Chapter 5
The Myriad Forms of *p*-Hacking

Dorota Reis and Malte Friese

Abstract In the present chapter, we are going to discuss several *p*-hacking practices as part of the broader category of questionable research practices. It has become clear that *p*-hacking can have detrimental consequences—particularly an increase in false-positive rates—that ultimately damage the trustworthiness and robustness of psychological science. What can any researchers do to confirm that they did not engage in questionable research practices? The solution is surprisingly simple. It lies in the transparent distinction between a priori planned, confirmatory steps of data analysis and exploratory, additional steps. The line between the two can be drawn easily by adhering to the open science practices outlined in this chapter, particularly the detailed preregistration of all measures, manipulations, hypotheses, and planned analysis steps. Open science practices are surely not the solution to all challenges psychological science currently faces, but they are a pretty good and easy-to-implement solution to prevent *p*-hacking. Let's do it.

Keywords Meta science · *p*-hacking · Questionable research practices

Credibility Concerns about (Clinical) Psychological Science

Alice is an experienced psychotherapist. For many years, she has worked in an outpatient facility specializing in the treatment of chronic pain. Being a passionate practitioner, Alice is continuously educating herself on how to use state-of-the-art treatment methods to best benefit her patients. Therefore, Alice is thrilled when she reads about a new therapy and its impressive treatment response in a prestigious scientific clinical journal. "With this new approach," Alice feels, "I will be able to

Dorota Reis and Malte Friese are contributed equally to this work.

D. Reis (✉) · M. Friese (✉)
Saarland University, Saarbrücken, Germany
e-mail: dorota.reis@uni-saarland.de; malte.friese@uni-saarland.de

© Springer Nature Switzerland AG 2022

W. O'Donohue et al. (eds.), *Avoiding Questionable Research Practices in Applied Psychology*, https://doi.org/10.1007/978-3-031-04968-2_5

have a substantial additional impact on the well-being of my patients!" She invests time and money to be certified as a specialist in this new approach and starts implementing the new intervention strategy in her clinical work.

A few months later, Alice receives the treatment evaluations. They are sobering. Although she closely adhered to the therapy manual, the desired reduction of symptoms remains far behind her expectations. Even more, the new therapy appears to be less effective than the conventional "gold standard" treatment previously applied in the facility. Although the evaluations confirm the subjective impressions she obtained during the therapy sessions and match those reported by colleagues who have also implemented the new technique, Alice is frustrated. Instead of improving the treatment for her patients, the changes to the protocol appear to have backfired. The success rate even falls below that of previous treatments. Some patients begin dropping out early, whereas others begin to take even longer than before to attain noticeable treatment results. What happened here?

In the last decade, scientific psychology has seen a multitude of scenarios similar to the one described in the opening paragraphs. Large-scale replication projects (Klein et al., 2014; Open Science Collaboration, 2015, see also Chap. 18, this volume) and countless primary studies have shown disturbingly low replication rates (see also Chap. 4, this volume). Psychology is not alone. Other disciplines have reported similar problems, including the neurosciences (Button et al., 2013), economics (Camerer et al., 2016), cancer research (Begley & Ellis, 2012), and drug research (Prinz et al., 2011), to just name a few. Although some disciplines are more affected than others, low replicability appears to be a problem in many fields.

In psychology, what began as a "replication crisis" has quickly become a more general "credibility crisis." As a result, psychological science is under scrutiny (Lilienfeld & Waldman, 2017). This is not just academic ivory tower talk. The credibility of psychological research has profound real-world consequences. Particularly in clinical research and practice, unreliable findings can affect the (mental) health of people who rely on the trustworthiness of the science that informed their treatments. Interventions that were believed to be effective but actually are not imply that patients and clients will experience less symptom reduction and need more time to reduce distress than necessary.

Various issues have been discussed as undermining the credibility of psychological science. These include low statistical power (Bertamini & Munafò, 2012), an over-reliance on *p*-values (Wasserstein & Lazar, 2016, see also Chap. 7, this volume), maladaptive incentives (Lilienfeld, 2017; Nosek et al., 2012), hypothesizing after the results are known (HARKing, Kerr, 1998, see also Chap. 8, this volume), publication bias (Bakker et al., 2012, see also Chap. 10, this volume), and *p*-hacking (John et al., 2012; Simmons et al., 2011), among others.

We cannot know why the new treatment that Alice was so enthusiastic about in our fictitious introductory example was less effective in Alice's facility than what seemed to be realistic on the basis of the associated scientific publication. In the present chapter, we will focus on one specific issue that jeopardizes the credibility and robustness of psychological science: *p*-hacking. We will discuss what *p*-hacking is, which scientific practices are subsumed under this umbrella term, its prevalence

and detection (see also Chap. 6, this volume), its consequences, and how it can be prevented. We will close by making a case for open science practices that we argue are an effective remedy for several of the challenges that scientific psychology currently faces.

If you are a clinical psychologist, you may wonder why you should go on reading. Isn't the replication crisis and the associated use of problematic research practices something for other psychological subdisciplines to worry about? It is true that the extent of replicability problems and the use of problematic research practices vary across subdisciplines (John et al., 2012), but this does not mean that clinical psychology is free of concerns (Leichsenring et al., 2017; Tackett et al., 2019). In fact, there are a few weak spots that endanger the replicability and robustness of clinical research as well. (For an overview of research biases in psychotherapy research, in particular, see Leichsenring et al., 2017.)

