Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Critical Values Robust to P-hacking

Adam McCloskey, Pascal Michaillat

March 2024

Abstract. P-hacking is prevalent in reality but absent from classical hypothesis-testing theory. We therefore build a model of hypothesis testing that accounts for p-hacking. From the model, we derive critical values such that, if they are used to determine significance, and if p-hacking adjusts to the new significance standards, spurious significant results do not occur more often than intended. Because of p-hacking, such robust critical values are larger than classical critical values. In the model calibrated to medical science, the robust critical value is the classical critical value for the same test statistic but with one fifth of the significance level. (JEL code: C12)

Affiliations. McCloskey: University of Colorado–Boulder. Michaillat: University of California–Santa Cruz.

Acknowledgments. We thank Isaiah Andrews, Tim Bollerslev, Brian Cadena, Kenneth Chay, Andrew Chen, Garret Christensen, Pedro Dal Bo, Stefano DellaVigna, Peter Hull, Larry Katz, Miles Kimball, Megan Lang, Jonathan Libgober, Carlos Martins-Filho, Andriy Norets, Emily Oster, Bobak Pakzad-Hurson, Wenfeng Qiu, Jonathan Roth, Jesse Shapiro, and Yanos Zylberberg for helpful discussions and comments. This work was supported by the Institute for Advanced Study.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

I. Introduction

Definition of p-hacking. P-hacking occurs when scientists engage in various behaviors that increase their chances of reporting statistically significant results (Simonsohn, Nelson, & Simmons, 2014; Wasserstein & Lazar, 2016). Typical p-hacking practices include running many small-sample experiments rather than one large-sample experiment; reporting studies with significant results but suppressing studies with insignificant results; collecting data until a significant result is obtained; dropping inconvenient observations or outcomes from a study; and searching for statistical specifications that produce significant results (Nosek, Spies, & Motyl, 2012; Lindsay, 2015; Christensen, Freese, & Miguel, 2019; Stefan & Schoenbrodt, 2023).

Prevalence of p-hacking. P-hacking is prevalent in science (online appendix B.1). Scientists readily admit to it. It is visible in meta-analyses: the distributions of test statistics in entire literatures show that scientists tinker with their analyses to obtain significant results. And it appears when tracking cohorts of scientific studies: studies finding significant results are almost certain to be reported, whereas studies finding insignificant results are likely to remain unreported.

Reasons for p-hacking. That p-hacking is so prevalent is unsurprising because scientists face strong incentives to p-hack. First, significant results are more rewarded than insignificant ones (online appendix B.2). This is because scientific journals prefer publishing significant results. Publications, in turn, determine a scientist's career path, including promotions, salary, and honorific rewards. Second, scientists enjoy a lot of flexibility in data collection and analysis (online appendix B.3). Hence, even when the null hypothesis is true, they have ample opportunity to obtain significant results without violating scientific norms.

Problems caused by p-hacking. Despite its prevalence, p-hacking is not accounted for in classical hypothesis-testing theory. Therefore, classical critical values set a standard for significance that is too lax: a true null hypothesis is rejected more often than purported by the test's significance level.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

This is problematic because hypothesis tests are informative only insofar as a true null hypothesis is not rejected more often than the significance level. For instance, hypothesis tests are used to evaluate scientific theories and paradigms (Kuhn, 1957; Akerlof & Michaillat, 2018). They allow scientists to identify instances when theory does not accord well with empirical observations. Unbridled p-hacking threatens scientific progress. It leads to excessive rejection of established paradigms and to the unwarranted adoption of new paradigms. As such, it threatens the credibility of science. One manifestation of uncontrolled p-hacking is the replication crisis in science (Ioannidis et al., 2014; Christensen & Miguel, 2018).

Existing corrections for p-hacking. A few corrections for p-hacking in hypothesis testing have been discussed (Anscombe, 1954; Lovell, 1983; Glaeser, 2008). But these corrections take the scientist's p-hacking behavior as fixed, whereas in reality the scientist would change her p-hacking behavior as soon as the correction is implemented. Consider for instance a hypothesis test with a significance level of 5%. Classical critical values are constructed such that if the scientist conducted one experiment, a true null hypothesis would be rejected no more than 5% of the time. But if a scientist conducted more than one experiment, performed hypothesis tests in each experiment separately, and reported the best result, a true null hypothesis would be rejected more often than 5% of the time. Existing corrections take the number of experiments as given and compute a more stringent critical value based on this number. But this is insufficient to resolve the problem. Just as scientists may conduct more than one experiment under the classical critical value, they may conduct more experiments than anticipated under the new critical value, overwhelming the proposed correction.

This paper's correction for p-hacking. In this paper, we start by developing a model of hypothesis testing that accounts for p-hacking. From the model, we derive critical values such that, if they are used to determine significance, and if scientists optimally p-hack in response to the new significance standards, spurious significant results do not occur more often than intended. Unlike classical critical values, these robust critical values deliver the promised probability of type 1 error. Once the robust

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

critical values are in place, scientists continue to p-hack, but readers can be confident that true null hypotheses are not rejected more often than the advertised significance level.

Model of hypothesis testing with p-hacking. We consider a scientist who tests a hypothesis by conducting an experiment. If she obtains a significant result from the experimental data, she obtains a high payoff. By contrast, if she obtains an insignificant result, she obtains a lower payoff. The difference in payoff between significant and insignificant results reflects the facts that significant results are more likely to be published, and publications yield rewards to scientists. Therefore, if the scientist obtains an insignificant result, and if she still has resources to devote to the project, she has the incentive to conduct another experiment to try to obtain a significant result using the second experiment's data. Conducting a second experiment without disclosing the first experiment constitutes p-hacking.[1]

Optimal p-hacking strategy. Using optimal stopping theory, we find that the scientist's optimal strategy is to conduct experiments until finding a significant result (Ferguson, 2007). Not all projects yield significant results, however, because resources that a scientist can devote to any project are finite (Chen, 2021). If the scientist runs out of resources before finding a significant result, she reports an insignificant result.

Probability of type 1 error. We begin by computing the expected number of experiments run by a scientist when the null hypothesis is true, as a function of the prevailing critical value. From this we compute the probability of type 1 error as a function of the critical value. The critical value influences the probability of type 1 error in two ways. First, it determines the probability that a true null hypothesis is rejected in each experiment—as in classical statistics. Second, it influences the number of experiments that the scientist collects—a feature unique to our model.

[1]Because the number of experiments is not observable, multiple-testing corrections cannot be used to correct for p-hacking.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Critical value robust to p-hacking. From these results we compute the critical value such that type 1 errors occur at the intended rate—given by the significance level. This critical value is robust to p-hacking, and it is given by a nonstandard form of Bonferroni correction. For any test statistic and any significance level, the robust critical value is the classical critical value for the same test statistic with the significance level divided by the expected number of experiments when the robust critical value is in place. Accordingly, the robust critical value is larger than the classical critical value for the same test statistic and significance level. An advantage of the model is that the expected number of experiments when the robust critical value is in place, and the robust critical value itself, are solely determined by two parameters: significance level and probability of completing an experiment before running out of resources.

Numerical illustration. To illustrate the amount of correction that p-hacking might require, we calibrate the completion probability using evidence from medical science (Dwan et al., 2008). We obtain the rule of thumb that the robust critical value for any test statistic is the classical critical value for the same test statistic with one fifth of the significance level. Hence, the robust critical value for a significance level of 5% is the classical critical value for a significance level of 5%/5 = 1%. For a z-test with a significance level of 5%, and similarly for a large-sample t-test with a significance level of 5%, this means that the robust critical value is 2.33 instead of 1.64 if the test is one-sided, and 2.58 instead of 1.96 if the test is two-sided.

