DISCUSSION PAPER SERIES

DP16942

Incentive-Compatible Critical Values

Pascal Michaillat and Adam McCloskey

LABOUR ECONOMICS

PUBLIC ECONOMICS

CEPR

ISSN 0265-8003

Incentive-Compatible Critical Values

Pascal Michaillat and Adam McCloskey

Discussion Paper DP16942
Published 21 January 2022
Submitted 20 January 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

Incentive-Compatible Critical Values

Abstract

Statistically significant results are more rewarded than insignificant ones, so researchers have the incentive to pursue statistical significance. Such p-hacking reduces the informativeness of hypothesis tests by making significant results much more common than they are supposed to be in the absence of true significance. To address this problem, we construct critical values of test statistics such that, if these values are used to determine significance, and if researchers optimally respond to these new significance standards, then significant results occur with the desired frequency. Such incentive-compatible critical values allow for p-hacking so they are larger than classical critical values. Using evidence from the social and medical sciences, we find that the incentive-compatible critical value for any test and any significance level is the classical critical value for the same test with approximately one fifth of the significance level—a form of Bonferroni correction. For instance, for a z-test with a significance level of 5%, the incentive-compatible critical value is 2.31 instead of 1.65 if the test is one-sided and 2.57 instead of 1.96 if the test is two-sided.

JEL Classification: C12, C18

Keywords: Hypothesis testing, Academic incentives, p-Hacking, Statistical significance, Optimal stopping

Pascal Michaillat - pascalmichaillat@brown.edu
Department of Economics, Brown University and CEPR

Adam McCloskey - adam.mccloskey@colorado.edu
University of Colorado, Boulder

Acknowledgements
We thank Isaiah Andrews, Brian Cadena, Kenneth Chay, Garret Christensen, Pedro Dal Bo, Stefano DellaVigna, Alexander Frankel, Peter Hull, Miles Kimball, Megan Lang, Jonathan Libgober, Edward Miguel, 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.

Incentive-Compatible Critical Values

Adam McCloskey, Pascal Michaillat

January 2022

Statistically significant results are more rewarded than insignificant ones, so researchers have the incentive to pursue statistical significance. Such p-hacking reduces the informativeness of hypothesis tests by making significant results much more common than they are supposed to be in the absence of true significance. To address this problem, we construct critical values of test statistics such that, if these values are used to determine significance, and if researchers optimally respond to these new significance standards, then significant results occur with the desired frequency. Such incentive-compatible critical values allow for p-hacking so they are larger than classical critical values. Using evidence from the social and medical sciences, we find that the incentive-compatible critical value for any test and any significance level is the classical critical value for the same test with approximately one fifth of the significance level—a form of Bonferroni correction. For instance, for a z-test with a significance level of 5%, the incentive-compatible critical value is 2.31 instead of 1.65 if the test is one-sided and 2.57 instead of 1.96 if the test is two-sided.

McCloskey: University of Colorado–Boulder. Michaillat: Brown University. We thank Isaiah Andrews, Brian Cadena, Kenneth Chay, Garret Christensen, Pedro Dal Bo, Stefano DellaVigna, Alexander Frankel, Peter Hull, Miles Kimball, Megan Lang, Jonathan Libgober, Edward Miguel, 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.

1. Introduction

P-hacking occurs when researchers engage in various behaviors that increase their chances of reporting statistically significant results (Simonsohn, Nelson, and Simmons 2014; Lindsay 2015; Wasserstein and Lazar 2016; Christensen, Freese, and Miguel 2019). Typical p-hacking practices include suppressing inconvenient experiments, halting data collection at a convenient time, dropping inconvenient observations or treatments or outcomes, applying convenient transformations to the data, choosing convenient covariates in regressions, or choosing convenient statistical specifications.

P-hacking is prevalent across scientific fields. Researchers readily admit to engaging in it (John, Loewenstein, and Prelec 2012). It is visible in the lifecycle of scientific studies: significant results are almost certain to be reported, whereas insignificant results are likely to remain unreported (Dwan et al. 2008; Franco, Malhotra, and Simonovits 2014). And its effects appear in meta-analyses: the distributions of test statistics in entire literatures show that researchers tinker with their analysis to obtain significant results (Hutton and Williamson 2000; Head et al. 2015; Brodeur et al. 2016; Vivalt 2019; Brodeur, Cook, and Heyes 2020; Elliott, Kudrin, and Wuthrich 2021).

P-hacking is so prevalent because researchers face strong incentives to p-hack (Glaeser 2008; Nosek, Spies, and Motyl 2012; Bakker, van Dijk, and Wicherts 2012). First, significant results are more rewarded than insignificant ones. This is because scientific journals prefer publishing significant results (Sterling 1959; Bozarth and Roberts 1972; Begg and Berlin 1988; Csada, James, and Espie 1996; Jennions and Moeller 2002; Song et al. 2000; Ioannidis and Trikalinos 2007; Head et al. 2015; Fanelli, Costas, and Ioannidis 2017; Christensen, Freese, and Miguel 2019; Andrews and Kasy 2019). Publications, in turn, determine a scientist's career path, including promotions, salary, and honorific rewards (Hagstrom 1965; Skeels and Fairbanks 1968; Merton 1973; Katz 1973; Siegfried and White 1973; Tuckman and Leahey 1975; Hansen, Weisbrod, and Strauss 1978; Sauer 1988; Swidler and Goldreyer 1998; Gibson, Anderson, and Tressler 2014; Biagioli and Lippman 2020). At the same time, researchers enjoy a lot of flexibility in data collection and analysis. Hence, even when the null hypothesis is true, they have ample opportunity to obtain significant results without violating scientific norms (Cole 1957; Armitage 1967; Leamer 1983; Lovell 1983; Simmons, Nelson, and Simonsohn 2011; Humphreys, de la Sierra, and van der Windt 2013; Huntington-Klein et al. 2021).

Despite its prevalence, p-hacking is not taken into account in hypothesis testing theory: the critical values used to determine significance assume no p-hacking. Therefore,

the critical values set a standard for significance that is too lax: significance is reached much more often than purported by the test's nominal significance level in the absence of true significance. This is problematic because hypothesis tests are informative only insofar as a true null hypotheses is not rejected more often than the significance level.

A number of corrections for p-hacking have been discussed (Anscombe 1954; Lovell 1983; Glaeser 2008). But these corrections face a conceptual challenge. They take the researcher's p-hacking behavior as fixed and offer a correction based on that behavior, whereas in reality the researcher would change her p-hacking behavior as soon as the correction is implemented.

Consider for instance an hypothesis tests with a 5% significance level and corresponding critical value. Classical critical values are constructed such that if the researcher examined one data sample, a true null hypothesis would be rejected no more than 5% of the time. But if a researcher collected $n > 1$ data samples, performed a hypothesis test on each sample, and reported the best of the n results—without mentioning the $n - 1$ other results—a true null hypothesis would be rejected more often than 5% of the time. Existing corrections take the number n of samples as given and compute a more stringent critical value based on it.

Yet this first step is not enough to resolve our problem. Just as researchers may collect more than one sample under the classical critical value, they may collect more than n samples under the new critical value, overwhelming the proposed correction. Indeed, researchers respond to incentives—this is why they p-hack in the first place.

This paper proposes a solution to this statistical problem. We construct critical values such that, if these values are used to determine significance, and if researchers optimally respond to the new significance standards, then significant results occur with the desired frequency. We call these *incentive-compatible critical values* (ICCV).

We consider a researcher who sequentially collects independent and identically distributed (iid) data samples, performs the hypothesis test on each sample, and reports the best result. Such data sampling is an egregious form of p-hacking, but it is a useful starting point for several reasons. First, it leads to a simple model of p-hacking in which the test statistics obtained at each p-hacking step are iid. This property makes it simple to solve the researcher's optimal p-hacking strategy and then compute the ICCV. Second, because each step of p-hacking requires to collect a new data sample, we are able to calibrate the p-hacking process from evidence on the lifecycle of studies in the social and medical sciences. The calibration then allows us to compute the ICCV. Third, unlike other forms of p-hacking that could be detected through specification

curves (Simonsohn, Simmons, and Nelson 2020; Young and Holsteen 2017) or multiverse analysis (Steegen et al. 2016), data sampling is undetectable. It is therefore particularly important to correct this form of p-hacking. Fourth, the ICCV obtained under the assumption of iid data sampling also controls the probabilities of type 1 error under many other common forms of p-hacking.

We first find researchers' optimal p-hacking strategy using results from optimal stopping theory (Ferguson 2007). The optimal strategy is to continue collecting new data samples until the result is significant. Not all studies report significant results, however, because each study must be completed before a deadline. If the researcher is unable to find a significant result before the deadline, she reports an insignificant result.

We then determine the number of data samples collected by a researcher and the probability of type 1 error, as a function of the prevailing critical value. The critical value influences the rate of type 1 error in two ways: by determining the probability that a true null hypothesis is rejected in each data sample, and by influencing the number of data samples that the researcher collects. Using these results, we compute the ICCV: it is the critical value such that when researchers p-hack optimally, type 1 errors occur at the desired rate, given by the nominal significance level.

We find that the ICCV is given by a Bonferroni correction. For any test and any significance level, the ICCV is the classical critical value for the same test with the significance level divided by the expected number of p-hacking steps at the ICCV. Accordingly, the ICCV is larger than the classical critical value for the same test and significance level.