For example, statistical power is often low, particularly for treatment/intervention research and clinical neuroscience (Button et al., 2013; Cuijpers, 2016; Reardon et al., 2019; Sakaluk et al., 2019). One reason for low power is small sample sizes. As later sections of this chapter will clarify, some *p*-hacking practices are particularly "effective" in small samples, making such studies vulnerable to considerable bias. In addition, in a scientific culture that values novel, statistically significant findings so much more than less novel and/or statistically nonsignificant findings, incentives to "find" something in the data of a given study are high, and this is particularly true when the study cannot be easily repeated or extended because it is very resource-intensive or relies on a difficult-to-reach sample. This applies to a lot of clinical psychology intervention studies, which is one reason why a lot of important studies cannot be easily replicated in single studies or in large-scale replication projects (Tackett & Miller, 2019). Based on novel evidential value metrics, such as rates of misreported statistics, power, and Bayes Factors, the replicability of empirically supported treatments seems to be remarkably low (Sakaluk et al., 2019). Thus, it cannot be assumed that actual therapy results will achieve the effectiveness that could be assumed on the basis of the scientific articles the interventions were published in. On the basis of their analysis, Sakaluk et al. (2019) concluded that whereas there is strong evidence behind a few therapies, "the evidence is mixed or consistently weak for many, including some classified by Division 12 of the APA as 'Strong'" (p. 500). Finally, there is evidence for considerable publication bias (also) in clinical psychology, another factor that undermines the robustness of published findings (Cuijpers et al., 2010; Rapport et al., 2013). Thus, yes, we are afraid the credibility crisis in general, and *p*-hacking in particular, are topics that should also be of interest to clinical psychologists.[1]

[1] We were initially invited to contribute a chapter that focuses on the implications of *p*-hacking for clinical psychology specifically. This is why most of our examples in this chapter come from this area of research. However, the general issues covered in this chapter also apply to other areas of (applied) psychology.

What Is *p*-Hacking?

When researchers collect and analyze data, they have many decisions to make. Unless they commit a priori to a specified and exhaustive set of decision outcomes, they have many so-called researchers' degrees of freedom (Wicherts et al., 2016). These degrees of freedom invite *p*-hacking. The term *p*-hacking (also called data dredging or inflation bias) refers to a family of practices that can be responsible for substantial biases. It is an umbrella term that "refers to nonprincipled decisions during data analysis that are aimed at reducing the *p*-value of a significance test and thus make the data look more robust than they actually are" (Friese & Frankenbach, 2020, p. 457). Researchers have known about such nonprincipled decisions for a long time: As early as 1956, De Groot described this approach as an "attempt to let the material speak [that] leads to ad hoc decisions in terms of processing" (de Groot, 1956/2014, p. 191). *p*-Hacking can take various forms (also known as "*p*-hacks") and can occur at various points during the data analysis process, even before formal data analysis has even begun (see Fig. 5.1 for a schematic overview). The main characteristic of these practices is that they are selectively employed to bring an originally nonsignificant *p*-value below the alpha level of 5%.

p-Hacking can occur intentionally and with full awareness that one is not following best scientific practices. However, importantly, *p*-hacking can also occur largely without awareness of the potentially detrimental consequences or even with genuinely honest intentions. To illustrate, we believe it is quite likely that before the seminal paper by Simmons et al. (2011) demonstrated the tremendous effect of *p*-hacking on false-positive rates, many researchers were aware that playing with the data too much would likely increase the chances of false positives, but they probably had little idea about the extent of this problem. Even more, researchers who are convinced of the validity of a particular hypothesis may want to uncover what they think is hidden like a precious, but hard-to-detect signal in the data (Heene & Ferguson, 2017; Nelson et al., 2018). Because of confirmation and hindsight bias,

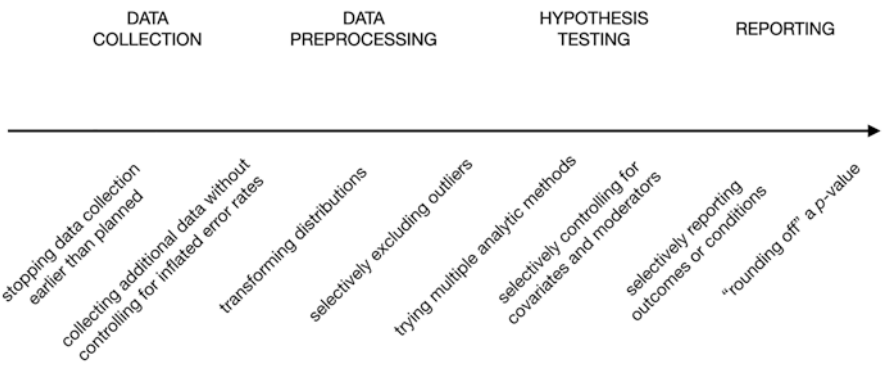

Fig. 5.1 Schematic depiction of exemplary *p*-hacking practices according to when they typically occur during the research process

researchers may believe that undesired outcomes result from suboptimal analyses (Munafò et al., 2017), so they try to optimize their analyses without any bad intent. Thus, the general notion of the present chapter is not to blame or denounce researchers for their presumably ill-intentioned behavior. Instead, we intend to provide information about the nature, consequences, and prevention of *p*-hacking, whether it occurs intentionally or not.

p-Hacking Practices

Various different practices are considered *p*-hacking. Our overview is not exhaustive. In Fig. 5.1, we arranged some exemplary strategies according to when they typically occur in the research process.

p-Hacking during Data Collection

During the data collection process, two types of *p*-hacking can occur: first, stopping data collection earlier than planned because preliminary analyses appear to reveal the result that one is looking for, and, second, collecting additional data without controlling for inflated error rates.

Each of these strategies can be problematic. When stopping early (i.e., with lower power than planned), it is more difficult to distinguish random variation from a true effect. Underpowered samples not only have a reduced chance of detecting a true effect, but the likelihood that a statistically significant effect reflects a true effect is also reduced (Button et al., 2013; Ioannidis, 2005). Moreover, if the true effect is zero, *p*-values are uniformly distributed: Every possible *p*-value is equally likely (Simonsohn et al., 2014). Hence, stopping earlier than intended can lead to uninformative results if statistical power remains low.

Collecting more data is generally a good thing as it increases power and the chances of revealing a true effect. However, looking at the data multiple times (and deciding to continue data collection) also increases the danger of false positives if researchers do not statistically control for the increased Type I error rate. If the true effect is zero, *p*-values are uniformly distributed. Hence, they will zigzag endlessly, and as a consequence, a result that is "approaching significance" may turn significant when a few more data points are added without actually revealing a true effect. Fortunately, because no *p*-value is more likely than another if the true effect is zero, in these cases, larger samples will also often reveal larger *p*-values. If there is a true effect, a larger sample increases the chance of obtaining a particularly small *p*-value, not one that barely crosses the 0.05 mark.