Extensions of the model. Our model of hypothesis testing is quite stylized, but it can be extended in various ways. In online appendix D, we add a cost of doing research, incurred by the scientist at each new experiment. In online appendix E, we add time discounting, which reduces the value of significant results obtained far into the future. And in online appendix F, we assume that consecutive experiments become more and more difficult to run, and thus less and less likely to be completed. In all these extensions, the robust critical value computed in the basic model continues to be operational: it maintains the probability of type 1 error below the significance level.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Other p-hacking strategies. In the model, scientists p-hack by repeatedly running experiments until they reach significant results. This p-hacking strategy appears to be quite common (Bakker, van Dijk, & Wicherts, 2012). However, the model can be adapted to describe a wider range of p-hacking strategies. In online appendix C.2, we consider scientists who pool data across experiments. In online appendix C.3, we consider scientists who remove more and more outliers until they reach significant results. In online appendix C.4, we consider scientists who successively examine different regression specifications so as to obtain significant results. Finally, in online appendix C.5, we consider scientists who successively examine different instruments to reach significant results. We find that the robust critical value computed under the repeated-experiment strategy remains useful under these other p-hacking strategies because it maintains the probability of type 1 error below the significance level.

Control of type 1 error rate for generic p-hacking strategies. More generally, the robust critical value derived in the basic model controls the type 1 error rate for any p-hacking strategy that induces positive dependence across test statistics (online appendix C.1). While the basic model assumes independent test statistics—each obtained from a separate experiment—real-world p-hacking often yields dependent test statistics. Nonetheless, our robust critical value remains valuable by maintaining the probability of type 1 error below the significance level even when p-hacking induces positive dependence across test statistics. Positive dependence results from various p-hacking strategies encountered in practice: when scientists pool data across experiments, when they remove outliers, or when they search across various statistical specifications. Our robust critical value can therefore be used even if the particular p-hacking strategies used by scientists are unknown, as long as these strategies can be expected to generate positive dependence across test statistics, and the completion probability is calibrated to the upper bound of plausible completion probabilities across strategies.

Difference between p-hacking and publication bias. P-hacking and publication bias are frequently conflated, but they represent distinct issues. P-hacking refers to the attempts by researchers to reach significance in individual studies, while publication bias refers to journals' preference for significant

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

results—their reluctance to publish insignificant results. P-hacking makes significance more likely within any published or unpublished study, while publication bias makes significant results more prevalent across bodies of the published literature. This paper solely tackles the distortions created by p-hacking in individual studies. Other methods are available to correct the distortions created by publication bias in meta-analyses (Begg & Berlin, 1988; Hedges, 1992; Duval & Tweedie, 2000; Stanley, 2005; Simonsohn, Nelson, & Simmons, 2014; Stanley & Doucouliagos, 2014; McCrary, Christensen, & Fanelli, 2016; Andrews & Kasy, 2019). Presumably these methods could continue to be used to debias meta-analytic estimates even if critical values robust to p-hacking replaced classical critical values.

II. Model of hypothesis testing with p-hacking

This section develops a simple model of hypothesis testing with p-hacking. A scientist runs experiments with the aim of reaching a significant result. Running experiments takes time, stamina, and money, which are all in finite supply. Because scientists must report results before running out of resources, not all projects yield significant results.

A. Hypothesis test

The scientist tests a null hypothesis H_0 against an alternative hypothesis H_1. The data are governed by a different probability distribution under each hypothesis. The scientist sets the test's significance level to $\alpha \in (0, 1)$. The significance level gives the intended probability of type 1 error—the error that occurs when a true null hypothesis is rejected. Common significance levels are 10%, 5%, and 1%.

B. Test statistic

To conduct the hypothesis test, the scientist collects a dataset from an experiment. From this dataset she constructs a test statistic T, whose realization is t. Under H_0, the cumulative distribution function

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

of the test statistic is F, its survival function is $S = 1 - F$, and its inverse survival function is $Z = S^{-1}$.[2]

C. *Classical critical value*

The null hypothesis is rejected when the test statistic exceeds the critical value z. If the scientist obtains a test statistic $t > z$, the null hypothesis is rejected: the result is significant. But if she obtains a test statistic $t \leq z$, the null hypothesis cannot be rejected: the result is insignificant. Accordingly, the probability of type 1 error is $S(z)$. The classical critical value is set such that the probability of type 1 error in one single test equals the significance level:

$$S(z) = \alpha, \tag{1}$$

or equivalently $z = Z(\alpha)$.

D. *Rewards from significant results*

The first nonclassical element of the model is the rewards accruing to significant results. To capture the facts that significant results are more likely to be published than insignificant results, and publications yield rewards to scientists, we assume that the expected rewards v^s from a study with significant results are higher than the expected rewards v^i from a study with insignificant results.

E. *Opportunities for p-hacking*

Scientists have ample opportunity to p-hack. However, their resources—time, money, manpower, stamina—are not infinite. Hence, they cannot systematically obtain significant results (Chen, 2021). We assume that it takes a random amount of resources to conduct an experiment, and the scientist must keep the cumulative resources used below a random limit L. Once the scientist has exhausted more resources than L, she must stop working on the project. The resource limit captures the many

[2]For simplicity we focus on simple null hypotheses. For composite null hypotheses, we would use the distribution under the null hypothesis's configuration that is the easiest to reject. For example, when testing $H_0 : \mathbb{E}(X) \leq \mu_0$ versus $H_1 : \mathbb{E}(X) > \mu_0$, we would use the distribution of the test statistic at the point $\mathbb{E}(X) = \mu_0$.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

resource constraints faced by scientists: limited access to data, limited funding, limited coauthor time, limited time before publication of similar results by competing research teams, limited stamina to work on specific projects, or limited time before the opportunity to work on more promising projects arises. Following Ferguson (2007, p. 4.12), we assume that the resource limit has an exponential distribution with rate $\lambda > 0$, so $\mathbb{P}(L > l) = \exp(-\lambda l)$ for any $l > 0$.[3]

F. P-hacking process

Experiments. The experiments are denoted by $n = 0, 1, 2, \ldots, \infty$, with $n = 0$ corresponding to not starting the research project. It takes a random amount of resources to conduct an experiment and collect a dataset. The cumulative amount of resources required to complete $1, 2, \ldots$ experiments is $D_1, D_2, \ldots$ given by a renewal process independent of the resource limit L. That is, the resources required for each experiment, $D_1, D_2 - D_1, D_3 - D_2, \ldots$, are independent and identically distributed (iid) according to a distribution independent of L.

First experiment. If resources are exhausted before the first experiment is completed, $L < D_1$, the scientist cannot obtain any results. If resources are not exhausted when the first experiment is completed, $L > D_1$, the scientist is able to collect a first dataset and construct a test statistic. This first test statistic is T_1, which is independent of the resource variables. The scientist then decides to submit the result to a journal, or to run another experiment.

Nth experiment. If the scientist chooses to run experiment $n \geq 2$, the scientist begins collecting a nth dataset of the same size and drawn from the same underlying population as previous datasets. If resources are exhausted before experiment n is completed, $L < D_n$, the scientist must stop the project before obtaining the nth dataset and submits the best result obtained up to the previous experiment, $\max\{T_1, \ldots, T_{n-1}\}$. If resources are not exhausted, $L > D_n$, the scientist obtains the nth dataset and

[3]Here research is costless to the scientist. But the robust critical values are not modified if the scientist incurs a cost of doing research (online appendix D).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

constructs the nth statistic, T_n, which is iid with $T_1, T_2, \ldots, T_{n-1}$.[4] She may then submit the best of the n test statistics, $\max\{T_1, \ldots, T_n\}$, or she may run yet another experiment.[5]

Infinite p-hacking. $n = \infty$ corresponds to running infinitely many experiments and never reporting any result.

G. Completion probability

Following Ferguson (2007, p. 4.13), we introduce the index of the first experiment that cannot be completed before resources are exhausted: $K = \min\{n \geq 1 : D_n > L\}$. Let γ be the probability that the first experiment can be completed:

$$\gamma = \mathbb{P}(D_1 < L) = \mathbb{E}(\exp(-\lambda D_1)).$$

The index K is independent of the test statistics $T_1, T_2, \ldots$, and it has a geometric distribution with success probability $1 - \gamma$, so $\mathbb{P}(K > k) = \gamma^k$ for $k = 0, 1, 2, \ldots$.[6]

H. Payoffs

No results. If the scientist does not start the research project, she receives a payoff normalized to $y_0 = 0$. If resources are exhausted before the end of the first experiment, the scientist does not obtain any result, so she receives the same payoff of $y_1 = 0$. If the scientist never concludes the research

[4] By modeling successive test statistics as independent, we are able to derive robust critical values that control the probability of type 1 error across a wide variety of common p-hacking strategies that induce positive dependence—without having to specify which particular p-hacking strategy was used by the scientist (online appendix C.1).