An advantage of our simple model is that the expected number of p-hacking steps when the ICCV is in place, and thus the ICCV, are solely determined by two parameters: the significance level and the probability that each p-hacking step is completed before the deadline. We compute the completion probability using evidence from the lifecycle of studies in the social and medical sciences (Dwan et al. 2008; Franco, Malhotra, and Simonovits 2014). An accurate rule of thumb for any significance level below 10% and any test is that the ICCV is the classical critical value for the same test with one fifth of the significance level. For instance, for a z-test with a significance level of 5%, the ICCV is 2.31 instead of 1.65 if the test is one-sided and 2.57 instead of 1.96 if the test is two-sided.

In the baseline model, researchers construct a sequence of test statistics by collecting a sequence of independent datasets; therefore, the test statistics are statistically independent. But in reality researchers often construct test statistics that are positively dependent—for instance when they pool data or when they examine various statis-

tical specifications. In these scenarios the ICCV obtained under the independence assumption remains useful: it keeps the type 1 error rate below the significance level.

2. Model of p-hacking

We develop a simple model of p-hacking and incorporate it into statistical hypothesis testing theory. The researcher samples data, with the aim of reaching a significant result. Sampling data and conducting hypothesis tests take time, and the researcher must report her results before a deadline, so not all researchers are able to obtain significant results.

2.1. Hypothesis test

The researcher wishes to test a null hypothesis H_0 against an alternative hypothesis H_1. The data are governed by a different probability distribution under each hypothesis.

The researcher sets the test's significance level to $\alpha \in (0, 1)$. The significance level gives the desired probability of type 1 error—the error that occurs when a true null hypothesis is rejected. Common significance levels are 10%, 5%, and 1%.

2.2. Test statistic

To conduct the hypothesis test, the researcher collects a dataset. She constructs a test statistic T from this dataset, whose realization is denoted t.

Under classical conditions, there is no p-hacking: the construction of the test statistic is transparent to all readers. Everyone therefore knows the distribution of the test statistic under the null hypothesis.[1] We denote the statistic's cumulative distribution function by F, its survival function by $S = 1 - F$, and its inverse survival function by $Z = S^{-1}$.

2.3. Classical critical value

The null hypothesis is rejected when the test statistic falls into the critical region. For simplicity, we assume that the critical region takes the form (z, ∞), where z is the critical value. If the researcher obtains a test statistic $t > z$, the null hypothesis is

[1] For composite null hypotheses, we 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 use the distribution of the test statistic at the point $\mathbb{E}(X) = \mu_0$.

rejected: the result is statistically significant. But if she obtains a test statistic $t \leq z$, the null hypothesis cannot be rejected, and the result is insignificant. Accordingly, the probability of type 1 error is bounded above by $S(z)$.

The critical value is defined such that the probability of type 1 error is no greater than the significance level. Hence, it is defined implicitly by

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

and explicitly by $z = Z(\alpha)$.

2.4. Rewards accruing to statistically significant results

The first element of the p-hacking model is the rewards accruing to statistically significant results. Empirically, statistically significant results are more likely to be published than insignificant results. Moreover, a published study is expected to yield higher rewards than an unpublished study. Based on this evidence, we assume that the expected rewards v^s from a study with statistically significant results are higher than the expected rewards v^i from a study with statistically insignificant results.[2]

2.5. Opportunities and limits to p-hacking

While researchers have ample opportunity to p-hack, they do not systematically obtain significant results because they must complete their work before a random deadline, L.[3] The deadline captures future events that may force the researcher to stop working on the project: losing access to the data, losing funding, departure of a coauthor, publication of similar results by competing research teams, being denied tenure, retirement, or the opportunity to work on more promising projects. Following Ferguson (2007, p. 4.12), we assume that the deadline has an exponential distribution with rate $\lambda > 0$, so $\mathbb{P}(L > l) = \exp(-\lambda l)$ for any $l > 0$.

[2] Appendix A surveys the evidence and justifies the assumption $v^s > v^i$ from it.

[3] The deadline can be reinterpreted as a finite amount of money or a finite amount of stamina without changing the results (appendix B). In the money interpretation, it takes a random amount of money to sample data and conduct hypothesis tests, and the researcher must keep the cumulative costs below a random budget L. In the stamina interpretation, it takes a random amount of stamina to sample data and conduct hypothesis tests, and the researcher must keep the cumulative stamina expended below a random quantity L.

2.6. P-hacking timeline

The steps of the p-hacking process are denoted by $n = 0, 1, 2, \ldots, \infty$, with step 0 corresponding to not starting the research project. It takes a random amount of time to sample data and conduct hypothesis tests. The different steps of the p-hacking process are completed at dates $D_1, D_2, \ldots$ given by a renewal process independent of the deadline L. That is, the durations of each step, $D_1, D_2 - D_1, D_3 - D_2, \ldots$, are iid according to a distribution independent of L.

If the deadline occurs before the first step is completed, $L < D_1$, the researcher is not able to obtain any results. If the deadline has not occurred when the first step is completed, $L > D_1$, the researcher is able to collect a first sample of data and complete a first hypothesis test. The test statistic obtained in the first test is T_1, which is independent of the deadline and completion dates. The researcher then decides to submit this result to a scientific journal, or to engage in another step of p-hacking.

If the researcher moves to step 2, she starts by collecting a second sample. This second sample has the same size as the first, and it is drawn from the same underlying population. The researcher conducts the same hypothesis test on this second sample. Once again, if the deadline occurs before the second step is completed, $D_1 < L < D_2$, the researcher must stop the project before obtaining the second test statistic. If the deadline has not occurred before the second step is completed, $L > D_2$, the researcher obtains a second test statistic, T_2. This statistic is independent from T_1, and from the deadline and completion dates, but it has the same distribution as T_1.

As the researcher stores all the data that she collects and keeps tracks of all the statistical analyses, she is able to recall past results. If the deadline occurs in the middle of the second step, the researcher does not obtain a second test statistic, but she may still submit the statistic T_1 to a scientific journal. If the deadline does not occur, she obtains the second statistic, and she may submit the best result from the two hypothesis tests, $\max\{T_1, T_2\}$. If she does not want to submit the result, she may engage in another step of p-hacking.

More generally, at step n, the researcher collects a nth sample, of the same size and drawn from the same underlying population as all the previous samples. She conducts the same test on that sample. If the deadline occurs before step n is completed, the researcher can only report the best result obtained up to the previous stage, $\max\{T_1, \ldots, T_{n-1}\}$. If the deadline does not occur before step n is completed, the researcher obtains the nth statistic, T_n, which is iid with $T_1, T_2, \ldots, T_{n-1}$. She may then submit the best of the n test statistics, $\max\{T_1, \ldots, T_n\}$, or she may proceed to the next

step of p-hacking.

Step ∞ corresponds to collecting infinitely many samples, conducting infinitely many tests, and never reporting any result.

2.7. Completion probability

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

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

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 $k = 0, 1, 2, \ldots$.

2.8. Payoffs

If the researcher does not start the research project, she receives a payoff $y_0 = 0$. If the deadline occurs before the end of the first step, the researcher does not obtain any results, so she receives the same payoff of $y_1 = 0$. If the researcher never concludes the research project and keeps on p-hacking forever, she also receives a payoff $y_\infty = 0$. In all other cases, she will receive a positive payoff.

The researcher is not able to continue p-hacking past the deadline. To capture this constraint, we set all payoffs after the deadline to zero: $y_n = 0$ in any step $n > K$. With these payoffs, the researcher never continues past step K.

The deadline occurs right before the end of step K. The researcher cannot obtain the Kth test statistic, but she can still submit the best test statistic from the previous $K - 1$ hypothesis tests, $\max\{T_1, \ldots, T_{K-1}\}$. If that statistic is significant, the payoff is $y_K = v^s$; if it is insignificant, the payoff is $y_K = v^i$.

Any step $n < K$ is completed before the deadline, so the researcher can submit for publication the best test statistic from the n previous hypothesis tests, $\max\{T_1, \ldots, T_n\}$. If that statistic is significant, the payoff is $y_n = v^s$; if it is insignificant, the payoff is $y_n = v^i$.[4]

[4]Here research is costless to the researcher: the university or lab where she works covers all the costs. But the results are not modified if the researcher incurs a research cost—be it a monetary cost or a psychic cost (appendix C). Moreover, the researcher does not discount the future, so a significant result yields the same payoff irrespective of when it is obtained. But the results are not modified if the researcher discounts future payoffs (appendix D).

3. Optimal stopping time

The researcher p-hacks as long as she wishes. At each step, after observing all previous test statistics, she may decide to stop and receive a payoff, or she may decide to continue to the next step. If she is able to complete the next p-hacking step before the deadline, she will observe another test statistic. The researcher's problem is to choose a time to stop p-hacking to maximize expected payoffs. We now solve this problem.

3.1. Researcher's problem

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

3.2. Reported statistic

As long as she has had time to complete at least one test, the researcher 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 researcher may be able to publish it or not.

3.3. Characteristics of the optimal stopping time

An optimal stopping time $N(z)$ exists because two conditions are satisfied (Ferguson 2007, chapter 3). Let Y_n denote the random payoff received by the researcher when she stops at time n. First, $Y_n \leq v^s$ a.s., so $\sup_n Y_n < \infty$ a.s. Second, because the deadline inevitably occurs, $Y_n \xrightarrow{as} 0 = y_\infty$ as $n \to \infty$. Furthermore, the optimal stopping time is given by the principle of optimality of dynamic programming: it is optimal to stop as soon as the payoff is at least as high as the best payoff that can be expected by continuing.

3.4. Finding the optimal stopping time

If the researcher does not start the research project, she receives $Y_0 = 0$. In contrast, if she starts she earns a non-negative payoff: 0 if the deadline occurs before completion of the first step, 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.