For trial researchers, both the practices of stopping early and collecting more data after looking at the results are known as sequential investigation (Armitage et al., 2002) or the sequential stopping rule (Dienes, 2008; Lakens, 2014). Sequential

(group) investigations are particularly important for clinical trials because it may be ethical to terminate the trial early when there is strong evidence in favor of, or against, the treatment under investigation. The determination of when data collection will end is defined as the stopping rule for a study. Problems arise if the decision of when to terminate—or collect more data—is not specified a priori. Checking data more often (than once) increases the actual α level because each test that is conducted offers a new chance to reject the null hypothesis. In the extreme, a stopping rule that implies "I will continue running the experiment until the test is significant" guarantees a significant finding even when the null hypothesis is true (Dienes, 2008).

p-Hacks During Data Preprocessing

Several *p*-hacks can occur during data analysis. We arranged these *p*-hacks according to whether they most typically happen during data preprocessing or during hypothesis testing. In reality, the separation between these stages is not strict. All strategies can be employed at any point when trying to make the data reveal the most about the proposed hypotheses.

We discuss two types of *p*-hacking that refer to data preprocessing and exploratory data analysis: transforming distributions and (selectively) excluding outliers. Data transformations can be useful for normalizing the data. For example, log-transforming the data may give a parametric test more power and—as a result—lower *p*-values. However, such transformations must be specified in advance. *p*-Hacking occurs when an analyst runs the analyses on raw data first and, after trying one or even several transformations, reports the results with the smallest *p*-value (Lew, 2020).

Similar problems can occur when outliers are excluded. Excluding a few data points that are not representative of the rest of the distribution can be useful when these data points exert an extraordinarily strong influence on the inferences researchers draw from the data. Problems arise when the decisions about whether to exclude data points and which ones to exclude are based on how much the various decisions change the *p*-value toward significance. Admittedly, decisions about the exclusion of outliers can be inherently ambiguous. There are several approaches that explain how to exclude outliers (e.g., standard deviations from the mean, median absolute deviation, Boxplot analysis), and within each approach, there are several choices (e.g., 2.5 or 3 standard deviations/median absolute deviations), opening up an extensive array of options. It can be challenging to decide which of these paths is the best choice in a particular study. However, what is clear is that picking the specific path on the basis of the resulting *p*-value will increase the danger of believing that the effect that was found is more robust than it actually is. We will discuss the detrimental consequences of this and other *p*-hacking practices in a later section as well as how to prevent them from occurring.

p-Hacks During Hypothesis Testing

When researchers test confirmatory hypotheses, they may try out multiple analytical approaches (e.g., a *t* test for dependent measures, a robust *t* test such as the Yuen-Test, and an analysis based on change scores). Again, applying diverse methods may happen in good faith in an attempt to identify the most adequate method for analyzing the particular data set. However, running a bunch of analyses and reporting only the method yielding the lowest *p*-value capitalizes on chance. Indeed, recent endeavors have shown that even different well-intentioned analysts can analyze the same data in widely different ways and arrive at conclusions that differ greatly (Silberzahn et al., 2018).

A similar reasoning applies when researchers' decisions to selectively control for covariates or moderators (e.g., gender, age) are based on whether or not this reduces their focal *p*-value instead of a priori theoretical reasoning that it will be advisable to do so. This practice highlights the similarities between *p*-hacking and overfitting. Overfitting refers to situations in which sample-specific noise is misinterpreted as a true signal that can be generalized to the population. Yarkoni and Westfall (2017) consider *p*-hacking to be a form of procedural overfitting because it takes place either prior to or in parallel with hypothesis testing or model estimation.

p-Hacks During the Reporting of Studies

Another subset of *p*-hacking practices refers to decisions about whether to exclude data after looking at the impact of doing so on the results. These decisions may pertain to single data points (e.g., outliers, see above), ghost variables (i.e., dependent variables assessed during data collection but not reported in the publication itself; Bishop & Thompson, 2016), or experimental conditions (e.g., dropping, combining, splitting groups). Conceptually, the dropping of conditions is a borderline case that falls between *p*-hacking and publication bias (i.e., dropping whole studies). Similarly, failing to disclose experimental conditions (e.g., when the results are inconsistent with theoretical predictions) is considered *p*-hacking because this may impact the *p*-value of some analyses (e.g., dropping a condition in a one-way ANOVA).

The term *ghost variables* describes the situation when researchers do not specify in advance their hypotheses about which specific measure will differ between groups or will show a substantial association with a chosen predictor. If researchers report only the significant ones and assign a ghost status to the remaining variables, this is considered *p*-hacking (Bishop & Thompson, 2016). This type of *p*-hacking is problematic because, due to the problems that arise from multiple testing, the inferential statistics reported in such cases will be misleading. Conceptually, dropping dependent variables is also linked to publication bias on the outcome level instead of the study level.

Multiple testing refers to situations in which researchers perform a "family" of tests. When performing one *t* test, the null hypothesis can be rejected at the 5% α level. But when performing two *t* tests, the probability of making a Type I error increases to 9.75%. Therefore, when running more than one test on the same hypothesis, researchers need to control the overall Type I error rate (i.e., the family-wise error). This can be done by correcting (i.e., reducing) the α level for every single test in the family. Researchers sometimes try to avoid having to make such a correction by not reporting some of the tests they ran or by dropping some of their (dependent) variables. Consequently, they report results as significant even when these tests would have missed the corrected threshold if the proper procedures had been followed.

A final *p*-hacking practice that may occur during the reporting of studies is the rounding off of *p*-values. In practice, this means reporting values slightly above 0.05 as equal to or even less than 0.05, and hence, reporting the results as significant when in fact they are not. Thus, other than the previously discussed techniques, this *p*-hack does not even lead to a formally significant result. It only pretends to do so.

The Consequences of *p*-Hacking

p-Hacking has a whole range of implications. In this section, we will discuss two: an increase in false-positive findings and an overestimation of effect sizes.