[5] Here the scientist analyzes the datasets obtained from successive experiments in isolation. The scientist might instead pool the datasets and analyze the pooled data. Thankfully, the robust critical values computed here maintain the probability of type 1 error below the significance level with data pooling (online appendix C.2).

[6] Here each experiment is completed with the same probability γ. Experiments might instead become more and more difficult to run and less and less likely to be completed. Fortunately, the robust critical values computed here maintain the probability of type 1 error below the significance level with increasingly difficult experiments (online appendix F).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

project and keeps on p-hacking forever, she also receives a payoff $y_\infty = 0$. In all other cases, she receives a positive payoff.

Exhausted resources. The scientist cannot continue p-hacking once the project resources are exhausted. To capture the constraint, we set to zero all payoffs once resources are exhausted: $y_n = 0$ in any step $n > K$. With these payoffs, the scientist never continues past step K. At step K, the scientist cannot obtain a new test statistic, but she can submit for publication the best test statistic from the previous $K-1$ hypothesis tests, $\max\{T_1, \ldots, T_{K-1}\}$. If the statistic is significant, the payoff is $y_K = v^s$; if the statistic is not significant, the payoff is $y_K = v^i$.

Non-exhausted resources. Any experiment $n < K$ can be completed before running out of resources, so the scientist can submit the best statistic from the n previous tests, $\max\{T_1, \ldots, T_n\}$. If the statistic is significant, the payoff is $y_n = v^s$; if not, the payoff is $y_n = v^i$.[7]

III. Optimal stopping time

The scientist p-hacks as long as she wishes. At each experiment, she may decide to stop and receive a payoff, or she may decide to continue to the next experiment. If she is able to complete the next experiment, she computes another test statistic. The scientist's problem, which we now solve, is to choose a time to stop p-hacking so as to maximize expected payoffs.

A. *Scientist's problem*

The stopping rule chosen by the scientist, the critical value z, and the random research events determine the random time $N(z)$ at which the scientist stops p-hacking. The problem of the scientist is to choose a stopping time to maximize expected payoffs.

[7]Here the scientist does not discount the future, so a significant result yields the same payoff irrespective of when it is obtained. But the robust critical values are not modified if the scientist discounts future payoffs (online appendix E).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

B. *Reported statistic*

As long as she is able to complete at least one hypothesis test, the scientist reports a random statistic $R(z)$ upon stopping. This is the best test statistic that she has been able to obtain through p-hacking. It may be significant or insignificant, and the scientist may be able to publish it or not.

C. *Existence of the optimal stopping time*

An optimal stopping time $N(z)$ exists because two conditions are satisfied (Ferguson, 2007, p. 3.3). Let Y_n denote the random payoff received by the scientist when she stops at time n. The first condition is that $\sup_n Y_n < \infty$ almost surely. This is because $Y_n \leq v^s$ almost surely. The second condition is that $Y_n \to y_\infty$ almost surely as $n \to \infty$. This is because resources inevitably run out, after which point the payoff is $Y_n = 0$, which is just the same as the payoff $y_\infty = 0$ if the scientist never stops experimenting.

D. *Characteristics of the optimal stopping time*

Because $\sup_n Y_n < \infty$ almost surely and $Y_n \to y_\infty$ almost surely as $n \to \infty$, the optimal stopping time is given by the principle of optimality: the scientist optimally stops as soon as she receives a payoff that is at least as high as the best payoff that can be expected by continuing (Ferguson, 2007, pp. 3.6–3.7). We now characterize the optimal stopping time by considering the various situations faced by the scientist.

Starting the research project. If the scientist does not start the research project, she receives $Y_0 = 0$. In contrast, if she starts she earns a nonnegative payoff: 0 if resources are exhausted before the first experiment is completed; v^i if she obtains an insignificant result; or v^s if she obtains a significant result. Hence it is always optimal to start the research project.

Continuing after insignificant results. How does the scientist behave when she still has resources to allocate to the project? A first possibility is that the result at experiment n and all the results

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

before that are insignificant. Since the best result found by the scientist is insignificant, the scientist earns $Y_n = v^i$ by stopping at experiment n. All possible payoffs are more than the payoff received for an insignificant result, v^i, so all expected payoffs are more than v^i. Since the scientist is expected to obtain more than v^i by continuing, it is not optimal to stop without obtaining a significant result.

Stopping after a significant result. If the result of test n is significant, the best result found by the scientist is significant, so the scientist earns $Y_n = v^s$ by stopping at experiment n. All possible payoffs are less than the payoff received for a significant result, v^s, so all expected payoffs are less than v^s. Hence, the scientist cannot do better by continuing. She optimally stops at experiment n and reports $R(z) = \max\{T_1, \ldots, T_n\} > z$. In fact, the principle of optimality indicates that she should stop at the first occurrence of a significant result.

Stopping when resources are depleted. Once resources are depleted, the scientist must stop p-hacking. Hence, she stops at step K if she had not stopped before. Two possibilities emerge. If $K = 1$, resources are depleted before the first experiment, so the scientist has nothing to report. If $K > 1$, the scientist submits the best test statistic that she has collected. The best result is necessarily insignificant, otherwise she would have stopped before. So she reports $R(z) = \max\{T_1, \ldots, T_{K-1}\} \leq z$.

Summary. The optimality principle gives the following results:

Lemma 1. *The scientist stops when she obtains a significant result or when she runs out of resources, whichever comes first. In the former case the scientist reports a significant result; in the latter case she reports an insignificant result. So the scientist p-hacks: she never stops at insignificant results, unless she runs out of resources.*

IV. Critical value robust to p-hacking

Based on the scientist's p-hacking strategy, we compute the critical value robust to p-hacking. This critical value ensures that the probability of type 1 error remains below the significance level even as the scientist adjusts her p-hacking behavior to the critical value itself.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

A. *Distribution of the optimal stopping time*

We compute the distribution of the optimal stopping time. Since the distribution is used to calculate the robust critical value, we compute it under the null hypothesis.

Probability of finding a significant result at experiment n. Under the null hypothesis, the probability that the test statistic from experiment n reaches the critical value z is given by the test statistic's survival function: $\mathbb{P}(T_n > z) = S(z)$, where $\mathbb{P}$ denotes the probability measure under H_0. Conversely, the probability that the test statistic does not reach z is given by the test statistic's cumulative distribution function: $\mathbb{P}(T_n < z) = 1 - S(z) = F(z)$.