How does the researcher behave before the deadline? A first possibility is that the result at step n and all the results before that are insignificant. Since the best result found by the researcher is insignificant, the researcher earns $Y_n = v^i$ by stopping at step 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 researcher is expected to obtain more than v^i by continuing, it is not optimal to stop without obtaining a significant result.

If the result of test n is significant, the best result found by the researcher is significant, so the researcher earns $Y_n = v^s$ by stopping at step 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 researcher cannot do better by continuing. It is therefore optimal to stop at step n and report $R(z) = \max\{T_1, \ldots, T_n\} > z$. In fact, the principle of optimality indicates that it is optimal to stop the first time that a significant result occurs.

What happens at the deadline? The deadline occurs right before the end of step K. It makes no sense to continue p-hacking past the deadline, so the researcher stops at step K if she had not stopped before. Then there are two possibilities. If $K = 1$, the deadline has occurred before the first step, so the researcher has nothing to report. If $K > 1$, the researcher submits the best test statistic that she has collected before the deadline. This best result is necessarily insignificant, otherwise she would have stopped before. So she reports $R(z) = \max\{T_1, \ldots, T_{K-1}\} \leq z$.

In sum, the optimality principle gives the following results:

LEMMA 1. *The researcher stops when she obtains a significant result or when the deadline has occurred, whichever comes first. In the former case the researcher reports a significant result; in the latter case she reports an insignificant result. So there is p-hacking: the researcher never stops at insignificant results unless she is stopped by the deadline.*

4. Incentive-compatible critical value

Based on the researcher's optimal p-hacking strategy, we compute the ICCV.

4.1. Distribution of optimal stopping time

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

Under the null hypothesis, the probability that the test statistic at step n reaches the critical value z is simply 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.

The researcher continues p-hacking past any step if the deadline has not occurred during the step, which happens with probability γ, and the result is insignificant, which happens with probability $1 - S(z)$. The two events are independent, so the probability that the researcher continues p-hacking is $\gamma[1 - S(z)]$. Conversely, the probability that the researcher stops p-hacking at any step is

$$1 - \gamma[1 - S(z)]. \tag{2}$$

At each step, the probability of stopping p-hacking is constant, given by (2). The optimal stopping time therefore has a geometric distribution with success probability (2). Accordingly, the probability that the optimal stopping time is $n \geq 1$ is

$$\mathbb{P}(N(z) = n) = \left[\gamma - \gamma S(z)\right]^{n-1} \left[1 - \gamma + \gamma S(z)\right].$$

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

PROPOSITION 1. *Under the null hypothesis, the expected number of p-hacking steps is*

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

where $\mathbb{E}$ denotes the expectation operator under H_0. P-hacking is prevalent ($\mathbb{E}(N(z)) > 1$). Moreover, researchers p-hack more when the standards for significance are more stringent (higher z).

Classical critical values are defined by (1). Accordingly, the expected number of p-hacking steps under the null hypothesis and with classical critical values is

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

P-hacking is more prevalent when the significance level is lower (lower α).

What happens if the alternative hypothesis is true instead of the null? In (4), $1 - \alpha$ represents the probability to obtain an insignificant result at each step 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 to obtain an insignificant result at each step becomes β, where $1 - \beta$ is the power of the hypothesis test. Hence, if the alternative hypothesis is true, the expected number of p-hacking steps is simply $1/(1 - \beta\gamma)$. In many fields, hypothesis tests are acceptable only if their power is above 80% (Duflo, Glennerster, and Kremer 2007; Christensen 2018). Setting power to $1 - \beta = 80\%$, we find that the expected number of p-hacking steps under the alternative is $1/(1 - 0.2 \times \gamma) < 1/(1 - 0.2) = 1.25$: there is almost no p-hacking. This is unsurprising. If the alternative hypothesis is true and the study is well powered, the null hypothesis is rejected most of the time, which makes p-hacking unnecessary. Hence, if we see a lot of p-hacking, either the alternative hypothesis is false, or the alternative hypothesis is true but tests have little power (Ioannidis 2005).

4.2. 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 study is*

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

The probability of type 1 error is larger when researchers p-hack ($S^*(z) > S(z)$). In fact, the probability of type 1 error grows linearly with the expected number of p-hacking steps ($\mathbb{E}(N(z))$):

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

The proof is not difficult: it relies on appropriate applications of the law of total probability and Bayes' rule. But it is not particularly interesting so we relegate it to appendix E.

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

COROLLARY 1. *Under classical critical values, the probability of type 1 error is larger than the significance level α:*

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

When researchers p-hack, the probability of type 1 error given by classical critical values is larger than the significance level α. Hence, the standards for significance set

by classical critical values are too low: significance is reached much 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.

4.3. Incentive-compatible critical value

Equation (6) shows that changing the critical value z has two effects on the probability of type 1 error. First, there is a mechanical effect, whereby a higher critical value makes it less likely that a given test statistic exceeds it ($S(z)$ is decreasing in z). Second, there is a behavioral effect, whereby the optimal stopping time and thus reported test statistic are altered by the critical value. Indeed, when the critical value is larger, researchers p-hack more in hope of reaching significance ($\mathbb{E}(N(z))$ is increasing in z). The novelty of this analysis is to propose a critical value that accounts for the behavioral effect: the ICCV. The mere fact that p-hacking exists implies that the behavioral effect is important.

The ICCV is such that the probability of type 1 error equals the significance level α when researchers p-hack. Since the probability of type 1 error with optimal p-hacking is given by (5), the ICCV z^* is implicitly defined by

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

From this implicit definition we obtain the following results (details of the proof are relegated to appendix E):

PROPOSITION 3. *For any hypothesis test with significance level α, the ICCV is given by*

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

where Z is the inverse survival function of the test statistic. The ICCV is always larger than the classical critical value $Z(\alpha)$.

The ICCV corrects for the distortion introduced by p-hacking without eliminating or even reducing p-hacking. In fact, because the significance standards imposed by the ICCV are more stringent than classical standards, researchers p-hack more under the ICCV. Combining (3) and (8), we obtain the following corollary:

COROLLARY 2. *The average number of p-hacking steps under the ICCV is*

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

4.4. Bonferroni correction

To provide another perspective on the significance standards imposed by the ICCV, we compute the type 1 error rate that would be achieved if the ICCV was used in classical conditions. The classical significance level required to deal with p-hacking is closely related to the amount of p-hacking under the ICCV:

COROLLARY 3. *Achieving a significance level α under p-hacking requires to set the critical value at the level that would be appropriate for a significance level*

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

under classical conditions.

This relation is obtained by evaluating (6) at z^*, and using $\alpha^* = S(z^*)$ and $S^*(z^*) = \alpha$.

Hence, our correction for p-hacking is just a Bonferroni correction. Unlike in a typical Bonferroni correction, however, the number of p-hacking steps used for the correction is not observed, and it is not the number of p-hacking steps prevailing under a standard critical value. Rather, it is the average number of p-hacking steps under the ICCV when the null hypothesis is true. Thanks to the model, we can link this number of steps to the probability γ, which we can calibrate by observing the lifecycle of scientific studies (section 6).

4.5. Influence of the completion probability

To finish the analysis, we discuss how the results are influenced by the completion probability—the main parameter of the model. In particular, we describe what happens when the completion probability goes to 1. We obtain the following two corollaries:

COROLLARY 4. *For a given critical value (z), when researchers p-hack faster or face a looser deadline (higher γ), researchers conduct more p-hacking steps (higher $\mathbb{E}(N(z))$), so the probability of type 1 error is higher (higher $S^*(z)$). As a result, when researchers p-hack faster or face a looser deadline, the ICCV is higher (higher z^*).*

Corollary 4 implies that ICCVs should be higher in fields in which p-hacking is easier. It also implies that ICCVs should be raised when technological progress renders p-hacking easier. An example of such progress is the advent of online surveys and experiments in social science, which have greatly simplified the task of collecting survey and experimental data. Finally, the corollary indicates that it would make sense to set higher ICCVs for research teams with more resources—more time, more money, or more manpower. Such teams are less likely to be forced to interrupt a study before completion, so they can p-hack more. To control their type 1 error rate properly, a higher ICCV is required.

COROLLARY 5. *For a given critical value (z), when researchers p-hack infinitely fast or have an infinite amount of time before the deadline ($\gamma \to 1$), researchers conduct on average $1/\alpha$ steps of p-hacking ($\mathbb{E}(N(z)) \to 1/\alpha$), and the probability of type 1 error reaches 1 ($S^*(z) \to 1$). Hence, when researchers p-hack infinitely fast or have an infinite amount of time before the deadline, the ICCV continues to exist but it reaches infinity ($z^* \to \infty$). The average number of p-hacking steps under the ICCV also reaches infinity ($\mathbb{E}(N(z^*)) \to \infty$).*

Corollary 5 implies that if researchers can complete any number of tests, they will sample data until they reach significance. Since all null hypotheses are eventually rejected, the probability of type 1 error is 1. At this limit, researchers sample data to reach a foregone conclusion (Anscombe 1954). Yet, the ICCV continues to exist when the completion probability γ gets close to 1: it becomes arbitrarily large to offset the arbitrarily large amount of p-hacking.

5. Other forms of p-hacking

We have so far assumed that the test statistics sequentially formed by the researcher are statistically independent—which happens when researchers form statistics by collecting a sequence of independent datasets. However, a p-hacker often collects test statistics that are positively dependent: for instance, when pooling data or when examining various specifications of a statistical model. We show that in these scenarios, the ICCVs obtained under the independence assumption continue to maintain control of the type 1 error rate.