Increase in False-Positive Findings

The most tangible consequence of *p*-hacking is a sharp increase in false-positive findings—hence, the reference to *p*-hacking as "inflation bias." Put differently, *p*-hacking practices "wreak havoc with a method's error probabilities. It becomes easy to arrive at findings that have not been severely tested" (Mayo, 2018, p. 439). *p*-Hacking leads researchers to believe they found a real effect when in reality there was none, at least not one strong enough to reach statistical significance.

In an impressive demonstration of this consequence, Simmons et al. (2011) convincingly showed how strongly *p*-hacking inflates actual false-positives rates. These authors simulated scenarios for four *p*-hacking practices: choosing from among outcome variables, optional stopping, including covariates, and excluding experimental conditions. Also, they evaluated various combinations of these four practices by taking into account the possibility that several practices may occur jointly. Their simulations showed that applying a single *p*-hack can easily double the factual false-positive rate (under the specific conditions employed by Simmons et al. in their simulation). At the same time, researchers still assume that their results are quite unlikely (≤0.05) under the null hypothesis. The most disturbing finding from

the analyses demonstrated that false-positive rates increased to 61% when the four *p*-hacking practices were combined. Consequently, in this situation, the probability that researchers would erroneously conclude and report a significant finding was higher than the probability that they would correctly reject the null hypothesis. They would even have been better off by flipping a coin.[2]

On a larger scale, one may wonder what it means if a literature is built substantially on studies that, in reality, did not reveal significant findings but were false positives. False-positive findings may be particularly detrimental in small, emerging literatures with a few landmark studies that may give the impression of a robust effect when it is much weaker in reality. One may hope that in the long run, as a literature grows and matures, the weight of individual studies will decrease. Still, if a literature (for whatever reason) remains small, a few false-positive findings can bias perceptions of this literature for a long time. Obviously, the more prevalent and severe *p*-hacking is in a given literature, the more damaging the consequences.

Overestimation of Effect Sizes

A second detrimental consequence of *p*-hacking involves the overestimation of effect sizes. *p*-Hacking means using any of the aforementioned practices to bring an originally nonsignificant *p*-value down to significance. For many of these practices, this essentially means obtaining a larger effect size estimate that will cross the significance threshold (e.g., by including a covariate or excluding some outliers). This effect will be particularly pronounced in small studies with low power because only large effect sizes become significant in such studies.

Imagine a research group that ran a relatively small study with a striking result: evidence for a hitherto unknown effect Y. They published the study in a high-impact journal. When analyzing the data, they tried many different things (i.e., they used researchers' degrees of freedom) and settled on a solution that they believed was most appropriate (in fact, they overfitted the analysis to the data, resulting in an effect size that overestimated the true effect). In addition to (in this case unintended) *p*-hacking, the study is haunted by another problem: Effect sizes are additionally exaggerated in small, underpowered studies such as the one our fictitious research group ran, a statistical phenomenon called the winner's curse (Button et al., 2013). It means that the research group is "cursed" by overestimating the magnitude of the effect in the population due to random error that is more pronounced in underpowered studies (see also Chap. 11, this volume). Other researchers trying to replicate the initial finding Y will then suffer from a "decline effect" (Protzko & Schooler, 2017), indicating that attempts to replicate the effect will likely find considerably

[2]You can experience the power of *p*-hacking yourself by using the *p*-hacker Shiny app (Schönbrodt, 2016).

smaller effects or even end up with a null finding. In this situation, further *p*-hacking when analyzing the replication attempt becomes more likely, particularly in a culture that incentivizes significant results. The researchers' assumption that there must be a true effect and that it's only hidden due to a suboptimal analysis will motivate them to dig deeper and "dredge" the data. This illustrates how nonrobust findings can initiate a vicious cycle that—in combination with maladaptive incentive structures—motivates the continued use of *p*-hacking and other questionable research practices.

A recent large-scale simulation study examined the extent to which *p*-hacking can bias effect size estimates of whole literatures, either on its own or in combination with publication bias (Friese & Frankenbach, 2020). Publication bias is another questionable research practice that arises when studies that did not produce the desired outcomes are less likely to be published than studies that "worked" (Franco et al., 2014; Ioannidis et al., 2014). Hence, publication bias occurs at the level of studies (is a study published or not?), whereas *p*-hacking refers to the data collection practices and analyses used in a study.

The results of Friese and Frankenbach's (2020) study revealed that *p*-hacking and publication bias result in different threats to the robustness of findings. Whereas *p*-hacking can dramatically increase the rate of false positives in a given literature, high levels of publication bias can lead to a considerable distortion of (meta-analytic) effect size estimates. Perhaps surprisingly, in the absence of publication bias, *p*-hacking does little to distort meta-analytic effect size estimates. However, the two phenomena interact: *p*-hacking adds considerable bias to effect size estimates at medium levels of publication bias—particularly in literatures where the true effects in question are small. At low and high levels of publication bias, *p*-hacking hardly contributes any bias to meta-analytic effect size estimates. With increasing true effect sizes, literatures are more and more shielded against the effect size bias introduced by publication bias and *p*-hacking (Fig. 5.2).

Increased rates of false-positive findings and bias in effect size estimates have palpable implications for the literature's meta-analytical record. Large numbers of seemingly positive (but factually false-positive) results create the impression of a robust and accurate literature. As the inflated effect sizes from single studies will be included and summarized in meta-analyses, they bias the meta-analytical effect size (provided that there is some publication bias). As a consequence, researchers conducting new work who base their expectations on these biased estimates will inadvertently run underpowered studies because they believe that the effects of interest are more robust than they actually are. This combination of increased rates of false-positive findings and biased meta-analytical effect size estimates impedes the accumulation of knowledge. For one, it leads researchers who are trying to build upon previous work astray. In addition, practitioners relying on biased literature might not be able to provide the best solutions to those they are working with. The result is a lamentable waste of resources.