Probability of continuing after experiment n. The scientist continues p-hacking after any experiment if she has not run out of resources during that experiment, which happens with probability γ, and the latest result is insignificant, which happens with probability $1 - S(z) = F(z)$. The two events are independent, so the probability that the scientist continues p-hacking is $\gamma F(z)$. Conversely, the probability that the scientist stops at any experiment is

$$1 - \gamma F(z). \tag{2}$$

Geometric distribution of the optimal stopping time. The probability of stopping at each experiment is constant, given by (2). The optimal stopping time therefore has a geometric distribution with success probability (2). The probability that the optimal stopping time is $n \geq 1$ is

$$\mathbb{P}(N(z) = n) = [\gamma F(z)]^{n-1} [1 - \gamma F(z)] .$$

Expected number of experiments. Given that the optimal stopping time has a geometric distribution with success probability (2), we obtain the following result:

Proposition 1. *When the critical value is set to z, the expected number of experiments under the*

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

null hypothesis is

$$\mathbb{E}(N(z)) = \frac{1}{1 - \gamma F(z)}, \tag{3}$$

where $\mathbb{E}$ *denotes the expectation operator under* H_0*. P-hacking is prevalent (*$\mathbb{E}(N(z)) > 1$*). And scientists p-hack more when the standard for significance is more stringent (*$\mathbb{E}(N(z))$ *is higher when* z *is higher).*

Since classical critical values are defined by (1), we infer the following result:

Corollary 1. *With a classical critical value z, the expected number of experiments under the null hypothesis is*

$$\mathbb{E}(N(z)) = \frac{1}{1 - \gamma(1 - \alpha)}. \tag{4}$$

Scientists p-hack more when the significance level is lower ($\mathbb{E}(N(z))$ *is higher when* α *is lower).*

P-hacking under the alternative hypothesis. In (4), $1 - \alpha$ represents the probability of obtaining an insignificant result from an experiment when the classical critical value is used to determine significance and the null hypothesis is true. When the alternative hypothesis is true instead, the probability of obtaining an insignificant result becomes β, where $1 - \beta$ is the power of the hypothesis test. Hence, if the alternative hypothesis is true, the expected number of experiments is $1/(1 - \beta\gamma)$. In many fields, hypothesis tests are acceptable only if their power is above 80% (Duflo, Glennerster, & Kremer, 2007, p. 3928). Setting power to $1 - \beta = 80\%$, we find that the expected number of experiments under the alternative hypothesis is $1/(1 - 0.2 \times \gamma) < 1/(1 - 0.2) = 1.25$. So there is almost no p-hacking—which is unsurprising. If the alternative hypothesis is true and the study is well powered, the null hypothesis is rejected most of the time, which renders p-hacking unnecessary. Hence, if we observe a lot of p-hacking, either the alternative hypothesis is false, or the alternative hypothesis is true but tests have low power (Ioannidis, 2005).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

B. Probability of type 1 error

Next, we compute the probability of type 1 error as a function of the critical value.

Proposition 2. *When the critical value is set to z, the probability of finding a type 1 error in a reported hypothesis test is*

$$S^*(z) = \frac{S(z)}{1 - \gamma F(z)}. \tag{5}$$

The probability of type 1 error is larger when scientists p-hack ($S^*(z) > S(z)$). In fact, the probability of type 1 error grows linearly with the expected number of experiments under the null hypothesis:

$$S^*(z) = S(z) \times \mathbb{E}(N(z)). \tag{6}$$

The proof is in online appendix A.1; it relies on an appropriate application of the law of total probability. Since classical critical values are defined by (1), we infer the following:

Corollary 2. *Under a classical critical value z, the probability of type 1 error is larger than the significance level α:*

$$S^*(z) = \frac{\alpha}{1 - \gamma(1 - \alpha)} > \alpha. \tag{7}$$

When scientists p-hack under classical critical values, the probability of type 1 error exceeds the significance level. Hence, the standards for significance set by classical critical values are too low: significance is reached more often than purported by the test's significance level. This is problematic because hypothesis tests are only informative insofar as true null hypotheses are not rejected more often than the significance level.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

C. Robust critical value

Influence of the critical value on the type 1 error rate. The critical value influences the probability of type 1 error through two channels (equation (6)). The first is a mechanical channel: a higher critical value reduces the probability that a test statistic exceeds it ($S(z)$ is decreasing in z). The second is a behavioral channel: a higher critical value pushes scientists to p-hack more in hope of finding a significant result ($\mathbb{E}(N(z))$ is increasing in z). The novelty of our correction for p-hacking is to take into account this behavioral channel.

Computing the robust critical value. The robust critical value ensures that the probability of type 1 error equals the significance level α when scientists p-hack. Since the probability of type 1 error with p-hacking is given by (5), the robust critical value z^* is implicitly defined by

$$\frac{S(z^*)}{1-\gamma F(z^*)} = \alpha. \tag{8}$$

From this definition we obtain the following result (proof details are in online appendix A.2):

Proposition 3. *For an hypothesis test with significance level α, the robust critical value is*

$$z^* = Z\left(\alpha \cdot \frac{1-\gamma}{1-\alpha\gamma}\right). \tag{9}$$

The robust critical value is always larger than the classical critical value ($z^* > Z(\alpha)$).

P-hacking with a robust critical value. The robust critical value corrects the distortion introduced by p-hacking without eliminating p-hacking. Because the significance standard imposed by the robust critical value is more stringent than the classical standard, scientists p-hack more under the robust critical value (proposition 1). In fact, combining (3) and (8), we obtain the following result:

Corollary 3. *With a robust critical value z^*, the expected number of experiments under the null*

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

hypothesis is

$$\mathbb{E}(N(z^*)) = \frac{1-\alpha\gamma}{1-\gamma}. \tag{10}$$

Scientists p-hack more when the significance level is lower ($\mathbb{E}(N(z^*))$ is higher when α is lower).

D. Bonferroni correction

Our correction for p-hacking can be formulated as a nonstandard Bonferroni correction:

Corollary 4. *The critical value that achieves a significance level α under p-hacking is the critical value that achieves a significance level*

$$\alpha^* = \frac{\alpha}{\mathbb{E}(N(z^*))} \tag{11}$$

under classical conditions, where $\mathbb{E}(N(z^*))$ is the expected number of experiments under the null hypothesis when the robust critical value is in place.

Equation (11) is obtained by evaluating (6) at z^*, and using $\alpha^* = S(z^*)$ and $S^*(z^*) = \alpha$. Unlike a standard Bonferroni correction, the number of experiments used for the correction is not observed. Rather, it is the expected number of experiments under the robust critical value when the null hypothesis is true. Thanks to the model, we can link this number to the probability γ, which we can then calibrate (section V).

E. Influence of the completion probability

Finally, we discuss how the results are influenced by the completion probability γ, which is the main parameter of the model.

Higher completion probability. From equations (3), (5), and (9), we obtain the following:

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Corollary 5. *Consider a situation with a higher completion probability. For a given critical value, scientists p-hack more, so type 1 errors occur more frequently. As a result, the robust critical value is higher.*

Hence, critical values should be higher for research teams with more resources—more time, more money, or more manpower. Research teams with more resources are less likely to be forced to interrupt an experiment before completion, so they can p-hack more. To control their type 1 error rate properly, a higher critical value is required. Critical values should also be raised when technological progress makes p-hacking easier. An example of such progress is the advent of online surveys and experiments in the social sciences, which have simplified the task of collecting data. Finally, critical values should be higher in fields in which p-hacking is easier.

Completion probability of 1. From (4), (7), (9), and (10), we obtain the following results:

Corollary 6. *Assume that the completion probability converges to 1. With a classical critical value, the expected number of experiments under the null hypothesis converges to* $1/\alpha$*, and the probability of type 1 error converges to 1. The robust critical value continues to exist but it diverges to infinity. When the robust critical value is in place, the expected number of experiments under the null hypothesis diverges to infinity.*

So if scientists can complete any number of experiments, they will continue experimenting until they reach significance. Since any null hypothesis is eventually rejected, the probability of type 1 error is 1. That is, scientists experiment to reach a foregone conclusion (Anscombe, 1954). At this limit, the robust critical value continues to exist, but it becomes arbitrarily large to offset the arbitrarily large amount of p-hacking.

V. Numerical illustration

To illustrate the amount of correction that p-hacking might require, we calibrate the completion probability γ from the lifecycle of studies in medical science. We then compute the resulting robust critical values.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

A. *Completion probability in medical science*

Calibration method. In the model, with probability $1-\gamma$, the first experiment cannot be completed before running out of resources. The probability $1-\gamma$ therefore is the share of studies that stop before completion, while the probability γ is the share of studies that are completed. We use data collected by Dwan et al. (2008) to calibrate γ (table 1). Dwan et al. review 16 metastudies that each follow a cohort of medical studies. The studies are followed from protocol approval to publication, so we can measure the fraction of studies that were stopped before completion and thus γ.

Studies that never started. Overall the data include 5736 approved studies. We focus on the 4563 studies whose fate is known. The information about these studies is obtained both by surveying the scientists who conducted them and by searching the literature. In the pool, 658 studies never started, or 658/4563 = 14.4%.