PROPOSITION 4. *Rather than assuming that the sequence of test statistics $T_1, \ldots, T_n$ are independent, we assume that*

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) \leq \mathbb{P}(T_n > z) = S(z) \tag{12}$$

for all $z \geq 0$. *Then the probability of type 1 error under the ICCV* (9) *does not exceed the significance level* α.

In the common case of sequential *t*-tests, a simple condition on the covariances between successive *t*-statistics guarantees that proposition 4 applies:

PROPOSITION 5. *Suppose the sequence of test statistics are distributed as follows under* H_0*:* $(T_1, \ldots, T_n) \sim \mathcal{N}(0, \Omega(n))$*, where all the variances* $\Omega_{1,1}(n)$*, …,* $\Omega_{n,n}(n)$ *equal 1 and all covariances* $\Omega_{1,n}(n)$*,…,* $\Omega_{n-1,n}(n)$ *are non-negative. Then condition* (12) *is satisfied so proposition 4 applies.*

The proofs of the propositions are relegated to appendix E, but the intuition is simple. The optimal p-hacking strategy described by lemma 1 remains the same. Indeed, the derivation of the optimal stopping time does not rely on the independence of the test statistics; it remains valid even if the test statistics are dependent. What does change are the stochastic properties of the optimal stopping time and of the reported test statistic. However, under (12), we can guarantee that the ICCV given by (9) keeps the type 1 error rate below the significance level.

The distributional assumption in proposition 5 is satisfied by the large-sample joint distribution of a sequence of positively correlated *t*-statistics under the null hypothesis. Such positive correlation appears under common forms of p-hacking. This section provides three examples: pooling data that are collected sequentially, examining various regression specifications, and examining various instruments.

5.1. Pooling data

The researcher studies a parameter μ. The null hypothesis is $\mu = \mu_0$, and the alternative hypothesis is $\mu > \mu_0$. At each step the researcher adds data to the previous sample; the additional data are collected from the same underlying population and independent. In step n the researcher constructs an estimate $\hat{\mu}_n$ of the parameter by taking a mean from the pooled sample:

$$\hat{\mu}_n = \frac{1}{m_n} \sum_{j=1}^{m_n} X_j, \tag{13}$$

where m_n is the size of the pooled sample, and $X_1, X_2, \ldots$ are iid random variables. This situation approximately describes for instance a standard linear regression coefficient μ estimated from a large sample of size m_n.

Since the researcher accumulates data at each step, $m_n > m_q$ for all $n > q$. Hence, using (13) for $n \geq q$, we obtain

$$\text{cov}(\hat{\mu}_q, \hat{\mu}_n) = \frac{1}{m_q m_n} \sum_{r=1}^{m_q} \sum_{j=1}^{m_n} \text{cov}(X_r, X_j) = \frac{\text{var}(X)}{m_n}. \tag{14}$$

Here we used the assumption that $X_1, X_2, \ldots$ are iid, so that $\text{cov}(X_r, X_j) = 0$ for all $r \neq j$ and $\text{cov}(X_r, X_r) = \text{var}(X)$ for all r.

Under the null hypothesis and in a large sample, we can consistently estimate $\text{var}(X)$ and approximate the nth t-statistic as

$$T_n \approx \frac{\sqrt{m_n}(\hat{\mu}_n - \mu_0)}{\sqrt{\text{var}(X)}} \sim \mathcal{N}(0, 1). \tag{15}$$

Combining (14) and (15) for $q \leq n$, we find

$$\text{cov}(T_q, T_n) \approx \frac{\text{cov}(\sqrt{m_q}\hat{\mu}_q, \sqrt{m_n}\hat{\mu}_n)}{\text{var}(X)} = \sqrt{\frac{m_q}{m_n}} > 0.$$

Therefore, the assumptions of proposition 5 are satisfied when the researcher p-hacks by pooling data.

5.2. Examining various regression specifications

For this example, we assume that the researcher uses ordinary least squares in the standard linear regression model to estimate an effect of interest. A typical effect of interest would be the population value of a regression coefficient. The researcher uses different regression specifications at each p-hacking step, so the effect of interest is no longer fixed.

Specifically, at step n the researcher uses ordinary least squares to estimate a regression coefficient in a regression of W_n on X_n from a set of m iid data points $(W_{n1}, \ldots, W_{nm})$ and $(X_{n1}, \ldots, X_{nm})$ so

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m} X_{nj} W_{nj}}{\sum_{j=1}^{m} X_{nj}^2}.$$

Here, X_n represents the regressor of interest after it has been projected off of the space spanned by the covariates included in the nth regression model, following the procedure described in the Frisch-Waugh-Lovell theorem.

At each step n, the dependent variable W_n, the regressor of interest, and the covariates may be specified differently in the regression model. The researcher implicitly sets the object of interest at step n to the population regression coefficient

$$\mu_n = \frac{\mathbb{E}(X_n W_n)}{\mathbb{E}(X_n^2)}.$$

Standard assumptions on the underlying data-generating process yield the following large-sample distributional approximation:

$$\hat{\mu}_n \sim \mathcal{N}(\mu_n, \text{var}(\hat{\mu}_n)) \quad \text{where} \quad \text{var}(\hat{\mu}_n) = \frac{\text{var}(X_n W_n)}{m\,\mathbb{E}(X_n^2)^2}.$$

In large samples, a law of large numbers for the sample means of X_q^2 and X_n^2 provides

(16)
$$\text{cov}(\hat{\mu}_q, \hat{\mu}_n) = \text{cov}\left(\frac{\left(\sum_{j=1}^m X_{qj} W_{qj}\right)/m}{\left(\sum_{j=1}^m X_{qj}^2\right)/m}, \frac{\left(\sum_{j=1}^m X_{nj} W_{nj}\right)/m}{\left(\sum_{j=1}^m X_{nj}^2\right)/m} \right) \approx \frac{\text{cov}(X_q W_q, X_n W_n)}{m\,\mathbb{E}(X_q^2)\,\mathbb{E}(X_n^2)}.$$

With different specifications at each step, the researcher implicitly tests a different null hypothesis $H_{0,n} : \mu_n = \mu_{0,n}$ corresponding to each specification, leading to the following t-statistic at the nth p-hacking step:

$$T_n = \frac{\hat{\mu}_n - \mu_{0,n}}{\text{se}(\hat{\mu}_n)}.$$

If the researcher can consistently estimate $\text{var}(X_n W_n)$ and $\mathbb{E}(X_n^2)$, under H_0 we can approximate the nth t-statistic in large samples as

(17)
$$T_n = \frac{\sqrt{m}(\hat{\mu}_n - \mu_{0,n})}{\sqrt{\text{var}(X_n W_n)}/\,\mathbb{E}(X_n^2)} \sim \mathcal{N}(0, 1).$$

Combining (16) and (17), we find that in large samples

$$\text{cov}(T_q, T_n) \approx \frac{\text{cov}(\sqrt{m}\hat{\mu}_q, \sqrt{m}\hat{\mu}_n)}{\sqrt{\text{var}(X_q W_q)\,\text{var}(X_n W_n)}/[\mathbb{E}(X_q^2)\,\mathbb{E}(X_n^2)]} \approx \frac{\text{cov}(X_q W_q, X_n W_n)}{\sqrt{\text{var}(X_q W_q)\,\text{var}(X_n W_n)}}.$$

Thus, $\text{cov}(T_q, T_n) \geq 0$ if and only if $\text{cov}(X_q W_q, X_n W_n) \geq 0$.

To conclude, when the researcher p-hacks by examining various regression specifications, the assumptions of proposition 5 are approximately satisfied in large samples

as long as any two specifications $q < n$ satisfy

$$\text{cov}(X_q W_q, X_n W_n) \geq 0.$$

This condition is testable. It presumably holds for most regression specifications aimed at estimating similar population regression coefficients μ_q and μ_n.

5.3. Examining various instruments

By modifying some of the definitions in the previous example, we can also cover the case for which the researcher uses two-stage least squares in a standard linear regression model to estimate the effect of interest. Assuming that the instruments are both strong and valid, we can simply modify the definition of X_n to equal the regressor of interest after all regressors have been projected onto the space spanned by the instruments used at the nth p-hacking step, and then the resulting regressor of interest has been projected off of the space spanned by the covariates included in the nth regression model. If the researcher uses the same dependent variable and second stage covariates at each step and only changes the set of instruments used, and if the regression model is correctly specified, the null hypotheses are identical at each step since each μ_n simply equals the true second stage regression coefficient.

6. Calibration of the completion probability

We now calibrate the completion probability γ, which determines all the numerical results in the model. We use the result that with probability $1 - \gamma$, the first research step cannot be completed before the deadline, so the researcher does not obtain any result. We can therefore measure the probability $1 - \gamma$ from data on the lifecycle of studies: it is the share of studies that were started and never reported results.

6.1. Social sciences

Franco, Malhotra, and Simonovits (2014, table 2) report that among 249 social-science studies funded by the National Science Foundation, 20 went missing: they did not report any results and were never written up. An additional 3 studies were written up but without any results. So 23/249 = 9.2% of the studies did not obtain any results. Based on this number, we would calibrate the completion probability to $\gamma = 1 - 9.2\% = 90.8\%$.

In addition to the 23 studies that did not obtain any results, 45 studies were never written up. Some researchers may have never written up their study for strategic reasons: because they did not find a significant result, and it takes some time to write up results. But this seems unlikely. First, some studies that were not written up had isolated significant or strongly significant results. Second, a study that is left unwritten does not contribute to the authors' CVs. They are also invisible to other researchers, so they do not contribute to the literature. As such, it seems likely that in most cases, researchers were forced to abandon their project by external forces. If we include these studies, we find that (23 + 45)/249 = 27.3% of the studies were not completed. Based on this number, we would calibrate the completion probability to $\gamma = 1 - 27.3\% = 72.7\%$.