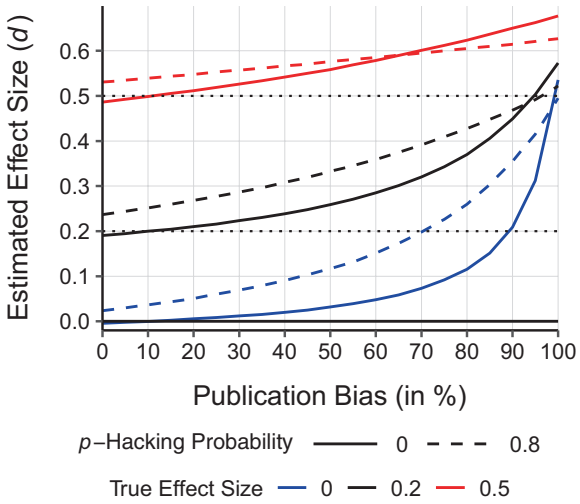

Fig. 5.2 Meta-analytic effect size estimates as a function of degrees of *p*-hacking, publication bias, and true effect size
Note. In the absence of publication bias, *p*-hacking does little to distort meta-analytic effect size estimates. By contrast, high degrees of publication bias do distort these estimates even in the absence of *p*-hacking. Together, *p*-hacking and publication bias interact such that *p*-hacking adds considerable bias when publication bias is moderate. Bias is greatest when the true effect size is very small. Larger true effects act as a shield against the deleterious effects of *p*-hacking and publication bias on meta-analytic effect size estimates. Figure reprinted from Friese and Frankenbach (2020).

The Prevalence and Detection of *p*-Hacking

The prevalence of *p*-hacking in different disciplines has been discussed repeatedly (Fiedler & Schwarz, 2016; John et al., 2012, see also Chap. 6, this volume). There are essentially three approaches that seek to determine how frequently researchers *p*-hack. One directly surveys researchers about their practices, whereas the others attempt to obtain statistical indicators of *p*-hacking based on published literature or by comparing planned with reported analyses.

Prevalence Estimates Based on Self-Reports

John et al. (2012) aimed to estimate the prevalence of *p*-hacking (and other Questionable Research Practices, QRPs) in a few ways (see also Chaps. 1 and 2, this volume). One way was to acquire self-admission rates for 10 QRPs (e.g., "rounding off" a *p*-value, deciding whether to exclude data after determining how such an

exclusion would impact the results, or deciding whether to collect more data after looking to see whether the results were significant; John et al., 2012, p. 525). The second way involved asking participants to estimate the percentage of other psychologists who had engaged in the behavior. There was large variability across the 10 QRPs for both indicators. For example, up to 58% of participants indicated that they had at least once decided "whether to collect more data after looking to see whether the results were significant." Conversely, for claiming in a paper "that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)" (p. 525), only 4.5% indicated that they had done this at least once. The prevalence estimates of other psychologists engaging in these practices were often higher than the self-admissions. With respect to participants working in clinical psychology, the mean self-admission rate across all 10 QRPs was 27%.

The findings by John et al. (2012) have been frequently cited but also criticized for exaggerating actual prevalence rates because the authors used lifetime prevalence rates to conclude that some research practices "constitute the prevailing research norm" (p. 524). However, lifetime prevalence rates are unable to distinguish between researchers who engaged in a QRP once and only once in their lifetime and researchers who engage in the same QRP regularly. A conceptual replication among German psychologists decomposed the prevalence of QRPs into their two multiplicative components: the proportion of researchers who ever committed a given practice and, if so, how frequently (Fiedler & Schwarz, 2016). This survey found prevalence estimates that were a lot lower than those reported by John et al. (2012).

John et al. (2012) also suggested that prevalence rates of QRPs are not uniformly distributed across the subdisciplines of psychology. Even within subdisciplines, there are different subfields with potentially different research cultures that may be more or less susceptible to certain QRPs. Of particular relevance for the present purposes is a recent survey among faculty and students in clinical and counseling psychology doctoral programs (Swift et al., 2020). In this survey, over 64% of faculty and 48% of students indicated engaging in at least one of 12 QRPs at least once during their career. The *p*-hacking practices that participants admitted to engaging in included rounding off a *p*-value (12.8% of faculty and 8.2% of doctoral students) and excluding data after looking at the impact of doing so (11.8% of faculty and 9.1% of doctoral students). These admission rates were considerably lower than the rates reported in other surveys (e.g., 22% for at least once rounding-off a *p*-value and 40% for at least once excluding data; Fiedler & Schwarz, 2016).

Prevalence Estimates Based on Analyses of the Published Literature

Assuming that self-reported data underestimate socially undesirable behavior, other approaches attempt to obtain statistical indicators of *p*-hacking on the basis of published literature. For example, some researchers have suggested that when looking

at empirical *p*-value distributions, clusters of *p*-values just below 0.05 may indicate that researchers engaged in *p*-hacking strategies until their results were (barely) significant. Indeed, this pattern was found by some large-scale analyses of *p*-value distributions across multiple sciences, suggesting that *p*-hacking is widespread (e.g., Head et al., 2015).

These findings have been disputed for two reasons: First, other researchers have argued that a bump in the number of *p*-values just below 0.05 is a sufficient but not necessary condition for the presence of specific forms of *p*-hacking (Hartgerink, 2017) and that *p*-value distributions that reveal evidence of *p*-hacking likely look different (Lakens, 2015a). *p*-Value distributions depend on additional factors, such as power and publication bias. With some types of *p*-hacking (e.g., multiple testing and reporting the analysis that yielded the smallest *p*-value), the *p*-value distributions are not likely to reveal clusters just below 0.05 (Hartgerink, 2017; Lakens, 2015a, b). Second, re-analyses of the data by Head et al. (2015) and other studies did not find convincing evidence of a bump in the number of *p*-values just below 0.05 (Hartgerink, 2017; Lakens, 2015a, b).

Prevalence Estimates Based on Planned Versus Reported Analyses

A third approach is somewhat broader and does not apply to all of the *p*-hacking practices we discussed, but only to a subset. It compares records of studies that were openly available before publication with the final published paper. Thereby, this approach can reveal so-called selective reporting practices because it can detect the omission of variables or conditions that yielded undesired results, the underreporting of null results (publication bias), and HARKing (Cairo et al., 2020).

In one study, Franco et al. (2014) looked at a database of empirical studies that had been submitted for review at a National-Science-Foundation-sponsored program. They found that the publication probability of null findings was remarkably lower than for studies that yielded the desired results (a difference of approximately 40%). Hence, Franco et al.'s results indicate the presence of publication bias.