Studies that started but stopped without analysis. In addition, not all the studies that started were completed. Of the 3905 studies that started, 228 were still ongoing when the metastudies were written, so 3677 studies started and stopped. Of these, 243 stopped before any analysis could be conducted, or 243/3677 = 6.6%.

Calibrated value of the completion probability. Adding the studies that never started to those that stopped without analysis, we find that $14.4\% + (1-14.4\%)\times 6.6\% = 20.0\%$ of the approved studies could not be completed. This yields a completion probability of $\gamma = 1-20.0\% = 80.0\%$.

B. *Robust critical values*

We now compute robust critical values using the Bonferroni correction (11) and the completion probability observed in medical science, $\gamma = 80\%$.

Simple Bonferroni correction. Since the significance level α is always less than 10%, and since γ is less than 1, $1-\alpha\gamma$ is close to 1, and the expected number of experiments under the null hypothesis

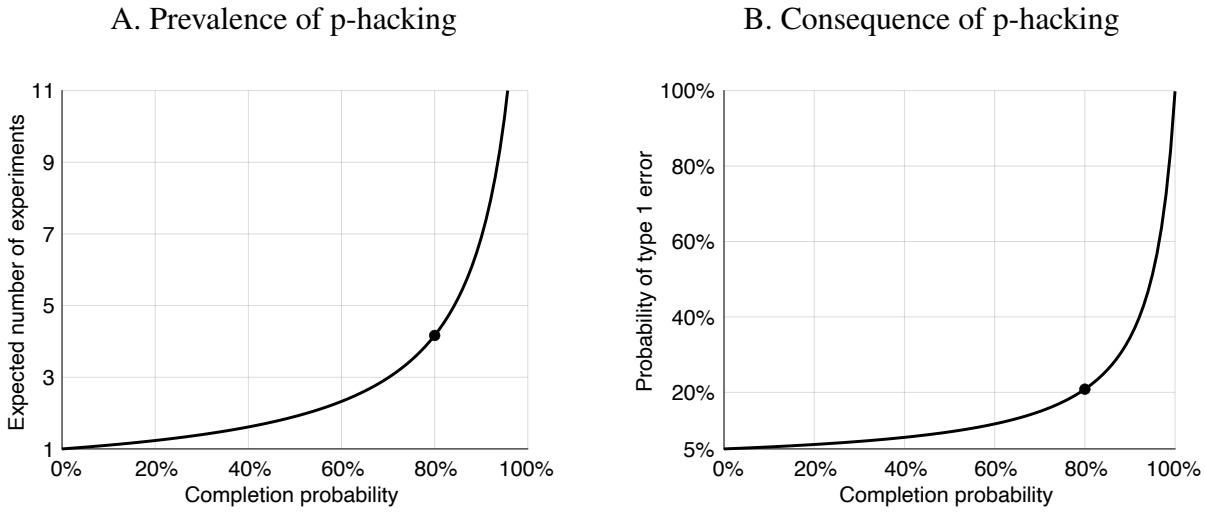
TABLE 1. INCOMPLETE STUDIES IN MEDICAL SCIENCE

Metastudy	Location in Dwan et al. (2008)	Years	Number of studies				
			Approved	With information	Never started	Ongoing	Stopped without analysis
Chan et al. (2004a)	figure 3	1994–2003	304	274	24	2	38
Easterbrook et al. (1991)	figure 4	1984–1990	715	500	113	42	28
Dickersin, Min, & Meinert (1992)	figure 5	1980–1988	921	698	184	NA	NA
Dickersin & Min (1993)	figure 6	1979–1988	310	270	17	NA	NA
Stern & Simes (1997)	figure 7	1979–1992	748	520	100	63	64
Cooper, DeNeve, & Charlton (1997)	figure 8	1986–	178	159	4	0	2
Wormald et al. (1997)	figure 9	1963–1997	61	56	5	2	10
Ioannidis (1998)	figure 10	1986–1996	109	109	0	35	8
Pich et al. (2003)	figure 11	1997–2001	158	154	11	20	20
Cronin & Sheldon (2004)	figure 12	1993–1998	101	71	0	0	NA
Decullier, Lheritier, & Chapuis (2005)	figure 13	1994–2002	976	649	68	16	51
Decullier & Chapuis (2006)	figure 14	1997–2003	142	114	21	29	12
Hahn, Williamson, & Hutton (2002)	figure 15	1990–1995	56	40	3	0	10
Chan et al. (2004b)	figure 16	1990–2003	108	105	0	17	NA
Ghersi (2006)	figure 17	1992–	318	318	92	0	0
von Elm et al. (2008)	figure 18	1988–2006	531	526	16	2	NA
Aggregate		1963–2006	5736	4563	658	228	243

Source: Dwan et al. (2008). The studies with information are all approved studies minus studies for which there is no information and studies that are excluded from the metastudy (for instance because the researchers declined to participate in the metastudy). The studies that stopped without analysis are all studies that stopped early minus studies in which an interim analysis was conducted. NA indicates that the information is not available in the metastudy.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

FIGURE 1. P-HACKING WITH A SIGNIFICANCE LEVEL OF 5% AND CLASSICAL CRITICAL VALUE



A: The curve gives the expected number of experiments run by a scientist as a function of the probability of completing an experiment when the significance level is 5% and significance is determined by a classical critical value. It is obtained from (4) with $\alpha = 5\%$. B: The curve gives the probability of type 1 error as a function of the probability of completing an experiment when the significance level is 5%, significance is determined by a classical critical value, and the scientist optimally p-hacks. It is obtained from (7) with $\alpha = 5\%$. The black dots mark the calibrated value of the completion probability: $\gamma = 80\%$.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

with the robust critical value in place is close to $1/(1-\gamma)$ (equation (10)). Using equation (11), we therefore obtain a simple Bonferroni correction against p-hacking. The classical significance level α^* required to correct p-hacking is approximately $1-\gamma$ times the actual significance level α:

$$\alpha^* \approx (1-\gamma)\alpha. \tag{12}$$

Numerical application. With $\gamma = 80\%$, the classical significance level required to address p-hacking is one fifth of the actual significance level:

$$\alpha^* = (1-0.8) \times \alpha = \frac{\alpha}{5}.$$

For instance, the critical value that achieves a significance level of 5% under p-hacking is the critical value that yields a significance level of 5%/5 = 1% under classical conditions. The rule of thumb works for any test statistic. For a z-test with a significance level of 5%, the robust critical value is 2.33 instead of 1.64 if the test is one-sided, and 2.58 instead of 1.96 if the test is two-sided. These robust critical values also apply to a large-sample t-test with a significance level of 5%.

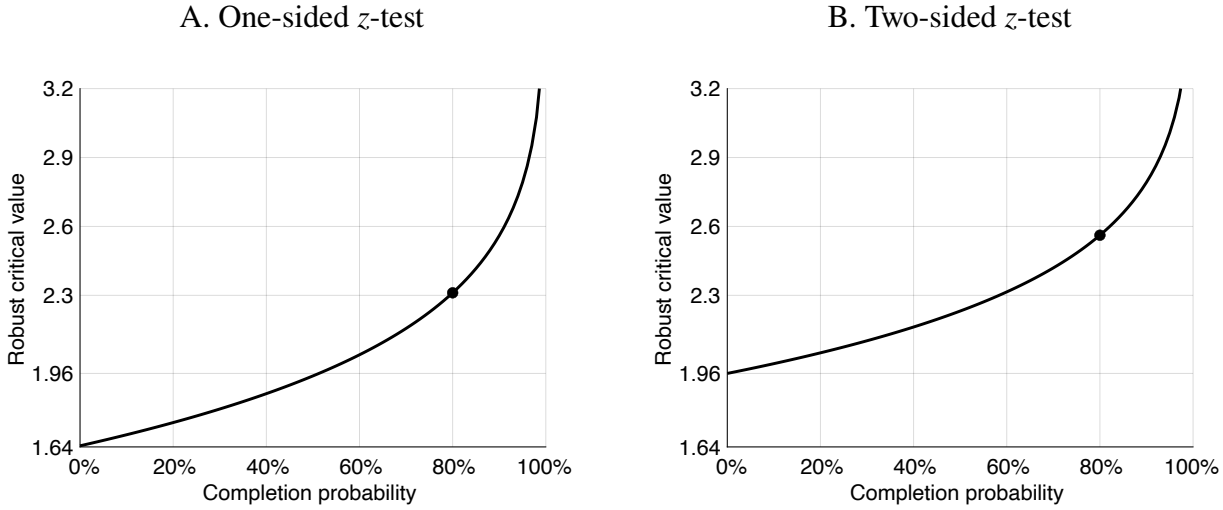
Comparison with the Benjamin et al. (2018) proposal. To address the replication crisis in science, Benjamin et al. (2018) propose that scientists replace the standard significance level of 5% by a lower significance level of 0.5%. Such tenfold reduction in the significance level is a more aggressive response to p-hacking than the fivefold reduction obtained in this numerical exercise. However, a tenfold reduction in significance level would be appropriate for a completion probability of $\gamma = 90\%$ (equation (12)). In that way, our analysis provides a theoretical underpinning for proposals to reduce the significance levels used in science. It also links the proposed reductions to the amount of resources available to scientists for p-hacking.

