simple in the abstract (are all researcher decisions data-independent or not?) but often complex and nuanced in practice. Most research involves a combination of planned and exploratory analyses; you might record a pre-analysis plan for a primary analysis and then also conduct a number of unplanned analyses to explore what else you can learn from your dataset. Exploratory analyses can range from highly principled and still fairly constrained (e.g., deciding to include a single, carefully chosen covariate in one's analysis to boost power after finding a nonsignificant key result; see Wang, Sparks, Gonzales, Hess, & Ledgerwood, 2017) to completely unconstrained (e.g., examining a giant correlation table to see if any interesting correlations pop out). And planned analyses can turn out to be inappropriate once you fully examine (and explore!) your actual data; for instance, you might plan to run a *t*-test on the amount participants choose to donate to charity, only to find that the actual distribution of this variable is binary rather than continuous. Always check the assumptions underlying your statistical tests and always graph your data (preferably in a way that allows you to see the distribution of the actual data points)—you should never blindly follow a pre-analysis plan or let it prevent you from also exploring what your actual data look like. Instead, use pre-analysis plans to help you (a) think carefully about various data collection and

analysis decisions, and (b) accurately distinguish and transparently report which of your analyses are data-independent.

Planning for Unexpected Results

Regardless of whether the study you are designing is exploratory or confirmatory in its predictions and its analyses, it is worth thinking carefully about how to maximize what you will learn from any possible pattern of results. For example, you may be designing a study to test the prediction that Manipulation X will increase Measure Y. But what will you learn if you find an unexpected effect in the opposite direction, or no effect at all?

Often, it is possible to design a study so that it pits two interesting predictions against each other—perhaps Theory 1 predicts that X will increase Y, whereas Theory 2 predict the opposite. Such an approach ensures that a difference in either direction will be interesting and informative. It can be helpful to graph or think through the various possible patterns of results you might see in a given study and ask what you would learn from each one. If most patterns of results would not be informative—if they would not lead you to update your beliefs in some way and/or point the way to the next study idea—then you may want to rethink your study design or the way that your variables are operationalized (i.e., manipulated or measured).

It can be more difficult—but just as important—to think about how to make a potential null result informative. One important tool is to maximize the statistical power of your study (see next section), so that a null result is less likely to reflect a Type II error (i.e., a false negative, or failing to detect an effect that is in fact there). Another useful tool is to assess the validity of the manipulation by including a manipulation check, either as the dependent measure in a pilot study or as an additional measure in the main study (van’t Veer & Giner-Sorolla, 2016). If Manipulation X affects the manipulation check but not the dependent variable Y, you could

reasonably infer that X does not affect Y. If, on the other hand, Manipulation X affects neither the manipulation check nor the dependent variable Y, you could instead infer that you were not successful at manipulating your construct of interest. The inclusion of a manipulation check allows you to pull apart these two possible explanations for a null effect of X on Y (Finkel, 2016). It can also be useful to think about why a particular manipulation might not work in a given sample (e.g., a well-established manipulation of cognitive load in the lab might not work for an online sample if online participants simply write down the long digit string they are asked to remember in the high load condition, or if online participants are highly distracted in all conditions; see also Zhou & Fishbach, 2016) and to include exploratory questions that could help assess that potential explanation.

In general, null results are more informative when the manipulations and measures used in a study have been carefully validated. Researches often think about the importance of using well validated measures in their research (although perhaps not as much as they should; see Flake, Pek, & Hehmann, 2017, for recommendations on best practices for validating measures). Ideally, researchers would also attend carefully to the construct validity of their manipulations, either by using previously validated manipulations in their studies or by conducting validation work themselves to ensure that a particular manipulation is successfully influencing only the intended construct of interest. Papers conducting such careful validation work for manipulations of common research constructs will hopefully become more prevalent in the future; arguably, they should be just as highly prized and cited as papers validating widely used measures.

Finally, whereas traditional statistical analyses using the most typical form of Null Hypothesis Significance Testing make it notoriously difficult for researchers to draw strong inferences from null results, other analytic approaches provide possible solutions. For instance,

equivalence testing allows researchers to conclude that two group means or two correlations are not meaningfully different from each other (Rogers, Howard, & Vessey, 1993; see Lakens, 2017, for concrete instructions). Meanwhile, Bayesian statistics allow researchers to test carefully specified null and alternative hypotheses (e.g., that a given effect size is $d = 0$ vs. $d = .30$) and to assess the extent to which the evidence favors one versus the other (Etz, Haaf, Rouder, & Vandekerckhove, in press).

Maximizing Statistical Power

Although statisticians have long emphasized the importance of power (e.g., Cohen, 1962; Maxwell, 2004; Rossi, 1997), many researchers only recently began appreciating how crucial it is to think about power in a careful way. Boosting statistical power helps to increase the informational value of an individual study, for reasons relating to both the likelihood of a Type II error (i.e., failing to detect a true effect) and the false positive rate (i.e., the proportion of effects in a set of significant findings that reflect spurious results rather than true effects). Furthermore, thinking carefully about the statistical power of a study—regardless of whether it is low or high—helps you draw better inferences from the results.