This approach has been further refined within organizational and management research (O'Boyle et al., 2017) and social psychology (Cairo et al., 2020). O'Boyle et al. (2017) vividly labeled the process of initial results (the ugly caterpillar) turning into a journal article (the beautiful butterfly) the "Chrysalis Effect." The authors compared 1978 hypotheses proposed in dissertations with hypotheses published in journal articles that were based on these dissertations. They found that 1000 hypotheses (!) were dropped in the process. The proportion of significant findings (the ratio of supported hypotheses to all contained hypotheses) increased by 21.0% (from 44.9% in dissertations to 65.9% in published articles). This inflation happened not only because hypotheses that did not yield the desired (i.e., significant) result had been dropped but also because new hypotheses had been added, the direction of predicted effects had been reversed, data had been altered, or variables had been selectively deleted or added (O'Boyle et al., 2017).

In social psychology, Cairo et al. (2020) looked at 100 dissertations, 373 published studies, and 1136 hypotheses and found that selective reporting practices were widespread. Supported hypotheses were four times more likely to end up in published journal articles than unsupported hypotheses and three times more likely to be reported unchanged. Again, the dropping of unsupported hypotheses alone resulted in a 20% inflation of significant findings in the published literature. In conclusion, the prevalence of *p*-hacking has been tackled via different approaches. All approaches have some merits but also some unresolved issues. Therefore, the actual frequency of *p*-hacking is unknown to date.

The Prevention of *p*-Hacking

Researchers have proposed, developed, and refined several solutions to prevent *p*-hacking practices. Some of them have been around for many years, for example, randomized controlled trials (RCTs). Others, such as preregistration, Registered Reports, and multiverse analyses, are more recent. In the following, we will describe some of the proposed solutions and discuss their good points and challenges.

RCTs

Medicine was one of the first disciplines to use registries. Initially, registries for clinical trials were aimed at facilitating the recruitment of patients, speeding up the dissemination of information, and reducing bias in the reporting of trials (Dickersin & Rennie, 2003). This idea is as timely as ever. Unfortunately, evaluations of RCTs have suggested that they often fall short of their potential. Of all registered trials, only about 50–60% end up in a journal, and those that find significant results have a higher probability of being published (Easterbrook et al., 1991; Tackett et al., 2019). This evidence of publication bias and low replication rates in registered clinical trials (e.g., Begley & Ellis, 2012; Prinz et al., 2011) has led to various attempts to improve the registration processes. New legal regulations, official statements (e.g., the Declaration of Helsinki), and technical advances have promoted centralized registries. These developments seem to have been successful at reducing selective reporting practices. For example, Kaplan and Irvin (2015) looked at large RCTs in drug research published between 1970 and 2012. They showed that after making the registration of primary outcomes obligatory on ClinicalTrials.gov in 2000, the percentage of positive results in the published trials dropped from 57% to 8%. They argued that both the obligatory prospective declaration of outcomes and improvements in transparency in the reporting standards may be responsible for this decline in the proportion of positive findings. Although one cannot be entirely certain that the inflation of positive findings before 2000 is purely due to *p*-hacking, it stands to reason that a flexible determination of the primary outcome after looking at the data

may have played a role here. Therefore, Kaplan and Irvin concluded that the required registration of studies accompanied by improvements in the transparency of the RCTs were the key for the sharp increase in null findings.

Preregistration

Several clinical research questions cannot be addressed with RCTs, but alternative solutions can help researchers avoid *p*-hacking. One of them is preregistration (see also Chap. 15, this volume). Preregistrations state research objectives, report the study design, describe the planned sample (size), and detail the planned analyses. Thus, they allow for a comparison between the studies that have been conducted with the studies that have been published. For the prevention of *p*-hacking, preregistration has numerous benefits. For one, it allows confirmatory research to be distinguished from exploratory research. Specifying which analyses were planned a priori and which were run post hoc helps to prevent (or at least to detect) practices such as including covariates or excluding or switching outcomes.

The idea of registering empirical studies a priori is itself not new. For example, de Groot (1956/2014) determined that "it is essential that these hypotheses have been precisely formulated and that the details of the testing procedure (which should be as objective as possible) have been registered in advance" (p. 188). Similarly, concerning decisions about whether to perform a one-sided versus a two-sided test, Bakan (1966) stated: "How should this be handled? Should there be some central registry in which one registers one's decision to run a one- or two-tailed test before collecting the data? Should one, as one eminent psychologist once suggested to me, send oneself a letter so that the postmark would prove that one had pre-decided to run a one-tailed test?" (p. 431). Hence, the awareness that a priori predictions are essential in science has been around for over 60 years. But only the recent development of online tools that allow for the time stamping and freezing of research plans, accompanied by the acknowledgment of an imperative change in culture, have substantially improved the feasibility of preregistrations. Researchers may now use platforms such as the Open Science Foundation (OSF) or AsPredicted.org to share their research plans openly. Moreover, several templates tailored for different purposes (e.g., experimental research, longitudinal and experience sampling studies, analysis of existing data) may lower the threshold for undertaking a registration (see https://osf.io/zab38/wiki/home/).

Recently, Benning et al. (2019) and Krypotos et al. (2019) introduced helpful guidelines directed at clinical psychology. Benning et al. (2019) spoke of a continuum of registration because study registrations may vary in the timing and their disclosure. Whereas preregistrations (such as clinical trial registrations, Registered Reports or grant proposals) occur before the data are collected, *coregistrations* disclose decisions made after researchers began collecting data but before any data were analyzed. *Postregistrations* occur after data analysis has begun but still offer the opportunity to disclose specific analytic choices. In all types of registrations,

researchers may register anything from a single specific aspect of data collection and analysis to complex decision trees that illustrate a series of decisions. Hence, Benning et al. (2019) presented registrations as a flexible framework for helping clinical researchers to increase the credibility of their work. To do this, Krypotos et al. (2019) provide a hands-on approach. They offer a step-by-step guide on preregistration, anonymizing data, and sharing both materials and data in psychopathology studies. The authors developed an open-source application based on the (free) statistical software package R (R Core Team, 2020) and git (a toolkit for tracking and merging changes) to facilitate version control and the time stamping of each step during the study. Researchers may thus use the same files throughout the study and easily track changes throughout the project.