C. Additional numerical results

Here we provide additional numerical results. We fix the significance level at 5%.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

FIGURE 2. CRITICAL VALUES ROBUST TO P-HACKING FOR z-TESTS WITH SIGNIFICANCE LEVEL OF 5%



A: The curve gives the critical value robust to p-hacking for a one-sided z-test with significance level of 5%, as a function of the probability of completing an experiment. It is obtained from (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard normal distribution. B: The curve gives the critical value robust to p-hacking for a two-sided z-test with significance level of 5%, as a function of the probability of completing an experiment. It is obtained from (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard half-normal distribution. The black dots mark the calibrated value of the completion probability: $\gamma = 80\%$.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Prevailing p-hacking. The amount of p-hacking under classical critical values is given by (4). For the completion probability of 80%, the expected number of experiments under the null hypothesis is 4.2 (figure 1A). Moreover, the amount of p-hacking is increasing with the completion probability. For instance, when the completion probability increases from 70% to 90%, the expected number of experiments grows from 3.0 to 6.9.

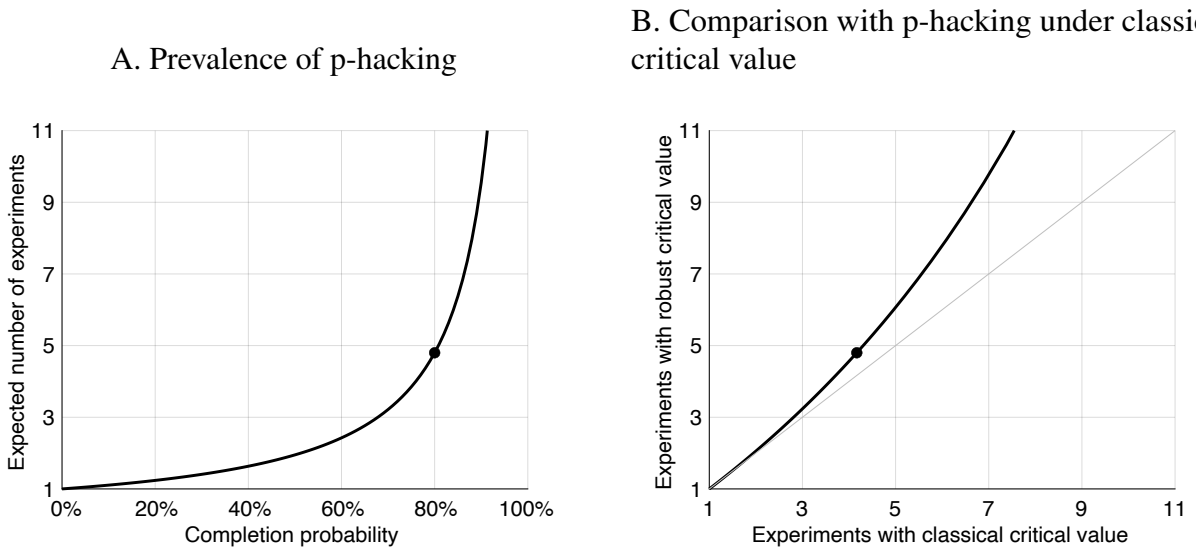
Prevailing probability of type 1 error. The probability of type 1 error under classical critical values is given by (7). For the completion probability of 80%, although the significance level is 5%, the probability of type 1 error is 21% (figure 1B). So in this case, p-hacking quadruples the probability of type 1 error. Moreover, the distortion caused by p-hacking is more severe when the completion probability is larger. For instance, when the completion probability increases from 70% to 90%, the probability of type 1 error increases from 15% to 34%.

Robust critical value for one-sided z-test. We calculate robust critical values when the underlying test statistic has a standard normal distribution under H_0, as in the common z-test, or in a t-test conducted from a large sample. We begin by calculating the robust critical value for a one-sided z-test. The critical value is given by (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard normal distribution: $Z(x) = \Phi^{-1}(1-x)$ where Φ is the standard normal cumulative distribution function. For the completion probability $\gamma = 80\%$, the robust critical value is 2.31, almost equal to the value of 2.33 given by the rule of thumb (12) (figure 2A).

Robust critical value for two-sided z-test. Next we calculate the robust critical value for a two-sided z-test. The critical value is now given by (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard half-normal distribution: $Z(x) = \Phi^{-1}(1-x/2)$. For the completion probability $\gamma = 80\%$, the robust critical value is 2.56, almost equal to the value of 2.58 given by the rule of thumb (12) (figure 2B).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

FIGURE 3. P-HACKING WITH A SIGNIFICANCE LEVEL OF 5% AND ROBUST CRITICAL VALUE



A: The curve gives the number of experiments run by a scientist, in expectation under the null hypothesis. The number is a function of the probability of completing an experiment, when the significance level is 5% and significance is determined by a robust critical value. It is obtained from (10) with $\alpha = 5\%$. B: The curve simultaneously gives the expected numbers of experiments run by a scientist under classical critical value (horizontal axis) and under robust critical value (vertical axis), for any probability of completing an experiment, and for a significance level of 5%. The numbers are obtained from (4) and (10) with $\alpha = 5\%$ and $\gamma \in (0, 1)$. The black dots mark the calibrated value of the completion probability: $\gamma = 80\%$.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Sensitivity to the completion probability. Robust critical values are increasing with the completion probability, but they are not very sensitive to it. For instance, as long as the completion probability remains between 70% and 90%, the robust critical value for one-sided z-tests remains between 2.16 and 2.56 (figure 2A), and the robust critical value for two-sided z-tests remains between 2.42 and 2.79 (figure 2B). This is reassuring: robust critical values remain close even in fields with different p-hacking intensity.

P-hacking with a robust critical value. The expected number of experiments under the null hypothesis and with a robust critical value is given by (10). For the completion probability of 80%, the expected number of experiments is 4.8 (figure 3A). Moreover, the amount of p-hacking is increasing with the completion probability. For instance, when the completion probability increases from 70% to 90%, the expected number of experiments grows from 3.2 to 9.6. Further, p-hacking is more prevalent under a robust critical value than under a classical critical value (figure 3B). At the completion probability of 80%, the expected number of experiments is 4.2 under a classical critical value but 4.8 under a robust critical value.