Many researchers think about power as being important for avoiding Type II errors: High power makes it more likely that an effect will be detected if it is there. This conceptualization of power can lead researchers to assume that although high power is desirable, low power is problematic only if you *fail* to see an effect. A researcher thinking about power in this way might conclude—understandably, but incorrectly—that if he runs an underpowered study and detects an effect, it represents especially trustworthy evidence for that effect ("I found it even with low power working against me!").

However, low statistical power also undermines researchers' ability to trust effects when they *do* see them. It turns out that reducing power also reduces the *positive predictive value* (PPV) of a significant finding (Button et al., 2013). PPV is the probability that a statistically significant result reflects a true positive (i.e., a real effect in the population). In other words, the PPV of all of your own findings in a given year would be the likelihood that any given significant effect that you detect in that year is real (that is, the proportion of all of your significant results in that year that are true positives). As the average power of your studies decreases, the number of true positives in your personal pool of significant results also decreases. The dwindling number of true positives means that the probability of any one of your significant results being true goes down.

Another way to think about underpowered studies is that they tend to produce "bouncy" results—effect size estimates will fluctuate more from one study to the next, and p-values will "dance" more dramatically (Cumming, 2012). That is, compared to the more precise and stable estimates provided by highly powered studies (see Ioannidis, 2008; Schönbrodt & Perugini, 2013), the estimates produced by underpowered research will tend to bounce more wildly from one study to the next. They will also tend to bounce more wildly from one subjective researcher decision to the next (e.g., decisions about which items to include in a scale or whether to exclude outliers). Low power therefore reduces the informational value of your results, and it can lead to problems later on when you or other researchers try to replicate your findings (Maxwell, 2004). In fact, when power drops below 50%, the average effect size estimate from significant studies starts to become dramatically inflated (because only the highly overestimated effect sizes will manage to hit significance), and when power drops below 10%, effect size estimates can actually be in the wrong direction (leading a researcher to conclude, for example, that an intervention

manipulation decreased a problem 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.

Taken together, these issues point to the crucial importance of estimating power when planning a study, so that you can (1) try to boost power when necessary and (2) acknowledge the uncertainty inherent in underpowered studies when achieving high power is not possible.

Estimating power. In order to calculate the approximate power of a planned study, you first have to estimate the effect size of interest. That might sound easy in theory, but there are a number of issues that can make it challenging in practice. First, if you are conducting the first study in a brand new line of basic research, you may have no idea what effect size to expect. Second, even if previous studies exist, they may not provide you with a good estimate. Unless sample sizes are very large, effect size estimates tend to fluctuate quite widely from one study to the next—which means that the correlation or mean difference observed in a single previous study or pilot study may be quite far away from the true population effect size. A useful heuristic for social-personality psychologists is that a total sample size of at least N = 240 provides a fairly stable estimate of a bivariate correlation or two-group mean difference that approximates the average effect size of r = .21 observed in social and personality psychology research (see Schönbrodt & Perugini, 2013). That means that the effect size point estimate from a Study 1 with N = 100 is probably not a good number to use when calculating power for Study 2 (instead, researchers should use the power-calibrated effect size approach; see McShane & Bockenholt, 2016, for a discussion and easy-to-use resources). Moreover, when combined with publication bias, this fluctuation of estimates from one study to the next means that published studies are likely to overestimate the size of an effect. If each study provides a guess about the

true population effect size, then some guesses are too high while some are too low; publication bias effectively truncates that distribution at the low end such that overestimated effect sizes are published while the underestimates that would have balanced them out end up in a file drawer.

Third and relatedly, although meta-analytic estimates of effect sizes can provide much more precise effect size estimates by aggregating across many individual studies, they can also be highly inflated due to publication bias. Such inflation is likely to occur 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). Because it is extremely difficult to adjust for publication bias in meta-analysis once the bias has occurred (see later section on Building Good Buildings), it is important to think carefully about whether a set of meta-analyzed effects was likely to be influenced by publication bias before assuming the resulting estimate is accurate. If publication bias is likely to be present, researchers may want to use methods for power analysis that help guard against the problem of overestimated effect sizes leading to underpowered subsequent studies; Anderson, Kelley, & Maxwell, 2017; Perugini, Gallucci, & Costantini, 2014). For example, Anderson et al. (2017) provide an R package and simple-to-use Shiny Web applications that help researchers easily calculate sample sizes for planned studies that account for both uncertainty and publication bias in published effect size estimates (see https://designingexperiments.com/shiny-r-web-apps).