6.2. Medical sciences

Dwan et al. (2008) review 16 metastudies that each follow a cohort of medical studies. The studies are followed from protocol approval to publication, so we can determine the fraction of approved studies that were abandoned. The 16 metastudies consider a total of 6903 approved studies. We focus on the 4563 studies for which the outcome is known—either by surveying the scientists who conducted the studies or by searching the literature. Of these studies, 658 were never started, or 658/4563 = 14.4%. Based on this number, we would calibrate the completion probability to $\gamma = 1 - 14.4\% = 85.6\%$.

In addition, not all the studies that started were completed. Of the 3905 studies that started, 228 were still ongoing when the cohort studies were written and 388 were stopped early. Hence, of the 3905 – 228 = 3677 studies that started and stopped, 388/3677 = 10.6% stopped early. Some researchers may have stopped their study early for strategic reasons: because they found an early significant result and tried to publish it. But this is unlikely in practice because the vast majority of the studies that stopped early did not include any analysis. If we add the studies that stopped early to those that never started, we find that $14.4\% + (85.6\% \times 10.6\%) = 23.5\%$ of the approved studies were not completed. This number yields a completion probability of $\gamma = 1 - 23.5\% = 76.5\%$.

6.3. Summary

In the social-science studies followed by Franco, Malhotra, and Simonovits (2014), the completion probability ranges between 72.7% to 90.8%—depending on how we count unwritten studies—with a midpoint of 81.8%. Given that these studies were fully funded, and that their data were collected by a high-quality survey firm, it is likely that they

faced better odds than standard studies. Accordingly, it is likely that the completion probability for regular social-science studies is at the lower end of the range.

In the medical-science studies followed by Dwan et al. (2008), the completion probability ranges between 76.5% and 85.6%—depending on how we count studies that stopped early—with a midpoint of 81.1%. The range of plausible completion probabilities in medical science is within the range in social science, and the midpoints of the two ranges are very close.

Hence, we simply calibrate the completion probability to $\gamma = 80\%$. We use the range 72.7%–90.8% to assess the sensitivity of the findings to the value of γ.

7. Obtaining the ICCV from a simple Bonferroni correction

We now propose a simple way to compute ICCVs as a Bonferroni correction. We apply the correction using the completion probability of $\gamma = 80\%$ derived in section 6.

7.1. 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. This gives a very simple Bonferroni correction to deal with p-hacking. From (11), we see that the classical significance level required to deal with p-hacking is just $1 - \gamma$ times the desired significance level: $\alpha^* \approx (1 - \gamma)\alpha$.

Since $\gamma \approx 80\%$, the classical significance level required to deal with p-hacking is just one fifth of the desired significance level: $\alpha^* = \alpha/5$. For instance, achieving a 5% significance level under p-hacking requires to use the critical value yielding a 1% significance level under classical conditions. This rule of thumb works for any hypothesis test.

7.2. Comparison with another proposal to address p-hacking

To deal with p-hacking, Benjamin et al. (2018) argue that researchers should replace the standard significance level of 5% by a lower significance level of 0.5%. According to (11), this tenfold reduction in significance level is appropriate for $(1 - \gamma)/(1 - \alpha\gamma) = 1/10$ or $\gamma = 90\%$, which is at at the top end of plausible completion probabilities. If the true completion probability is lower than 90%, then a significance level of 0.5% is too stringent: it would result in a type 1 error rate below 5%.

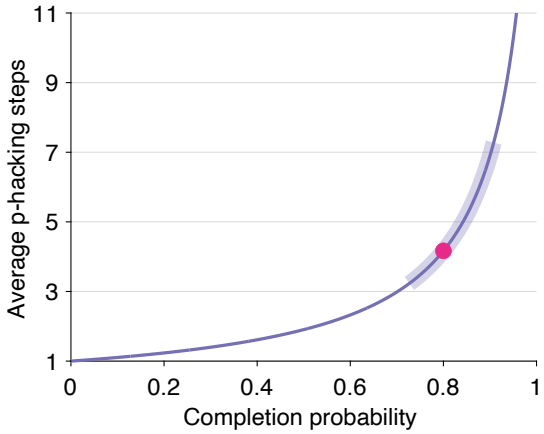


FIGURE 1. Amount of p-hacking at 5% significance level

The thick line indicates plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink point indicates our preferred calibration of the completion probability: $\gamma = 80\%$. The curve is obtained from (4) with $\alpha = 5\%$.

8. Sensitivity of the numerical results

We now describe the sensitivity of the ICCV to the completion probability γ. To do that, we fix the significance level to 5%, and we sometimes assume that the test statistic has a standard normal distribution under the null hypothesis.

8.1. Prevailing p-hacking

The amount of p-hacking under classical critical values is given by (4). For the median completion probability of 80% and a significance level of 5%, the expected number of p-hacking steps is 4.2 (figure 1). Moreover, the amount of p-hacking is increasing with the completion probability. When the completion probability increases from 72.7% to 90.8%, the average number of p-hacking steps grows from 3.2 to 7.3.

8.2. 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 2). So 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—because then there is more p-hacking. When the completion probability increases from 72.7% to 90.8%, the probability of type 1 error increases from

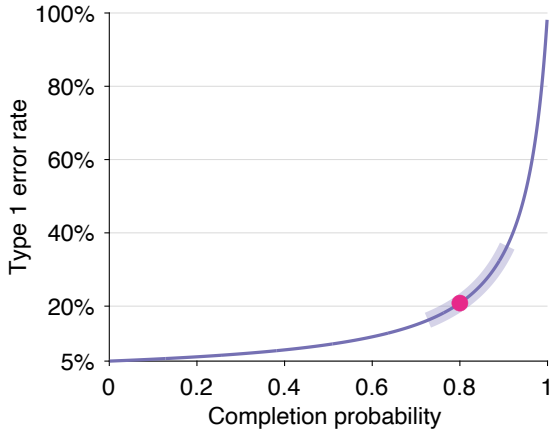


FIGURE 2. Type 1 error rate at 5% significance level

The thick line indicates plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink point indicates our preferred calibration of the completion probability: $\gamma = 80\%$. The curve is obtained from (7) with $\alpha = 5\%$.

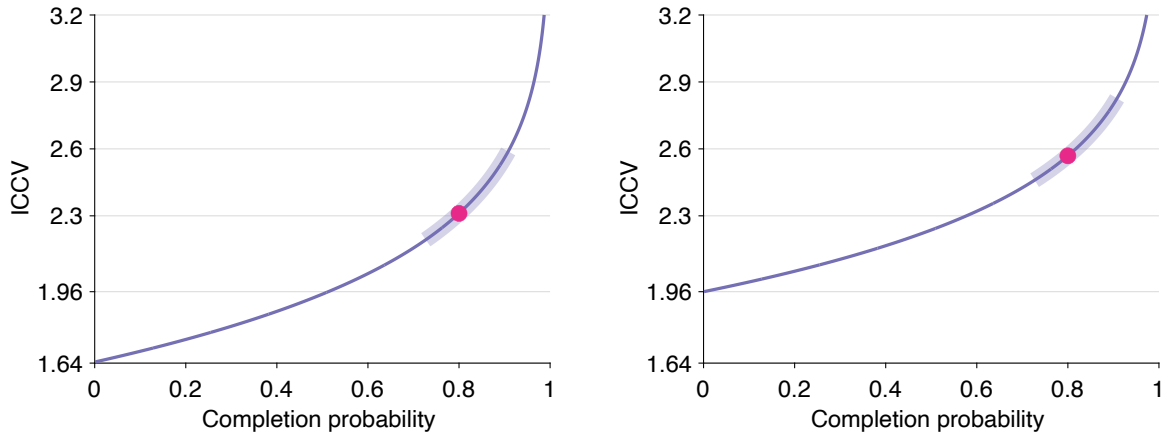
16% to 36%.

8.3. ICCV

We calculate the ICCV when the significance level is 5% and the underlying test statistic has a standard normal distribution, as in the common z-test, or in a t-test conducted from a large sample. We begin by calculating the ICCV for a one-sided test. The ICCV 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 of $\gamma = 80\%$, the ICCV is 2.31 (figure 3A). It is larger than the corresponding classical critical value of 1.65 but not by a tremendous amount.

Next we calculate the ICCV for a two-sided test. The ICCV is now given by (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard half-normal distribution: $Z(x) = \Phi^{-1}(1 - x/2)$. We use the standard half-normal distribution here because conducting a two-sided test when the test statistic follows a standard normal distribution is equivalent to conducting a one-sided test with the absolute value of the test statistic—which follows a standard half-normal distribution. For the completion probability of 80%, the ICCV is 2.57 (figure 3B). It is larger than the corresponding classical critical value of 1.96 but not by much.

For both one-sided and two-sided tests, the ICCV is increasing with the completion



A. One-sided z-test

B. Two-sided z-test

FIGURE 3. ICCV for z-test at 5% significance level

The thick lines indicate plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink points indicate our preferred calibration of the completion probability: $\gamma = 80\%$. A: The curve is obtained from (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard normal distribution. B: The curve is obtained from (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard half-normal distribution.

probability. But despite the broad range of p-hacking induced by plausible completion probabilities—from 3 to 7 p-hacking steps on average (figure 1)—the range of ICCVs is fairly compact. For one-sided tests, the ICCV remains between 2.19 and 2.59 (figure 3A); for two-sided tests, it remains between 2.46 and 2.83 (figure 3B). This is reassuring: ICCVs are not very different in fields with different completion probabilities and p-hacking intensity. It also means that the ICCV does not depend much on the exact value of the completion probability in a given field.