One final remark about preregistration strikes us as important: A preregistration is more useful and effective at preventing *p*-hacking the more clearly and precisely it lays out the plan for a study. At the same time, it always has to be clear that a preregistration is a "plan, not a prison" (DeHaven, 2017). Making a plan to the best of one's ability is great, but there can always be reasons why it became necessary to deviate from this plan. This poses no problem as long as these deviations are transparently reported and explained.

Registered Reports

A particular type of preregistration is a Registered Report (Chambers et al., 2014). Registered Reports refer to a type of preregistration that is presented in an article format and undergoes peer review before the data are collected. In a first step, the authors submit a Stage 1 part of the manuscript, including the Introduction, Method, and pilot study results if available (Chambers et al., 2014). After revisions proposed by reviewers and the editor, the authors are offered an in-principle acceptance if the Stage 1 manuscript is sound. An in-principle acceptance guarantees the final paper's publication regardless of the results as long as the authors adhere to the approved protocol. After collecting and analyzing the data, the authors submit their initial Stage 1 manuscript along with the Results and Discussion sections. This Stage 2 manuscript may contain any unplanned, additional analyses labeled as "exploratory."

This publishing model prevents *p*-hacking—in addition to HARKing, problems with low power (because Stage 1 manuscripts are only accepted if the planned study seems adequately powered), and publication bias. Registered Reports alleviate the pressure to produce novel and astounding results and emphasize rigor and reproducibility instead (Chambers et al., 2014). Moreover, due to the in-principle acceptance, Registered Reports help (early career) researchers disseminate their ideas more quickly and increase the visibility of these ideas. Given these benefits, it is not surprising that this new submission category has been introduced in over 250 journals by now (https://www.cos.io/initiatives/registered-reports).

Multiverse Analyses

Whereas unreviewed and reviewed preregistration types provide a solution for distinguishing confirmatory from exploratory research, multiverse analyses (Steegen et al., 2016) help to prevent biases in exploratory research. They involve running the same analyses across all reasonable combinations of different transformations, exclusions, and inclusions of data and variables to examine how they affect the results and conclusions. Thus, multiverse analyses address all the arbitrary decisions that have to be made during data processing. They demonstrate the sensitivity of the results to analysts' arbitrary choices. Hence, transparent multiverse analyses leave it to the scientific community to gauge the fragility of the conclusions and their credibility.

Concluding Remarks: A Case for Open Science

In the present chapter, we discussed several *p*-hacking practices as part of the broader category of questionable research practices. Throughout the sections, it became clear that *p*-hacking can have detrimental consequences—particularly an increase in false-positive rates—that ultimately damage the trustworthiness and robustness of psychological science. In these concluding remarks, we would like to add a final nuance to the previous considerations by asking: Are *p*-hacking practices—or any questionable research practices for that matter—necessarily blameworthy after all?

For some practices, the answer is clear. They are simply wrong. For example, there is no justification for generously rounding off a *p*-value to 0.05 to make the result look significant if the actual value is higher. The *p*-value should be reported precisely to the third decimal place (APA, 2020). However, other practices might not be inherently wrong. In fact, they can be quite sensible, useful, or even necessary. For example, in general, more data are better than less data. So, continuing data collection after peeking at the data may be a good idea. Including a covariate can make a lot of sense. Trying many different ways to analyze a data set can be highly informative and a sign of conscientiousness instead of a questionable research practice and so on. What can make these practices bad scientific practice is not that they are conducted at all. Rather, researchers are engaging in bad practice when their actions are not transparently reported to make clear what parts of the data analysis were planned a priori and what parts were added as exploratory analyses. In addition, bad practice occurs when the increased Type I error rates that result from massaging the data are not controlled for.

If researchers transparently disclose their a priori data analysis plan, where they deviated from this plan, why they did so, and how this affected the results, there is nothing wrong with amply exploring the data and reporting emerging insights that seem interesting. In fact, we encourage all researchers to explore their data sets, run

unplanned analyses, and come up with post hoc reasoning and new theoretical ideas—as long as these steps are labeled as such: post hoc. They can then be tested with confirmatory analyses in future research.

When appraising what is and is not questionable about questionable research practices, it becomes clear that some are not questionable, they are simply indefensible. Others might be better termed "questionable reporting practices," indicating that the problem lies in a lack of transparency more than in engaging in these practices per se (Wigboldus & Dotsch, 2016).

What can any researchers do to confirm that they did not engage in questionable research practices? The solution is surprisingly simple. It lies in the transparent distinction between a priori planned, confirmatory steps of data analysis and exploratory, additional steps. The line between the two can be drawn easily by adhering to the open science practices outlined above, particularly the detailed preregistration of all measures, manipulations, hypotheses, and planned analysis steps. Preliminary evidence suggests that this practice is remarkably effective. A recent analysis found that manuscripts published in the Registered Report format outperformed comparison papers published in the traditional format on 19 different criteria, including improvements in novelty, creativity, methodological rigor, and overall paper quality, among others (Soderberg et al., 2020).

Open science practices are surely not the solution to all challenges psychological science currently faces, but they are a pretty good and easy-to-implement solution to prevent *p*-hacking. Let's do it.

References

APA. (2020). *Publication manual of the American Psychological Association* (7th ed.). APA. https://apastyle.apa.org/products/publication-manual-7th-edition

Armitage, P., Berry, G., & Matthews, J. N. S. (2002). *Statistical methods in medical research* (4th ed.).