Another option in such cases is to identify the smallest effect size of interest (sometimes abbreviated SESOI) and use that effect size in your power calculations—to say, in essence, that you only care about the effect if it is larger than size X. Indeed, this practice of defining a smallest effect size of interest is what researchers do implicitly when they decide to run a study with a particular sample size without conducting a power analysis. For instance, if you decide

that a given research question is worth the resources it would take to conduct a two-condition experiment with a total N of 80 participants, you effectively are deciding that you are only interested in the effect if it is at least d = .63 (the effect size that can be detected with 80% power in a sample this size). It is worth developing an intuition about effect sizes—a Cohen's d of .63 is about the size of the difference between men and women in sprinting speed (Thomas & French, 1985), and it is bigger than about 70% of the effects studied in social psychology (based on effect size estimates found in meta-analyses; Lovakov & Agadullina, 2017). If you conduct a power analysis, you can make this kind of implicit decision process an explicit one, allowing you to consider whether it is worth running the study (maybe not, if you suspect the effect you are interested in may be more subtle than an easy-to-casually-observe effect such as sex differences in sprinting speed), whether you want to spend more resources, and how confident you can be in your conclusions. Alternatively, if you can identify the smallest effect size that would be theoretically or practically meaningful, powering your study to detect such a SESOI increases the ultimate informational value of a null result.

Another important consideration when estimating power is to recognize that the sample size necessary to conduct a well powered test of an interaction is often dramatically larger than the sample size necessary to detect a main effect. G*Power, a commonly used software for power computations, can produce misleading estimates for powering interactions when researchers rely on the kinds of effect size estimates that are typical for main effects (e.g., conventions for "small" and "medium" effect sizes like d = .2 and d = .5, or the effect size estimate from a main effect observed in Study 1). Instead, try using the rules of thumb summarized in Table 1 (see Giner-Sorolla, 2018, for a full discussion).

Table 1. Rules of thumb for estimating the sample size needed to examine a 2x2 between-subjects interaction in Study 2 that qualifies a between-subjects main effect observed in Study 1

Expected interaction type	**Required cell size (n) to have same power as Study 1**	**Required total sample size (N) to have same power as Study 1**
Reversal of main effect	Same as n in Study 1	2 x N in Study 1
Elimination of main effect	2 x n in Study 1	4 x N in Study 1
50% attenuation of main effect	7 x n in Study 1	14 x N in Study 1

Note. In this example, Study 1 tested a main effect (a two-group experiment) and planned Study 2 will test a potential moderator of this main effect (a 2x2 between-subjects factorial design).

Sequential analysis. Because it is often difficult to estimate accurately the size of an expected effect (especially when conducting initial studies in new lines of research), and because even relatively small variations in expected effect sizes can lead to dramatically different answers about the sample size needed to achieve adequate power, it can be challenging to decide on a target sample size for a new study. Your study could be woefully underpowered if you guess an effect size that is too high, but you would waste substantial resources if you guess too low. In such cases, sequential analyses are a valuable tool that enables you to adequately power your study to detect a potentially small effect size, but stop early and conserve resources if the effect turns out to be larger than anticipated (Lakens & Evers, 2014; Proschan, Lan, & Wittes, 2006).

For instance, you might decide that it is worth spending the resources to power your study to detect an effect as small as d = .18 (the effect size that captures the average tendency for women to be higher in conscientiousness than men; Feingold, 1994; you would need n = 486 men and n = 486 women to have 80% power to detect this effect), but that you would rather stop early if at all possible so that you have resources left to run Study 2. Sequential analysis allows you to do this by choosing ahead of time a planned total sample size as well as specific interim analysis points (see https://osf.io/qtufw/ for a step-by-step guide and Lakens, 2014, for an

illustrative example and additional discussion).[3] Whereas unplanned optional stopping inflates Type I error (Sagarin, Ambler, & Lee, 2014), planned optional stopping in a sequential analysis controls Type I error by portioning out the total desired alpha level (often .05) across interim and final analyses.

Beyond sample size: Other tools for boosting power. Large samples are one route to high power, but they are not the only one, nor necessarily the most efficient in some contexts. Understanding the other factors that affect power can provide you with a toolbox of different strategies for conducting well-powered research. For instance, when they are feasible, within-subjects designs can dramatically boost the power of an experiment, relative to between-subjects designs (Greenwald, 1976; Rivers & Sherman, 2018). Also, when conducting basic research, it may be possible to increase the size of the effect that you are studying (e.g., by using extreme groups or developing a stronger manipulation). Likewise, reducing measurement error by finding or creating more reliable measures of your constructs can substantially boost power, particularly in studies that examine small effects (Ledgerwood & Shrout, 2011; Stanley & Spence, 2014). Furthermore, when designing an experiment, it is worth thinking carefully about a potential covariate that could correlate strongly with the dependent variable. Such a covariate can boost power by soaking up noise in your dependent measure (Wang et al., 2017). For example, a researcher interested in a manipulation that could influence participants' nationalism might decide to include a measure of social dominance orientation (SDO) at the beginning of her experiment as a preregistered covariate, given that SDO and nationalism show strong correlations in large samples (Pratto, Sidanius, Stallworth, & Malle, 1994).

[3] Note that because sequential analyses can lead to overestimated effect sizes (Zhang et al., 2012), they are best suited for studies in which researchers are primarily interested in testing whether an effect exists rather than determining a precise estimate of the effect size itself (Lakens, 2014).