8.4. P-hacking under ICCV

Under the ICCV, p-hacking is more prevalent than under classical critical values. The average number of p-hacking steps under the ICCV is given by (10). For a completion probability of 80% and significance level of 5%, the average number of p-hacking steps increases from 4.2 under classical critical value to 4.8 under the ICCV (figure 4).

8.5. Classical significance level under the ICCV

Because the ICCV is higher than classical critical values, the classical significance level associated with the ICCV is below the nominal significance level—as shown by (11). For

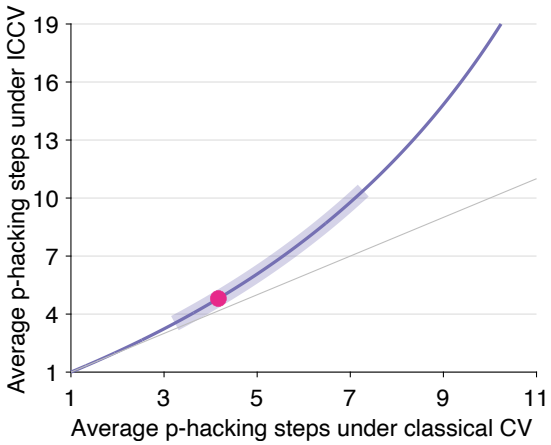


FIGURE 4. P-hacking under ICCV

The thick line indicates plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink point indicates our preferred calibration of the completion probability: $\gamma = 80\%$. The curve is obtained from (10) with $\alpha = 5\%$.

a completion probability of 80% and nominal significance level of 5%, for instance, the classical significance level is only 1.0% (figure 5).

The classical significance level happens to be almost linear in the completion probability, decreasing from 5% when the completion probability is 0 (no p-hacking) to 0% when the completion probability is 1 (reaching a foregone conclusion). This property makes it easy to assess the sensitivity of the classical significance level and hence ICCV to the underlying completion probability. For plausible completion probabilities, the classical significance level corresponding to the ICCV is between 1.4% and 0.5%.

9. Conclusion

To conclude, we summarize our results. We also compare our approach to p-hacking with the registration of pre-analysis plans (PAPs).

9.1. Summary

To correct the excessive rate of type 1 error caused by p-hacking in hypothesis tests, we construct critical values that are compatible with the incentives faced by scientists. These ICCVs deliver the promised rate of type 1 error. Once an ICCV is in place, researchers may conduct several studies to be able to publish their work; nevertheless, readers can be confident that true null hypotheses are not rejected more often than the

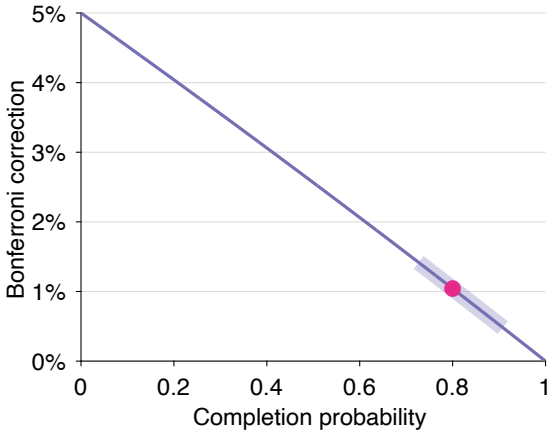


FIGURE 5. Bonferroni correction producing ICCV at 5% significance level

The thick line indicates plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink point indicates our preferred calibration of the completion probability: $\gamma = 80\%$. The curve is obtained from (11) with $\alpha = 5\%$.

advertised significance level.

ICCVs are larger than classical critical values. The ICCV for any test and any significance level is the classical critical value for the same test with roughly one fifth of the significance level. For instance, for a z-test with a significance level of 5%, the ICCV is 2.31 instead of 1.65 if the test is one-sided and 2.57 instead of 1.96 if the test is two-sided.

The ICCVs apply if researchers p-hack by collecting independent data samples. The ICCVs also apply if researchers p-hack by pooling the data samples, by examining various regression specifications, or by examining various instruments. Hence, the ICCVs apply against many of the p-hacking practices documented in the literature (appendix A).

9.2. Comparison of the ICCV and PAP approaches

The most popular solution to the p-hacking problem in hypothesis tests is to ask researchers to register a PAP (Miguel et al. 2014; Christensen and Miguel 2018; Nosek et al. 2018; Christensen, Freese, and Miguel 2019). Yet PAPs are not without limitations (Olken 2015; Coffman and Niederle 2015; Banerjee et al. 2020; Abrams, Libgober, and List 2021). First, PAPs prevent scientists from engaging in exploratory analysis, although exploration is an important source of scientific discovery. Second, PAPs are sometimes impractical, either because it is hard to formulate hypotheses before seeing the data, or because the scientist is already familiar with the data. ICCVs therefore seem more appropriate than PAPs when scientific exploration plays a key role, and with observational

data.

While PAPs prevent many forms of p-hacking (Adda, Decker, and Ottaviani 2020), they cannot prevent the data sampling that we consider in our model. Furthermore, PAPs cannot prevent researchers from strategically selecting a subset of the data that they collected in order to achieve statistically significant results (Lang and Qiu 2021). Hence, ICCVs could be used in conjunction with PAPs to eliminate the excessive type 1 error rates caused by strategic data selection.

References

Abrams, Eliot, Jonathan Libgober, and John A. List. 2021. "Research Registries and the Credibility Crisis: An Empirical and Theoretical Investigation." https://perma.cc/NJG8-55QV.

Adda, Jerome, Christian Decker, and Marco Ottaviani. 2020. "P-Hacking in Clinical Trials and How Incentives Shape the Distribution of Results Across Phases." *Proceedings of the National Academy of Sciences* 117 (24): 13386–13392.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–2794.

Anscombe, Francis J. 1954. "Fixed-Sample-Size Analysis of Sequential Observations." *Biometrics* 10 (1): 89–100.

Armitage, Peter. 1967. "Some Developments in the Theory and Practice of Sequential Medical Trials." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by Lucien M. Le Cam and Jerzy Neyman, vol. 4: 791–804. Berkeley, CA: University of California Press.

Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–554.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics." NBER Working Paper 26993.

Begg, Colin B., and Jesse A. Berlin. 1988. "Publication Bias: a Problem in Interpreting Medical Data." *Journal of the Royal Statistical Society (Series A)* 151 (3): 419–445.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Bjorn Brembs, Lawrence Brown, Colin Camerer et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6–10.

Biagioli, Mario, and Alexandra Lippman. 2020. "Metrics and the New Ecologies of Academic Misconduct." In *Gaming the Metrics: Misconduct and Manipulation in Academic Research*, edited by Mario Biagioli and Alexandra Lippman, 1–23. Cambridge, MA: MIT Press.

Bozarth, Jerold D., and Ralph R. Roberts. 1972. "Signifying Significant Significance." *American Psychologist* 27 (8): 774–775.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–3660.

Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: the Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Christensen, Garret. 2018. "Manual of Best Practices in Transparent Social Science Research." https://github.com/garretchristensen/BestPracticesManual/blob/65b77b1991e9b6d5360d3fc6aa2bb7528bedf7ff/Manual.pdf.

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

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–980.

Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (3): 81–98.

Cole, LaMont C. 1957. "Biological Clock in the Unicorn." *Science* 125 (3253): 874–876.

Csada, Ryan D., Paul C. James, and Richard H. M. Espie. 1996. "The 'File Drawer Problem' of Non-Significant Results: Does It Apply to Biological Research?" *Oikos* 76 (3): 591–593.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, edited by T. Paul Schultz and John A. Strauss, vol. 4, 3895–3962. Amsterdam: Elsevier.

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 et al. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLoS ONE* 3 (8): e3081.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wuthrich. 2021. "Detecting p-Hacking." *Econometrica*. https://www.econometricsociety.org/system/files/18583-3.pdf.

Fanelli, Daniele, Rodrigo Costas, and John P.A. Ioannidis. 2017. "Meta-Assessment of Bias in Science." *Proceedings of the National Academy of Sciences* 114 (14): 3714–3719.

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

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–1505.

Gibson, John, David L. Anderson, and John Tressler. 2014. "Which Journal Rankings Best Explain Academic Salaries? Evidence from the University of California." *Economic Inquiry* 52 (4): 1322–1340.

Glaeser, Edward L. 2008. "Researcher Incentives and Empirical Methods." In *The Foundations of Positive and Normative Economics: A Hand Book*, edited by Andrew Caplin and Andrew Schotter, chap. 13. New York: Oxford University Press.

Hagstrom, Warren. 1965. *The Scientific Community*. New York: Basic Books.

Hansen, W. Lee, Burton A. Weisbrod, and Robert P. Strauss. 1978. "Modeling the Earnings and Research Productivity of Academic Economists." *Journal of Political Economy* 86 (4): 729–741.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLoS Biology* 13 (3): e1002106.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research

Registration." *Political Analysis* 21 (1): 1–20.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3): 944–960.

Hutton, J. L., and Paula R. Williamson. 2000. "Bias in meta-analysis due to outcome variable selection within studies." *Applied Statistics* 49 (3): 359–370.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124.

Ioannidis, John P.A., and Thomas A. Trikalinos. 2007. "An Exploratory Test for an Excess of Significant Findings." *Clinical Trials* 4 (3): 245–253.