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423. https://doi.org/10.1037/h0020412

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533. https://doi.org/10.1038/483531a

Benning, S. D., Bachrach, R. L., Smith, E. A., Freeman, A. J., & Wright, A. G. C. (2019). The registration continuum in clinical science: A guide toward transparent practices. *Journal of Abnormal Psychology, 128*(6), 528–540. https://doi.org/10.1037/abn0000451

Bertamini, M., & Munafò, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science, 7*(1), 67–71. https://doi.org/10.1177/1745691611429353

Bishop, D. V., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ, 4*, e1715. https://doi.org/10.7717/peerj.1715

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cairo, A. H., Green, J. D., Forsyth, D. R., Behler, A. M. C., & Raldiris, T. L. (2020). Gray (literature) matters: Evidence of selective hypothesis reporting in social psychological research. *Personality and Social Psychology Bulletin, 46*(9), 1344–1362. https://doi.org/10.1177/0146167220903896

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered reports at AIMS neuroscience and beyond. *AIMS Neuroscience, 1*(1), 4–17. https://doi.org/10.3934/Neuroscience.2014.1.4

Cuijpers, P. (2016). Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies. *Evidence-Based Mental Health, 19*(2), 39–42. https://doi.org/10.1136/eb-2016-102341

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive-behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *The British Journal of Psychiatry: the Journal of Mental Science, 196*(3), 173–178. https://doi.org/10.1192/bjp.bp.109.066001

de Groot, A. D. (1956). 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. https://doi.org/10.1016/j.actpsy.2014.02.001

DeHaven, A. (2017). *Preregistration: A plan, not a prison*. https://www.cos.io/blog/preregistration-plan-not-prison.

Dickersin, K., & Rennie, D. (2003). *Registering clinical trials. Jama, 290*(4), 516–523. https://doi.org/10.1001/jama.290.4.516

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet, 337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-y

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Friese, M., & Frankenbach, J. (2020). P-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods, 25*(4), 456–471. https://doi.org/10.1037/met0000246

Hartgerink, C. H. J. (2017). Reanalyzing Head et al. (2015): Investigating the robustness of widespread p-hacking. *PeerJ, 5*, e3068. https://doi.org/10.7717/peerj.3068

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of P-hacking in science. *PLoS Biology, 13*(3). https://doi.org/10.1371/journal.pbio.1002106

Heene, M., & Ferguson, C. J. (2017). Psychological science's aversion to the null, and why many of the things you think are true, aren't. In *Psychological science under scrutiny* (pp. 34–52). Wiley. https://doi.org/10.1002/9781119095910.ch3

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010

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

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. https://doi.org/10.1371/journal.pone.0132382

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Krypotos, A.-M., Klugkist, I., Mertens, G., & Engelhard, I. M. (2019). A step-by-step guide on preregistration and effective data sharing for psychopathology research. *Journal of Abnormal Psychology, 128*(6), 517–527. https://doi.org/10.1037/abn0000424

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

Lakens, D. (2015a). Comment: What p-hacking really looks like: A comment on Masicampo and Lalande (2012). *Quarterly Journal of Experimental Psychology, 68*(4), 829–832. https://doi.org/10.1080/17470218.2014.982664

Lakens, D. (2015b). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ, 3*, e1142. https://doi.org/10.7717/peerj.1142

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Leweke, F., Luyten, P., Keefe, J. R., Midgley, N., Rabung, S., Salzer, S., & Steinert, C. (2017). Biases in research: Risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychological Medicine, 47*(6), 1000–1011. https://doi.org/10.1017/S003329171600324X

Lew, M. J. (2020). A reckless guide to P-values. In A. Bespalov, M. C. Michel, & T. Steckler (Eds.), *Good research practice in non-clinical pharmacology and biomedicine* (pp. 223–256). Springer International Publishing. https://doi.org/10.1007/164_2019_286

Lilienfeld, S. O. (2017). Psychology's replication crisis and the Grant culture: Righting the ship. *Perspectives on Psychological Science, 12*(4), 660–664. https://doi.org/10.1177/1745691616687745

Lilienfeld, S. O., & Waldman, I. D. (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. Wiley.

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

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

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. https://doi.org/10.1177/1745691612459058

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*(2), 376–399. https://doi.org/10.1177/0149206314527133

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

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

Protzko, J., & Schooler, J. W. (2017). Decline effects: Types, mechanisms, and personal reflections. In *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 85–107). Wiley Blackwell. https://doi.org/10.1002/9781119095910.ch6

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rapport, M. D., Orban, S. A., Kofler, M. J., & Friedman, L. M. (2013). Do programs designed to train working memory, other executive functions, and attention benefit children with ADHD? A meta-analytic review of cognitive, academic, and behavioral outcomes. *Clinical Psychology Review, 33*(8), 1237–1252. https://doi.org/10.1016/j.cpr.2013.08.005

Reardon, K. W., Smack, A. J., Herzhoff, K., & Tackett, J. L. (2019). An N-pact factor for clinical psychological research. *Journal of Abnormal Psychology, 128*(6), 493–499. https://doi.org/10.1037/abn0000435

Sakaluk, J. K., Williams, A. J., Kilshaw, R. E., & Rhyner, K. T. (2019). Evaluating the evidential value of empirically supported psychological treatments (ESTs): A meta-scientific review. *Journal of Abnormal Psychology, 128*(6), 500–509. https://doi.org/10.1037/abn0000421

Schönbrodt, F. D. (2016). *p-hacker: Train your p-hacking skills!*. http://shinyapps.org/apps/p-hacker/.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

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

Soderberg, C. K., Errington, T., Schiavone, S. R., Bottesini, J. G., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2020). Initial evidence of research quality of registered reports compared to the traditional publishing model. *MetaArXiv*. https://doi.org/10.31222/osf.io/7x9vy

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Swift, J. K., Christopherson, C. D., Bird, M. O., Zöld, A., & Goode, J. (2020). Questionable research practices among faculty and students in APA-accredited clinical and counseling psychology doctoral programs. *Training and Education in Professional Psychology*. https://doi.org/10.1037/tep0000322

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology, 15*(1), 579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710

Tackett, J. L., & Miller, J. D. (2019). Introduction to the special section on increasing replicability, transparency, and openness in clinical psychology. *Journal of Abnormal Psychology, 128*(6), 487. https://doi.org/10.1037/abn0000455

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

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01832

Wigboldus, D. H. J., & Dotsch, R. (2016). Encourage playing with data and discourage questionable reporting practices. *Psychometrika, 81*(1), 27–32. https://doi.org/10.1007/s11336-015-9445-1

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393