Aggregating across small studies can provide another useful tool for boosting statistical power. Researchers limited in the number of participants they can recruit in a particular time frame (e.g., an academic term) or setting (e.g., a public event) might choose to run two or more separate, small studies testing the same effect that, when aggregated, achieve adequate power. Likewise, researchers interested in conducting systematic replications to test the generalizability of an effect across various samples and/or stimuli might wish to conduct a series of smaller studies that vary these elements (rather than a single large study with only one sample and stimulus set) and then meta-analyze the studies to provide a better powered test (and more precise estimate) of the effect. Similarly, researchers who are constrained in their access to participants or research funds can conduct a multi-lab collaboration in which two or more research teams conduct a study using identical protocols and then meta-analyze the results. Importantly, because each individual study will be underpowered in such cases, researchers should plan to conduct and interpret analyses only at the meta-analytic level.

Discussing power in a manuscript. If you conduct a power analysis to determine the target sample size for your study, make sure to record the software you used, the specific effect size you estimated and where the estimate came from, and any other information that you will want to eventually report when writing up the study methods. If you do not conduct a power analysis, you should still consider discussing power in the study manuscript to help readers calibrate their conclusions appropriately (e.g., a null result is not very informative if a study is likely to be underpowered). In such cases, it may be helpful to conduct and report a *sensitivity power analysis*, which calculates the minimum effect size that you were powered to detect given your sample size and a particular level of power (e.g., 80%). It can also be useful to graph the relation between effect size and power given your actual sample size. The goal is to give readers

(and yourself!) an intuitive sense of the size of the effect that you could reasonably expect to detect in your study (see Box 1 for examples).[4]

Box 1. Examples of how to discuss power in the method section of a manuscript

Example 1: A priori power analysis
An *a priori* power analysis in R using the meta-analytic effect size estimate for the correlation between X and Y across all ten previous studies conducted by our lab (r = .3) indicated we would need a total sample size of N = 82 to achieve 80% power in our new study. Given that we had the resources to collect a larger sample and that we expected we would have to drop a few participants based on our *a priori* exclusion criteria, we decided to recruit at least 100 participants.

Example 2: Sensitivity analysis
We decided to collect data until the end of the semester, which resulted in a total sample size of N = 140 (n = 70 per condition). A sensitivity power analysis in G*Power indicated that a sample of this size would provide 80% power to detect an effect of Cohen's d = .48 and 60% power to detect an effect of Cohen's d = .38. For reference, the estimated median effect size in social psychological research is about d = .38 (Lovokov & Agadullina, 2018).

Example 3: Estimated power for a range of effect size estimates
We set an *a priori* target sample size of N = 200 (n = 100 per condition). A recent meta-analysis suggests that the effect of X on Y is in the range of d = .27 to d = .38 (Someone & Whosit, 2018, Table 3). Power analyses in G*Power indicated that a sample of N = 200 provided 48% power to detect an effect as small as d = .27 and 76% power to detect an effect as large as d = .38.

Conducting Programmatic Research: Direct, Systematic, and Conceptual Replications

Increasing the power of a study helps make the results more informative, but at the end of the day, it is still a single study. It can be useful to think of a single study as a data point that can contribute to a cumulative understanding of a phenomenon, to the extent that it is well designed and adequately powered, rather than something that can provide a definitive conclusion in isolation (Braver, Thoemmes, & Rosenthal, 2014; Ledgerwood & Sherman, 2012; Soderberg & Errington, 2018). Obviously, you want each study or data point to be as informative as possible,

[4] Do not calculate or report post-hoc (sometimes called "achieved" or "observed") power in a study using the effect size estimate from that same study; it is redundant with the *p*-value and can be extremely misleading (Hoenig & Heisey, 2001).

but you may also want to accumulate multiple data points that can together provide a more substantial contribution to a given research question. When thinking about the best way to construct 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 topic.

Direct replications (sometimes called "close" or "exact" replications) aim to repeat as closely as possible the procedures used in a prior study (Fabrigar & Wegener, 2016; Schmidt, 2009). Direct replications serve to increase confidence in an observed relationship between two or more operationalizations (i.e., the specific manipulations and/or measures used in a previous study). For instance, if exploratory analyses in a first study provide suggestive evidence for a particular pattern of results (e.g., you observed an interesting effect of your manipulation on your dependent measure, but only after an unanticipated change to your pre-analysis plan), conducting a direct replication would provide an opportunity to corroborate that pattern in an independent dataset. Suppose that in Study 1, a research team tests whether MTurk participants judge an ambiguous face as more threatening after seeing black versus white faces (Kreiglmeyer & Sherman, 2012). A direct replication would involve the same operationalizations of race (black faces vs. white faces) and threat judgments (responses to an ambiguous face), as well as the same stimuli and population, in order to assess the robustness of the relation between operationalizations observed in Study 1. In sum, if you wish to increase your confidence in a particular result, direct replication is often a useful next step.