D. Iterative correction for p-hacking

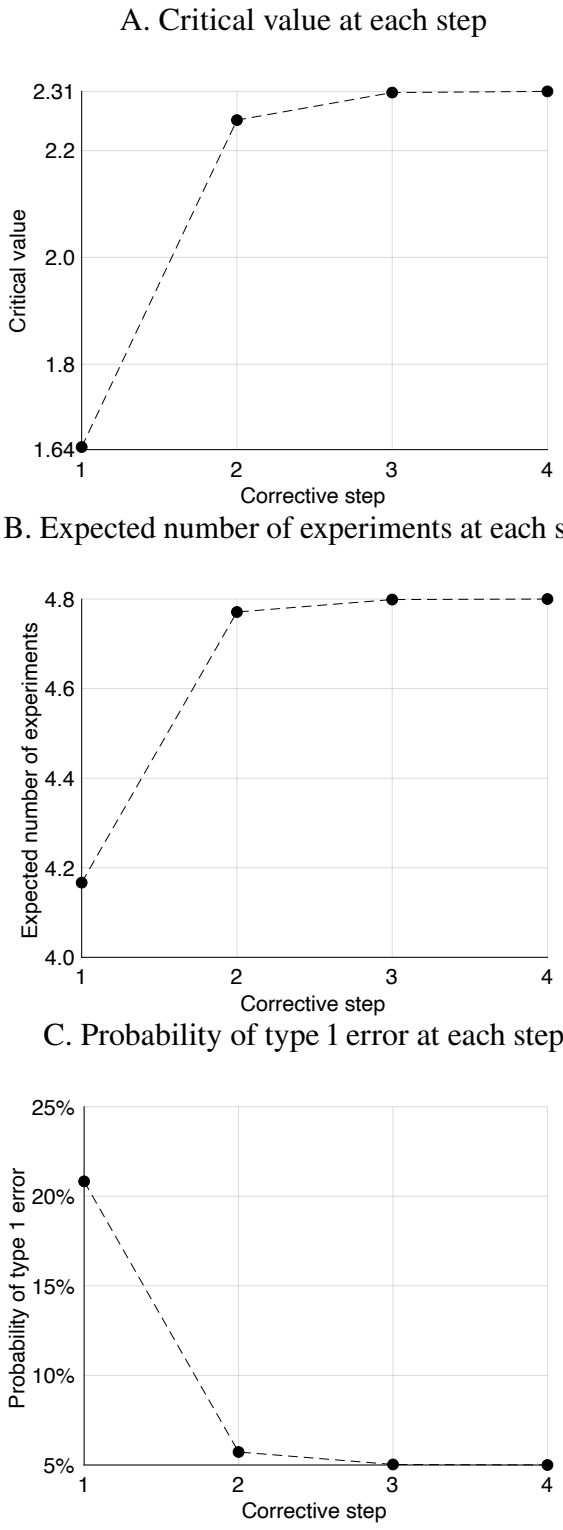
The corrections for p-hacking proposed by Anscombe (1954), Lovell (1983), and Glaeser (2008) take scientists' p-hacking behavior as fixed, whereas this paper's correction accounts for the fact that scientists would change their p-hacking behavior as soon as the correction is implemented. Here we numerically illustrate the difference between the two approaches.

Setup. For concreteness, we consider a one-sided z-test with a significance level of $\alpha = 5\%$. We calibrate the completion probability to $\gamma = 80\%$.

Initial step. The critical value is initially set to its classical value: $z_1 = \Phi^{-1}(95\%) = 1.64$. The expected number of experiments under the null hypothesis is $\mathbb{E}(N(z_1)) = 1/[1 - 0.8 \times \Phi(z_1)] = 4.17$, from (3). Accordingly, the probability of type 1 error is much greater than 5%: $S^*(z_1) = [1 - \Phi(z_1)] \times$

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

FIGURE 4. ITERATIVE BONFERRONI CORRECTION FOR P-HACKING IN ONE-SIDED z-TEST WITH SIGNIFICANCE LEVEL OF 5%



A: Step 1 uses the classical critical value, $z_1 = \Phi^{-1}(95\%) = 1.64$, where Φ is the standard normal cumulative distribution function. At step $i \geq 2$, the critical value is obtained from a Bonferroni correction using the number of experiments in the previous step: $z_i = \Phi^{-1}(1 - 5\%/\mathbb{E}(N(z_{i-1})))$. B: The expected number of experiments at each step, $\mathbb{E}(N(z_i))$, comes from (3) with $\gamma = 80\%$, $F = \Phi$, and $z = z_i$. C: The probability of type 1 error at each step comes from (5) with $\gamma = 80\%$, $F = \Phi$, $S = 1 - \Phi$, and $z = z_i$.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

$\mathbb{E}(N(z_1)) = 20.8\%$, from (6).

Bonferroni correction. In the next step, we apply the correction discussed in the literature. The critical value is obtained from a Bonferroni correction that uses the average number of experiments calculated in the initial step. So the critical value is $z_2 = \Phi^{-1}(1 - 5\%/4.17) = 2.25$. The expected number of experiments is $\mathbb{E}(N(z_2)) = 1/[1 - 0.8 \times \Phi(z_2)] = 4.77$, from (3). Since the critical value is higher than initially, the number of experiments is higher. Scientists p-hack more in response to the more stringent significance standard, which warrants additional correction. The probability of type 1 error remains greater than 5%, although it is much lower than without any correction: $S^*(z_2) = [1 - \Phi(z_2)] \times \mathbb{E}(N(z_2)) = 5.7\%$, from (6).

Next steps. We then iterate the Bonferroni correction. At any step $i \geq 2$, the critical value is given by a Bonferroni correction that uses the average number of experiments calculated in the previous step. So the critical value is $z_i = \Phi^{-1}(1 - 5\%/\mathbb{E}\left(N(z_{i-1})\right))$. The expected number of experiments is $\mathbb{E}(N(z_i)) = 1/[1-0.8\times\Phi(z_i)]$, from (3). The probability of type 1 error is $S^*(z_i) = [1-\Phi(z_i)]\times \mathbb{E}(N(z_i))$, from (6).

Results. The results of the iterative procedure are displayed in figure 4. The sequence of critical values given by the procedure rapidly approaches the robust critical value (figure 4A). By step 3, the critical value is very close to the robust critical value, $z^* = 2.31$, and the probability of type 1 error is very close to the significance level of 5% (figure 4C). We clearly see how scientists respond to an increase in critical value: by p-hacking more (figure 4B). Another take-away is that the iterative application of the Bonferroni correction converges to the robust critical value computed in proposition 3.

VI. Conclusion

We conclude by summarizing our results and comparing our approach with the registration of pre-analysis plans.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

A. *Summary*

Scientific journals prefer publishing significant results. Publications, in turn, determine a scientist's career path: promotions, salary, and honors. Scientists therefore have strong incentives to hunt for statistical significance. Such p-hacking reduces the informativeness of hypothesis tests, threatening the credibility of science. To address this problem, we develop a model of hypothesis testing with p-hacking. From it, we derive critical values robust to p-hacking, which guarantee that spurious significant results do not occur more often than intended. Robust critical values allow for p-hacking so they are larger than classical critical values. To quantify the correction that p-hacking might require, we calibrate the model using evidence from medical science. For a z-test with significance level of 5%, the robust critical values are 2.33 if the test is one-sided and 2.58 if the test is two-sided—somewhat higher than the classical critical values of 1.64 and 1.96.

B. *Comparison with the registration of pre-analysis plans*

A popular solution to p-hacking is to ask scientists to register pre-analysis plans (Miguel et al., 2014; Christensen & Miguel, 2018; Nosek et al., 2018; Adda, Decker, & Ottaviani, 2020). Although strict adherence to pre-analysis plans prevents certain forms of p-hacking, it also prevents scientific exploration, which is a prerequisite to scientific discovery (Gelman & Loken, 2014). By contrast, robust critical values can be used exactly like classical critical values, without preventing exploration. Another concern with pre-analysis plans is that they do not prevent scientists from repeating experiments. A plan could be registered for each experiment until an experiment delivers a significant result, which the scientist would then report with its accompanying pre-analysis plan. Therefore, even when pre-analysis plans are appropriate, it might make sense to use them in conjunction with robust critical values.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

References

Adda, Jerome, Christian Decker, and Marco Ottaviani,"P-hacking in Clinical Trials and How Incentives Shape the Distribution of Results Across Phases," *Proceedings of the National Academy of Sciences* 117 (2020), 13386–13392.

Akerlof, George A., and Pascal Michaillat,"Persistence of False Paradigms in Low-Power Sciences," *Proceedings of the National Academy of Sciences* 115 (2018), 13228–13233.