Jennions, Michael D., and Anders P. Moeller. 2002. "Publication Bias in Ecology and Evolution: An Empirical Assessment Using the 'Trim and Fill' Method." *Biological Reviews* 77 (2): 211–222.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–532.

Katz, David A. 1973. "Faculty Salaries, Promotions, and Productivity at a Large University." *American Economic Review* 63 (3): 469–477.

Lang, Megan, and Wenfeng Qiu. 2021. "Cherry Picking." https://doi.org/10.31222/osf.io/as9zd.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43.

Lindsay, D. Stephen. 2015. "Replication in Psychological Science." *Psychological Science* 26 (12): 1827–1832.

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

Merton, Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* 22 (6): 635–659.

Merton, Robert K. 1973. *The Sociology of Science*. Chicago: University of Chicago Press.

Miguel, Edward, Colin Camerer, Katherine Casey, Joshua Cohen, Kevin M. Esterling, Alan Gerber, Rachel Glennerster, Don P. Green, Macartan Humphreys, Guido Imbens et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspectives on Psychological Science* 7 (6): 615–631.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.

Sauer, Raymond D. 1988. "Estimates of the Returns to Quality and Coauthorship in Economic Academia." *Journal of Political Economy* 96 (4): 855–866.

Siegfried, John J., and Kenneth J. White. 1973. "Financial Rewards to Research and Teaching: A Case Study of Academic Economists." *American Economic Review* 63 (2): 309–315.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undis-

closed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–1366.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143 (2): 534.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. "Specification Curve Analysis." *Nature Human Behaviour* 4 (11): 1208–1214.

Skeels, Jack W., and Robert P. Fairbanks. 1968. "Publish or Perish: An Analysis of the Mobility of Publishing and Nonpublishing Economists." *Southern Economic Journal* 35 (1): 17–25.

Song, F., A. J. Eastwood, S. Gilbody, L. Duley, and A. J. Sutton. 2000. "Publication and Related Biases: A Review." *Health Technology Assessment* 4 (10): 1–115.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing transparency through a multiverse analysis." *Perspectives on Psychological Science* 11 (5): 702–712.

Sterling, Theodore D. 1959. "Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance–Or Vice Versa." *Journal of the American Statistical Association* 54 (285): 30–34.

Swidler, Steve, and Elizabeth Goldreyer. 1998. "The Value of a Finance Journal Publication." *Journal of Finance* 53 (1): 351–363.

Tuckman, Howard P., and Jack Leahey. 1975. "What Is an Article Worth?" *Journal of Political Economy* 83 (5): 951–967.

Vivalt, Eva. 2019. "Specification Searching and Significance Inflation Across Time, Methods and Disciplines." *Oxford Bulletin of Economics and Statistics* 81 (4): 797–816.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The ASA's Statement on P-Values: Context, Process, and Purpose." *American Statistician* 70 (2): 129–133.

Young, Cristobal, and Katherine Holsteen. 2017. "Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis." *Sociological Methods & Research* 46 (1): 3–40.

Appendix A. Evidence of p-hacking

Here we review evidence of p-hacking across scientific fields. We also discuss the two main reasons for p-hacking. The first is that statistically significant results are more rewarded, which gives researchers an incentive to p-hack. The second is that researchers have a lot of flexibility in their empirical work, which allows them to p-hack easily.

A.1. Observed p-hacking

P-hacking is prevalent in many sciences. First, a survey of 5964 psychologists at major US universities shows that p-hacking is common: 63% of respondents admit to failing to report all outcomes, 56% admit to deciding whether to collect more data after examining whether the results were significant, 46% admit to selectively reporting studies that "worked", 38% admit to deciding whether to exclude data after looking at the impact of doing so on the results, 28% admit to failing to report all treatments in a study, and 16% admit to stopping collecting data earlier than planned after obtaining the desired results (John, Loewenstein, and Prelec 2012, table 1).

P-hacking is also directly visible in the lifecycle of studies. Franco, Malhotra, and Simonovits (2014, table 3) track 221 experimental studies in the social sciences, from experimental design to publication. They find that 64.6% of the studies reporting insignificant results were never written up, whereas only 4.4% of the studies reporting strongly significant results were not written up. Clearly, scientists report results selectively: significant results are almost certain to be reported, whereas insignificant results are likely to remain unreported. In a large-scale analysis of the lifecycle of clinical trials, (Dwan et al. 2008) also find that statistically significant outcomes are more likely to be reported than insignificant outcomes.

Finally, the effects of p-hacking appear in meta-analyses. The distributions of test statistics or p-values across many studies in a literature show that researchers tinker with their econometric specifications in order to obtain statistically significant results (Hutton and Williamson 2000; Head et al. 2015; Brodeur et al. 2016; Vivalt 2019; Brodeur, Cook, and Heyes 2020; Elliott, Kudrin, and Wuthrich 2021).

A.2. Rewards from significant results

Publication bias. The main reason why statistically significant results are valuable is that scientific journals prefer publishing significant results. Such publication bias

was first identified in psychology journals (Sterling 1959; Bozarth and Roberts 1972). It has since been observed across the social sciences (Christensen, Freese, and Miguel 2019, chapter 3), medical sciences (Begg and Berlin 1988; Song et al. 2000; Ioannidis and Trikalinos 2007; Dwan et al. 2008), biological sciences (Csada, James, and Espie 1996; Jennions and Moeller 2002), and many other disciplines (Fanelli, Costas, and Ioannidis 2017).

Andrews and Kasy (2019, p. 2767) assess the magnitude of the bias in two literatures: experimental economics and psychology. They find that results significant at the 5% level are 30 times more likely to be published than insignificant results.

Rewards from publication. Publications, in turn, determine a scientist's career path, including promotion (Skeels and Fairbanks 1968) and salary (Katz 1973; Siegfried and White 1973; Tuckman and Leahey 1975; Hansen, Weisbrod, and Strauss 1978; Sauer 1988; Swidler and Goldreyer 1998; Gibson, Anderson, and Tressler 2014). In some countries, scientists are also rewarded with cash bonuses as high as $30,000 for publication in top journals (Biagioli and Lippman 2020, p. 6). Publications yield not only material rewards but also honorific rewards (Hagstrom 1965). One such reward is eponymy, "the practice of affixing the name of the scientist to all or part of what he has found" (Merton 1957). Beyond eponymy are prizes, medals, memberships in academies of sciences, and fellowships in learned societies (Merton 1957).

Computing the rewards from significant results. Since a study that presents statistically significant results is more likely to be published than one that presents insignificant results, and a published study yields higher rewards than an unpublished study, researchers have an incentive to obtain significant results by p-hacking.

Formally, let V be the random variable giving the rewards from a completed study. There are several sources of randomness. The study may not be published, or it may be published in one of many possible journals, from the most prestigious to the most obscure. And even when published in a journal of a given standing, the study's impact may greatly vary.

The expected rewards from a study with statistically significant results are $v^s = \mathbb{E}(V \mid \text{significant})$, and those from a study with statistically insignificant results are $v^i = \mathbb{E}(V \mid \text{insignificant})$.

Using the law of iterated expectations, we find

$$\begin{aligned} v^s = {} & \mathbb{E}(V \mid \text{published \& significant}) \times \mathbb{P}(\text{published} \mid \text{significant}) \\ & + \mathbb{E}(V \mid \text{unpublished \& significant}) \times \mathbb{P}(\text{unpublished} \mid \text{significant}). \end{aligned}$$

We note that $\mathbb{P}(\text{unpublished} \mid \text{significant}) + \mathbb{P}(\text{published} \mid \text{significant}) = 1$. And we assume that conditional on the publication status, the rewards are independent from statistical significance. Then we obtain

$$\begin{aligned} v^s = {} & \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished}) \right] \times \mathbb{P}(\text{published} \mid \text{significant}) \\ & + \mathbb{E}(V \mid \text{unpublished}). \end{aligned}$$

Following the same logic, we compute the expected rewards from a study with insignificant results. We find

$$\begin{aligned} v^i = {} & \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished}) \right] \times \mathbb{P}(\text{published} \mid \text{insignificant}) \\ & + \mathbb{E}(V \mid \text{unpublished}). \end{aligned}$$

Accordingly, the expected gain from reaching a significant result is

$$\begin{aligned} v^s - v^i = {} & \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished}) \right] \\ & \times \left[\mathbb{P}(\text{published} \mid \text{significant}) - \mathbb{P}(\text{published} \mid \text{insignificant}) \right]. \end{aligned}$$

Empirically, statistically significant results are more likely to be published than insignificant results, so $\mathbb{P}(\text{published} \mid \text{significant}) > \mathbb{P}(\text{published} \mid \text{insignificant})$. Moreover, a published study is expected to yield higher rewards than an unpublished study, so $\mathbb{E}(V \mid \text{published}) > \mathbb{E}(V \mid \text{unpublished})$. Based on this evidence, it is beneficial to obtain a significant result: $v^s > v^i$.

A.3. Opportunities for p-hacking

Researchers have a lot of flexibility in data collection and analysis (Huntington-Klein et al. 2021). Hence, they have opportunities to obtain statistically significant results, even when the null hypothesis is true. Indeed, researchers have found that it is easy to obtain statistically significant results when the null hypothesis is true, without violating prevailing scientific norms in biology (Cole 1957), medical science (Armitage 1967,

section 4), economics (Leamer 1983; Lovell 1983), psychology (Simmons, Nelson, and Simonsohn 2011), and political science (Humphreys, de la Sierra, and van der Windt 2013).