Systematic replications aim to vary presumably incidental aspects of the context in which a result was initially obtained, in order to test the critical assumption that those details are in fact irrelevant to the result (Kantowitz, Roediger, & Elmes, 2014). Systematic replications serve to increase confidence in the generalizability of an observed relation between particular

operationalizations. For instance, if a research team found an interesting effect of showing participants black versus white faces on threat judgments in Study 1, they might want to conduct a systematic replication using different stimuli in Study 2 in order to test whether Study 1's results were specific to a particular stimulus set (e.g., the particular black and white faces used in Study 1; Roediger, 2012; Westfall, Judd, & Kenny, 2015). Or, the research team might want to conduct a systematic replication that uses a different cover story and experimenter in order to test the generalizability of the results across different contextual details. If they observe an interesting effect in a sample of students at their own university, they might want to use StudySwap (2018) to systematically replicate the study at a different university or with older participants. An especially useful tool for thinking about and fostering systematic replications is to write a "Constraints on Generality" statement that explicitly articulates the extent to which you expect a set of finding to generalize across different stimuli, samples, and situations (Simons, Shoda, & Lindsay, 2017). Importantly, systematic replications help researchers combat confirmation bias in their research process by pushing them to explicitly consider and test whether variables presumed irrelevant for producing an effect might in fact be relevant. They 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 employed 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

similar patterns of findings, a researcher can be more confident that the results reflect something about the theoretical construct rather than the particular operationalization used to manipulate or assess it (Brewer & Crano, 2014; Cook & Campbell, 1979). For example, if our research team wanted to conduct a conceptual replication of their Study 1 finding that participants judge ambiguous faces as more threatening after seeing black (vs. white) faces, they might change the operationalizations of their conceptual independent and dependent variables by examining whether participants sit further away from a black (vs. white) confederate when waiting late at night in a waiting room. Conceptual replications are therefore useful when researchers are confident about the presence of a particular pattern of results between operational variables, but unsure if those operational variables are really tapping the theoretical constructs of interest. If the theoretical predictions hold up across a range of operationalizations, then researchers can be more confident that they are learning about the underlying concepts and theory rather than a specific instance of an effect (Crandall & Sherman, 2016; Fabrigar & Wegener, 2016).

Regardless of which type of replication you are conducting, it is important to appreciate how widely results can fluctuate from one study to the next due to chance, especially with imperfect measures and small samples (see Cumming, 2009; Stanley & Spence, 2014, for useful illustrations). When conducting a series of studies on a given research question, it is often best to take a cumulative approach that aggregates across studies rather than bean-counting each one in isolation as a “success” or “failure” (Braver et al., 2014; Fabrigar & Wegener, 2016). For instance, suppose you are embarking on a new line of basic research. You conduct a confirmatory (planned) analysis of a particular effect in three independent datasets, and find a significant result in one dataset and a non-significant result in the other two. The best approach to understanding these data will often 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 "succeeded" and the other two "failed." It can be tempting to think that some aspect of the method or sample of the significant study must have led it to "work" better than the other two, but unless the studies are very highly powered, a likely explanation for variability across study results is often chance fluctuation (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 to attain 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 use different sets of stimuli, 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 dataset 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.

Open and Transparent Science: Sharing Materials, Data, and Code

Another useful tool for improving the informational value of your research is to make your materials, data, and syntax as easily findable and understandable as possible. Doing so improves the study's informational value in a number of ways, including enabling other

researchers to easily use (and cite!) your materials in their own research, enabling readers of your manuscript to better understand the details of your study and the data behind your conclusions, and enabling scholars to find and include your data in a meta-analysis.

When seeking to enhance the openness and transparency of your research, several considerations merit careful attention (see Levenstein & Lyle, 2018; Meyer, 2018; for particularly helpful resources). For example, many kinds of data can in principle be shared with the public, but ethical and legal constraints can arise either due to unintended stumbling blocks (e.g., not using the right language in your consent form and IRB application; see Meyer, 2018, for how to avoid this problem) or due to the sensitive or confidential nature of some data (e.g., dyadic data where one partner could identify the other based on his/her responses, identifiable video recordings, etc.; see Gilmore, Kennedy, & Adolph, 2018; Joel, Eastwick, & Finkel, 2018; Levenstein & Lyle, 2018 for helpful resources). Researchers must also decide where to share study information among various options, including personal or institutional websites, in supplementary materials linked to a particular article, in a data repository such as Dryad (https://datadryad.org), or using an open-science service like OSF (https://osf.io/). These options vary in their ability to address different goals you may have, such as maximizing accessibility for various audiences, protecting the privacy of your participants when working with sensitive data, maximizing the likelihood that future meta-analysts or researchers in this topic area will find your study information, and/or streamlining efficiency within your lab.

One especially important issue to consider is whether the information you are sharing will be easily understandable by someone other than you. Uploading a set of files to an online site is easy, but it may not actually be transparent: The information you are trying to share needs to be findable, accessible, interoperable, and reusable by other people (Wilkinson et al., 2016).