Andrews, Isaiah, and Maximilian Kasy,"Identification of and Correction for Publication Bias," *American Economic Review* 109 (2019), 2766–2794.

Anscombe, Francis J.,"Fixed-Sample-Size Analysis of Sequential Observations," *Biometrics* 10 (1954), 89–100.

Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts,"The Rules of the Game Called Psychological Science," *Perspectives on Psychological Science* 7 (2012), 543–554.

Begg, Colin B., and Jesse A. Berlin,"Publication Bias: a Problem in Interpreting Medical Data," *Journal of the Royal Statistical Society (Series A)* 151 (1988), 419–445.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire,

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson,"Redefine Statistical Significance," *Nature Human Behaviour* 2 (2018), 6–10.

Chan, An-Wen, Asbjorn Hrobjartsson, Mette T. Haahr, Peter C. Gotzsche, and Douglas G. Altman,"Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles," *JAMA* 291 (2004a), 2457–2465.

Chan, An-Wen, Karmela Krleza-Jeric, Isabelle Schmid, and Douglas G. Altman,"Outcome Reporting Bias in Randomized Trials Funded by the Canadian Institutes of Health Research," *Canadian Medical Association Journal* 171 (2004b), 735–740.

Chen, Andrew Y.,"The Limits of P-hacking: Some Thought Experiments," *Journal of Finance* 76 (2021), 2447–2480.

Christensen, Garret, Jeremy Freese, and Edward Miguel, *Transparent and Reproducible Social Science Research: How to Do Open Science* (Oakland, CA: University of California Press) (2019).

Christensen, Garret, and Edward Miguel,"Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature* 56 (2018), 920–980.

Cooper, H., K. DeNeve, and K. Charlton,"Finding The Missing Science: The Fate of Studies Submitted for Review By a Human Subjects Committee," *Psychological Methods* 2 (1997), 447–452.

Cronin, Eugenia, and Trevor Sheldon,"Factors Influencing the Publication of Health Research," *International Journal of Technology Assessment in Health Care* 20 (2004), 351–355.

Decullier, Evelyne, and Francois Chapuis,"Impact of Funding on Biomedical Research: A Retrospective Cohort Study," *BMC Public Health* 6 (2006), 165.

Decullier, Evelyne, Veronique Lheritier, and Francois Chapuis,"Fate Of Biomedical Research Protocols and Publication Bias in France: Retrospective Cohort Study," *BMJ* 331 (2005), 19.

Dickersin, Kay, and Yuan-I Min,"NIH Clinical Trials and Publication Bias," Online Journal of Current Clinical Trials 50 (1993).

Dickersin, Kay, Yuan-I Min, and Curtis L. Meinert,"Factors Influencing Publication of Research

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Results: Follow-up of Applications Submitted to Two Institutional Review Boards," *JAMA* 267 (1992), 374–378.

Duflo, Esther, Rachel Glennerster, and Michael Kremer,"Using Randomization in Development Economics Research: A Toolkit," in T. Paul Schultz and John A. Strauss, eds., *Handbook of Development Economics*, vol. 4 (Amsterdam: Elsevier) (2007).

Duval, Sue, and Richard Tweedie,"Trim and Fill: A Simple Funnel-Plot–based Method of Testing and Adjusting for Publication Bias in Meta-Analysis," *Biometrics* 56 (2000), 455–463.

Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, Carrol Gamble, Davina Ghersi, John P. A. Ioannidis, John Simes, and Paula R. Williamson,"Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias," *PLoS ONE* 3 (2008), e3081.

Easterbrook, P. J., R. Gopalan, J. A. Berlin, and D. R. Matthews,"Publication Bias in Clinical Research," *Lancet* 337 (1991), 867–872.

Ferguson, Thomas S., *Optimal Stopping and Applications* (2007), https://web.archive.org/web/20200812154935/https://www.math.ucla.edu/~tom/Stopping/Contents.html.

Gelman, Andrew, and Eric Loken,"The Statistical Crisis in Science," *American Scientist* 102 (2014), 460–465.

Ghersi, Davina,"Issues in the Design, Conduct and Reporting of Clinical Trials That Impact on the Quality of Decision Making," PhD dissertation, School of Public Health, Faculty of Medicine, University of Sydney (2006).

Glaeser, Edward L.,"Researcher Incentives and Empirical Methods," in Andrew Caplin and Andrew Schotter, eds., *The Foundations of Positive and Normative Economics: A Hand Book* (New York: Oxford University Press) (2008).

Hahn, S., P. R. Williamson, and J. L. Hutton,"Investigation of Within-Study Selective Reporting in Clinical Research: Follow-up of Applications Submitted to a Local Research Ethics Committee," *Journal of Evaluation in Clinical Practice* 8 (2002), 353–359.

Hedges, Larry V.,"Modeling Publication Selection Effects in Meta-Analysis," *Statistical Science* 7

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

(1992), 246–255.

Ioannidis, John P. A.,"Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials," *JAMA* 279 (1998), 281–286.

Ioannidis, John P. A.,"Why Most Published Research Findings Are False," *PLoS Medicine* 2 (2005), e124.

Ioannidis, John P. A., Sander Greenland, Mark A. Hlatky, Muin J. Khoury, Malcolm R. Macleod, David Moher, Kenneth F. Schulz, and Robert Tibshirani,"Increasing Value and Reducing Waste in Research Design, Conduct, and Analysis," *Lancet* 383 (2014), 166–175.

Kuhn, Thomas S., *The Copernican Revolution* (Cambridge, MA: Harvard University Press) (1957).

Lindsay, D. Stephen,"Replication in Psychological Science," *Psychological Science* 26 (2015), 1827–1832.

Lovell, Michael C.,"Data Mining," *Review of Economics and Statistics* 65 (1983), 1–12.

McCrary, Justin, Garret Christensen, and Daniele Fanelli,"Conservative Tests under Satisficing Models of Publication Bias," *PLoS ONE* 11 (2016), e0149590.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan,"Promoting Transparency in Social Science Research," *Science* 343 (2014), 30–31.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor,"The Preregistration Revolution," *Proceedings of the National Academy of Sciences* 115 (2018), 2600–2606.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl,"Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability," *Perspectives on Psychological Science* 7 (2012), 615–631.

Pich, Judit, Xavier Carne, Joan-Albert Arnaiz, Begona Gomez, Antoni Trilla, and Juan Rodes,"Role of a Research Ethics Committee In Follow-Up and Publication of Results," *Lancet* 361 (2003), 1015–1016.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons,"P-curve: A Key to the File-Drawer,"

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01456

Journal of Experimental Psychology: General 143 (2014), 534.

Stanley, T. D.,"Beyond Publication Bias," *Journal of Economic Surveys* 19 (2005), 309–345.

Stanley, Tom D., and Hristos Doucouliagos,"Meta-Regression Approximations to Reduce Publication Selection Bias," *Research Synthesis Methods* 5 (2014), 60–78.

Stefan, Angelika M., and Felix D. Schoenbrodt,"Big Little Lies: A Compendium and Simulation of P-hacking Strategies," *Royal Society Open Science* 10 (2023), 220346.

Stern, Jerome M., and R. John Simes,"Publication Bias: Evidence of Delayed Publication In a Cohort Study of Clinical Research Projects," *BMJ* 315 (1997), 640.

von Elm, Erik, Alexandra Rollin, Anette Blumle, Karin Huwiler, Mark Witschi, and Matthias Egger,"Publication and Non-publication of Clinical Trials: Longitudinal Study of Applications Submitted to a Research Ethics Committee," *Swiss Medical Weekly* 138 (2008), 13–14.

Wasserstein, Ronald L., and Nicole A. Lazar,"The ASA's Statement on P-values: Context, Process, and Purpose," *American Statistician* 70 (2016), 129–133.

Wormald, R., J. Bloom, J. Evans, and K. Oldfield,"Publication Bias in Eye Trials," in *5th Annual Cochrane Colloquium* (Amsterdam: Cochrane Collaboration) (1997).