Are your materials in a file format that can be opened easily by other researchers, or are they saved in a program that few other scholars would have? Are the variables in your data file carefully labeled? Have you included a codebook that clearly explains any information secondary users will need to understand and use the data file? Is your syntax clearly annotated? Levenstein and Lyle (2018) review resources for data management practices that you can incorporate as a standard part of your workflow to enable transparent sharing.

Building Good Buildings: How to Maximize What We Learn from Research Synthesis

Ultimately, no matter how carefully a study is designed, it is only one study—one data point. In isolation, a single study always has limitations—for instance, it often contains only one possible set of operationalizations, in one particular kind of sample, in one particular kind of context. Research is cumulative; it involves incrementally adding individual studies together to build an increasingly clear picture of a given phenomenon or process. We want each individual study or data point to be as solid as possible, but we need multiple data points (including direct, systematic, and conceptual replications) to start to shed light on a research question.

Integrating or aggregating across findings is therefore an essential part of the research process. To integrate well, we need (1) solid individual data points (2) an unbiased set of data points, and (3) a good way of synthesizing the data points.

The second of these elements may in fact be the most challenging: One of the biggest threats to cumulative research is arguably publication bias (Fanelli, 2011; Ferguson & Heene, 2012; Dickersin, 1990; Sterling, 1959). There is considerable pressure throughout the scientific system (including authors, reviewers, and journal editors) to publish positive results while relegating negative results to a file drawer. Some of this pressure comes from the fact that a null finding is often difficult to interpret—does it reflect a Type II error, a failure of the manipulation

to influence the construct of interest, a true null effect (or an effect too small to matter), or something else? Thus, one important strategy for combating publication bias is to design studies that are informative regardless of how the results turn out (see section above on planning for unexpected results). But much of the pressure to publish only positive results comes from entrenched biases in the scientific publication system and incentive structure (Fanelli, 2010; Song, Eastwood, Gilbody, Duley, & Sutton, 2000). Moreover, bias against null findings can affect the research and publication process at multiple stages and in different ways (Greenwald, 1975).

Publication bias is therefore extremely difficult to model accurately (McShane, Bockenholt, & Hansen, 2016). Perhaps at a given time point in the field's history, a researcher would be twice as likely to write up a finding if the *p*-value is significant rather than "marginal," and ten times as likely to write it up if the finding is "marginal" than if $p > .10$. But those values might change depending on whether this is a first study or a third study in a line of work, and depending on whether there are other results in the same study that support a given conclusion. Furthermore, the values might change as prevailing norms about research practices change across time. Estimating these relative likelihoods of publication (rather than guessing at them) in any given research area would require a very large number of studies.

The prevalence and complexity of publication bias poses a challenge for cumulative science for two reasons (McShane et al., 2016). First, once publication bias contaminates a literature, it is extraordinarily difficult to figure out (and then to model accurately) how exactly the bias has operated. Second, meta-analytic techniques are very sensitive to the assumptions made by a given model of publication bias. In other words, if a meta-analyst does not use a model of publication bias that accurately captures the way publication bias actually operated on a

given set of results, the meta-analytic estimates will often do a poor job of recovering the true population parameters of interest.

There are many things that researchers can do to mitigate the problem of publication bias. For example, meta-analysts can avoid drawing strong conclusions about an overall average effect size, focusing instead on moderator analyses that are less likely to be affected by publication bias in a particular area (e.g., study-level moderators that would not have been considered in publication decisions about individual papers). Likewise, meta-analyses that aggregate ancillary findings (rather than primary findings that would have provided the basis for a decision about whether to publish a given paper) are less likely to be plagued by the problems of publication bias. The remainder of this chapter reviews additional tools that have featured prominently in recent discussions about improving research synthesis.

Within-Paper Meta-Analysis

One way to avoid the problem of publication bias is to meta-analyze all of the studies you have conducted in a given line of research. For example, if you conduct six studies testing a particular effect, and find a significant result in four out of the six studies, you could report all of the studies in a manuscript along with a within-paper meta-analysis that aggregates across the individual studies to provide the best estimate of the effect(s) of interest. Such within-paper meta-analyses can help researchers, reviewers, and editors move away from thinking about each individual study in isolation (e.g., “this study worked, but that study didn’t”) and toward focusing on cumulative, meta-analytic estimates that provide more stable, precise, and useful information about a research question. Increasingly, resources are available that enable researchers to easily conduct within-paper meta-analyses for a range of research designs (including an online application that allows researchers to synthesize data across studies with

different designs; McShane & Bockenholt, 2017; see also Braver, Thoemmes, & Rosenthal, 2014).[5] When you report a within-paper meta-analysis, you should transparently disclose whether you included all of the studies you conducted to test the research question (and if not, why not), since the informational value of such an approach depends on your ability to fully "empty the file drawer" and include all relevant studies.

Adjusting for Publication Bias in Meta-Analysis: Using Selection Methods to Generate a Range of Plausible Estimates

When you do not have full access to the file drawer (either because you are meta-analyzing other researchers' results or because you do not necessarily remember and have not systematically archived all the studies you have conducted on a given research question), you have to guess how publication bias may have operated on the set of studies you want to meta-analyze. This guess is captured by how you choose to model publication bias. (Note that "publication bias" in the context of meta-analysis means any bias in the set of studies that are *available for and included in* the meta-analysis; this set of studies may include some unpublished studies as well as the published ones.) For example, a classic fixed-effect or random-effects meta-analysis assumes no publication bias has occurred (and therefore gives upwardly biased effect size estimates in the common situation when publication bias does in fact characterize a given set of studies).[6] Hedges (1984) selection method, *p*-curve (Simonsohn, Nelson, & Simmons, 2014), and *p*-uniform (van Assen, van Aert, & Wicherts, 2015) all model a very simple form of publication bias: They assume that all statistically significant results are published (i.e., available for and included in the meta-analysis) and that no statistically

[5] Online app, example, and tutorials available at https://blakemcshane.shinyapps.io/spmeta

[6] When publication bias is present, regular meta-analysis also gives a false impression of effect size homogeneity, leading researchers to incorrectly conclude that effect size heterogeneity is not an issue when in fact it is.

nonsignificant results are published (i.e., available for and included in the meta-analysis).[7] Other selection methods model more complex forms of publication bias by assuming that both significant and nonsignificant results are published but with different likelihoods (e.g., Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995)—for example, one might imagine that the likelihood that a statistically nonsignificant result is published would increase as the *p*-value approaches significance (think, for example, of the common terms "marginally significant" or "approaching significance" or "trending in the expected direction").

Again, because many factors can influence whether a given result makes it into the published literature and/or the set of unpublished studies to which you have access, any one model of publication bias is unlikely to be accurate; moreover, even small variations in the choice of how to model publication bias can lead a meta-analysis to produce remarkably different estimates of an effect. Thus, if it is likely that a given set of results has been filtered by some form of publication bias, the most fruitful approach to meta-analysis may be to use a variety of selection methods—that is, to model a range of plausible forms of publication bias in order to generate a range of estimates that are consistent with the data under different reasonable assumptions (e.g., Inzlicht, Gervais, & Berkman, 2015; see McShane et al., 2016, for concrete resources). If these estimates vary considerably, it suggests that the results are telling you more about the assumptions of publication bias made by each model than about a true underlying effect. If the estimates tend to be consistent, you can be more confident that the different models are converging on useful information about the true underlying effect.

Conclusion

[7] These methods also either ignore effect size heterogeneity or assume that the researcher is not interested in generalizing beyond a specific set of studies to make inferences about a larger population (the goal of a random-effects meta-analysis); when these assumptions are violated, they perform poorly (McShane et al., 2016; van Aert, Wicherts, & van Assen, 2016).

Over the last decade, psychological science has emerged as a leader in the push to improve research methods and practices across scientific disciplines. Researchers are increasingly implementing the cutting edge methodological and statistical approaches outlined above. Meanwhile, the field continues to identify and hone additional tools for improving scientific methods and practices. By understanding the basic concepts underlying these tools and applying them to your own research, you can both maximize the informational value of each study you conduct as well as the picture that emerges from synthesizing across multiple findings, thereby helping to improve the quality of our cumulative and collaborative science.

References

Bandalos, D. L. (2018). Measurement Theory and Applications for the Social Sciences.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531-533.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407-425.

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

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: Cambridge University Press.

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*(5), 365-376.

Casey, K., Glennerster, R., & Miguel, E. (2011). *Reshaping institutions: Evidence on aid impacts using a pre-analysis plan* (No. w17012). National Bureau of Economic Research .

Chaiken, S., & Ledgerwood, A. (2011). A theory of heuristic and systematic information processing. *Handbook of theories of social psychology: Volume one*, 246-166.

Cohen J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153.

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

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

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93-99.

Cumming, G. (2009). Dance of the *p* values. Accessed 1/28/2018 from https://www.youtube.com/watch?v=ez4DgdurRPg

de Groot, A. D. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. Van der Maas]. *Acta Psychologica, 148*, 188-194.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, *263*(10), 1385-1389.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (in press). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science.*

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68-80.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PloS one*, *5*(4), e10271.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90, 891–904.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences, 115*, 2628-2631.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*(3), 429-456.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555-561.

Finkel, E. J. (2016). Reflections on the commitment-forgiveness registered replication report. *Perspectives on Psychological Science, 11,* 765-767.

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*(2), 275-297.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology, 113*, 244-253.

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

Forscher, B. K. (1963). Chaos in the brickyard. *Science*, *142*(3590), 339.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641-651.

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

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328-331.

Gerber, J. P., Wheeler, L., & Suls, J. (2017, November 16). A social comparison theory meta-analysis 60+ years on. *Psychological Bulletin*. Advance online publication.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement Theory for the Behavioral Sciences: Origin & Evolution*. WH Freeman & Company.

Gilmore, R. O., Kennedy, J. L., & Adolph, K. E. (2017). Practical solutions for sharing data and materials from psychological research. *Advances in Methods and Practices in Psychological Science*, 2515245917746500.

Giner-Sorolla, 2018: https://www.journals.elsevier.com/journal-of-experimental-social-psychology/news/announcement-of-new-policies-for-2018-at-jesp

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*(2), 314-320.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61-85.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*(1), 19-24.

Inzlicht, M., Gervais, W., & Berkman, E. (2015). *Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough*.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640-648.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 109-117.

Joel, S., Eastwick, P. W., & Finkel, E. J. (2018). Open sharing of data on close relationships and other sensitive social psychological topics: Challenges, tools, and future directions. *Advances in Methods and Practices in Psychological Science*, 2515245917744281.

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*(1), 54-69.

Kantowitz, B., Roediger III, H., & Elmes, D. (2014). *Experimental psychology*. Nelson Education.

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One*, *10*(8), e0132382.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480-498.

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

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

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*(3), 278-292.

Ledgerwood, A. (2014). Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, *9*(3), 275-277.

Ledgerwood, A. (2016). Introduction to the special section on improving research practices: Thinking deeply across the research cycle. *Perspectives on Psychological Science, 11*(5), 661-663.

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science, 7*(1), 60-66.

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*(6), 1174-1188.

Ledgerwood, A., Soderberg, C. K., & 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*, 33-58. Washington, DC, US: American Psychological Association.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255-275.

Levenstein, M. C., & Lyle, J. A. (2018). Data: Sharing is caring. *Advances in Methods and Practices in Psychological Science*, 2515245918758319.

Lin, W., & Green, D. P. (2016). Standard operating procedures: A safety net for pre-analysis plans. *PS: Political Science & Politics*, *49*(3), 495-500.

Lovakov, A., & Agadullina, E. (2017). Empirically derived guidelines for interpreting effect size in social psychology. *PsyArXiv*.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3-11.

Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science, 58*, 523-552.

McNutt, M. (2014). Journals unite for reproducibility. *Science, 346*(6210), 679.

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods, 21*, 47.

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

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.

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

Mills, J. L. (1993). Data torturing. *The New England Journal of Medicine, 329*, 1196-1199.

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*(6), 615-631.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 201708274.

Nyhan, B. (2015). Increasing the credibility of political science research: A proposal for journal reforms. *PS: Political Science & Politics*, *48*(S1), 78-83.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. Perspectives on Psychological Science, 9, 319-332.

Poincaré, H. (1902). *La science et l'hypothese*. Flammarion.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67*, 741.

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

Rivers, A. M., & Sherman, J. (2018, January 19). Experimental Design and the Reliability of Priming Effects: Reconsidering the "Train Wreck". *PsyArXiv*.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*, 9.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553-565.

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

Ross, M., & Wilson, A. E. (2000). Constructing and appraising past selves. *Memory, Brain, and Belief*. (pp. 231-258).

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 175–197). Mahwah, NJ: Erlbaum.

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology, 21*, 308.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 293-304.

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods*, *19*(3), 317-333.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609-612.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American psychologist*, *54*(2), 93-105.

Schwarz, N. (2016). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*.

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.

Simes, R. J. (1986). Publication bias: the case for an international registry of clinical trials. *Journal of Clinical Oncology*, *4*(10), 1529-1541.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.

Simons, D. J. (in press). Editorial. *Advances in Methods and Practices in Psychological Science*..

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123-1128.

Soderberg, C. K., & Errington, T. M. (in press). Replications and the social sciences. In J. Edlund & A. L. Nichols (Eds), *Advanced Research Methods and Statistics for the Behavioral and Social Sciences*. Cambridge University Press.

Song, F., Eastwood, A., Gilbody, S., Duley, L., & Sutton, A. (2000). Publication and related biases: a review. *Health Technology Assessment*, *4*(10).

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886-899.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305-318.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30-34.

StudySwap: A platform for interlab replication, collaboration, and research resource exchange. (2018, February 25). Retrieved from https://osf.io/view/studyswap/

Thomas, J. R., & French, K. E. (1985). Gender differences across age in motor performance: A meta-analysis. *Psychological Bulletin*, *98*(2), 260-282.

Tilburg University (2011). *Interim report regarding the breach of scientific integrity committed by prof. D.A. Stapel*. Tilburg University, 1-21.

van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on *p* values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*, 713-729.

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12.

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

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419-435.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274-290.

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

Welkowitz, J., Cohen, B. H., & Lea, R. B. (2012). *Introductory Statistics for the Behavioral Sciences*. John Wiley & Sons.

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*(3), 390-399.

Wilkinson M. D., Dumontier M., Aalbersberg IJ. J., Appleton G., Axton M., Baak A., , . . . Mons B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3,* Article 160018. doi:10.1038/sdata.2016.18

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.

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, 111,* 493-504.