Appendix B. Reframing the deadline in terms of money or stamina

In the model of p-hacking presented in section 2, the deadline can be reframed in terms of money or stamina instead of time without changing any of the results.

B.1. Money

Assume that it takes a random amount of money to sample data and conduct hypothesis tests. The cumulative amounts of money required to complete the different p-hacking steps are D_1, D_2, ... In other words, the monetary costs of each step are D_1, $D_2 - D_1$, $D_3 - D_2$, ... Furthermore, assume that the researcher's budget is L, so the researcher must complete the research project before the cumulative costs reach L. The size of the budget could be random.

If a p-hacking step is completed before the research budget is depleted, the researcher may report the results or decide to start a new step. If the researcher runs out of money in the middle of a p-hacking step, she must stop and is not able to complete the analysis. Such a model works exactly like the model in the main text and the ICCV is the same.

B.2. Stamina

The model with stamina works exactly like the model with money once some of the variables are reinterpreted. First, D_1, D_2, D_3, ...indicate the cumulative amount of stamina required to complete the different p-hacking steps. Second, the researcher cannot tolerate to expand more than a quantity L of stamina on a given project. Accordingly, the researcher must complete the research project before the cumulative amount of stamina expanded reaches L. Such a model works like the model in the main text and the ICCV is the same.

Appendix C. Adding a monetary or psychic cost of research

We extend the p-hacking model of section 2 by introducing a cost of research, incurred at each new p-hacking step. The cost could be monetary or psychic. Unlike in appendix B, the cost enters the researcher's payoffs. We find that the ICCV is not modified by this extension. Hence, our model of p-hacking is quite robust.

C.1. Assumptions

We introduce an expected cost of doing research, c. The cost could be monetary or psychic; it is incurred at each p-hacking step. Because we focus on fields in which research occurs, we assume that c is low enough relative to the rewards from research, v^i and v^s, such that it is optimal for researchers to engage in research.

C.2. Optimal stopping time and ICCV

Significant result. Since it is optimal to engage in research, the researcher starts a first p-hacking step. With probability γ, the step can be completed before the deadline, and the researcher obtains a test result. If the result is significant, the researcher obtains v^s, so she stops immediately. Indeed, she cannot obtain a higher payoff by continuing. The same is true at any future step: any time a researcher obtains a significant result, she immediately stops, since it is impossible to obtain a higher payoff in the future.

High research cost. What does the researcher decide if the result is insignificant? It depends on the research cost c. If the cost is high enough, the researcher stops right away. This happens when the possibility of obtaining a significant result in the future does not compensate the research cost. In that case, there is no p-hacking: the researcher conducts one step of the study and stops, irrespective of the result. The ICCV is then just the classical critical value.

Low research cost. As discussed in appendix A, however, p-hacking is prevalent in reality. So the most realistic scenario is that the research cost is low enough so the researcher starts a new step upon obtaining a first insignificant result. In that case, because the researcher faces exactly the same situation after each step, the researcher continues to p-hack until she obtains a significant result.

Summary. If the research cost is low enough that p-hacking occurs, the presence of the research cost does not modify the researcher's behavior. It is optimal for the researcher to p-hack until she reaches a significant result. Accordingly, everything remains the same in the model—including the ICCV.

C.3. Computing the cost boundaries

Given that the properties of the model remain the same with the research cost, we can use previous results to compute the expected payoffs from doing research and then the cost below which it is optimal to p-hack, and the cost below which it is optimal to engage in research. The expectations of the payoffs depends on the distribution of the test statistic, which in turn depends on which hypothesis is true. Here we assume that the researcher is conservative and therefore computes the payoff expectations under the null hypothesis.

Continuation value of research. We first compute the continuation value of research for a researcher who has already recorded an insignificant result. We denote this value V^i. Because the researcher's situation is invariant in time, this continuation value is the same at each p-hacking step.

When a researcher decides to continue p-hacking, three scenarios are possible. With probability $1-\gamma$, the researcher cannot complete the p-hacking step and must submit an insignificant result. She then collects v^i. With probability γ, she can complete the p-hacking step. Then with probability $S(z^*)$, her result is significant and she collects v^s. With probability $1-S(z^*)$, her result is insignificant once again and the continuation value at this point is V^i. In any case, she must incur a cost c to conduct the research step.

Aggregating these scenarios, we obtain the following continuation value:

$$V^i = (1-\gamma)v^i + \gamma S(z^*)v^s + \gamma[1-S(z^*)]V^i - c.$$

Hence the continuation value is

$$V^i = \frac{(1-\gamma)v^i + \gamma S(z^*)v^s - c}{1-\gamma[1-S(z^*)]}. \tag{A1}$$

Condition for p-hacking. From the continuation value (A1), we compute the cost below which it is optimal to p-hack. When a researcher has obtained one insignificant result,

it is optimal to continue p-hacking if $V^i > v^i$. After a few steps of algebra, this condition becomes

$$c < \gamma S(z^*)(v^s - v^i).$$

Hence, it is optimal to p-hack if the cost of each p-hacking step is below the threshold

$$c^p = \gamma S(z^*)(v^s - v^i).$$

Of course, the cost threshold is higher when significant results are more rewarded relative to insignificant results.

Condition for research. From the continuation value (A1), we also compute the cost below which it is optimal to engage in research. Given that we have normalized the outside option of the researcher to 0, it is optimal to engage in research if the expected value from it is positive.

When a researcher decides to start research, three scenarios are again possible. With probability $1-\gamma$, the researcher cannot complete the first research step and cannot submit any result; she then collects 0. With probability γ, she can complete the first research step. Then with probability $S(z^*)$, her result are significant and she collects v^s. With probability $1 - S(z^*)$, her result are insignificant and the continuation value at this point is V^i. In any case, she must incur a cost c to conduct the research step.

Aggregating these scenarios, we obtain the initial continuation value:

$$V^r = (1-\gamma) \times 0 + \gamma S(z^*)v^s + \gamma[1 - S(z^*)]V^i - c.$$

We rewrite the initial continuation value as

$$V^r = \gamma V^i + \gamma S(z^*)(v^s - V^i) - c.$$

Using the value of V^i given by (A1), we finally obtain

$$V^r = \frac{\gamma S(z^*)}{1-\gamma[1-S(z^*)]}v^s + \frac{(1-\gamma)\gamma[1-S(z^*)]}{1-\gamma[1-S(z^*)]}v^i - \frac{1}{1-\gamma[1-S(z^*)]} \cdot c. \tag{A2}$$

It is optimal to start a research project if $V^r > y_0 = 0$. From (A2), this condition becomes

$$c < \gamma S(z^*)v^s + (1-\gamma)\gamma[1-S(z^*)]v^i.$$

Hence, it is optimal to start research if the cost of each research step is below the threshold

$$c^r = \gamma S(z^*)v^s + (1-\gamma)\gamma[1-S(z^*)]v^i.$$

The cost threshold is higher when scientific results are more rewarded.

The threshold to engage in research is higher than the threshold to engage in p-hacking:

$$c^r = c^p + \gamma[1-\gamma(1-S(z^*))] > c^p.$$

Hence, for all cost between c^p and c^r, researchers engage in research but do not p-hack.

Appendix D. Adding time discounting

We introduce time discounting into the p-hacking model of section 2. When the researcher discounts the future, a result submitted early is more valuable than the same result submitted later. Yet, the researcher's behavior and ICCV are not modified.

D.1. Assumptions

We introduce a discount factor, $\delta \in (0,1)$. The discount factor cost is incurred at each new p-hacking step, so the value of a research result obtained at step n is discounted by δ^n. Because the returns to research are positive without discounting, they also are positive with discounting, so it is optimal for researchers to engage in research.

D.2. Optimal stopping time

As in appendix C, we find that the optimal stopping time is the same as in the basic model.

Significant result. Any time a researcher obtains a significant result, she immediately stops, since it is impossible to obtain a higher payoff in the future.

High discounting. What does the researcher decide if the result is insignificant? It depends on the value of the discount factor δ. If discounting is high enough, the researcher is better off stopping right away. This happens when the possibility of obtaining a significant result in the future does not compensate the time discounting. In that

case, there is no p-hacking: the researcher conducts one step of the study and stops, irrespective of the result. The ICCV is then just the classical critical value.

Low discounting. As discussed in appendix A, p-hacking is prevalent in reality. So the most realistic scenario is that discounting is low enough so the researcher starts a new step upon obtaining a first insignificant result. In that case, because the researcher faces exactly the same situation after each step, the researcher continues to p-hack until she obtains a significant result.

Summary. If time discounting is low enough that p-hacking occurs, the presence of discounting does not modify the researcher's behavior. It is optimal for the researcher to p-hack until she reaches a significant result. Accordingly, everything remains the same in the model—including the ICCV.

D.3. Computing discounting boundary

Given that all the properties of the model remain the same with discounting, we can use previous results to compute the discount factor below which it is optimal to p-hack.

The key step is computing the continuation value of research for a researcher who has already recorded an insignificant result. We denote this value V^i. Because the researcher's situation is invariant in time, this continuation value is the same at each new p-hacking step.

When a researcher decides to continue p-hacking, three scenarios are possible. With probability $1-\gamma$, the researcher cannot complete the new p-hacking step and must submit an insignificant result; she then collects δv^i. With probability γ, she can complete the new p-hacking step. Then with probability $S(z^*)$, her result are significant and she collects δv^s; with probability $1-S(z^*)$, her result are insignificant once again and the continuation value at this point is δV^i.

Aggregating these scenarios, we obtain the following continuation value:

$$V^i = (1-\gamma)\delta v^i + \gamma S(z^*)\delta v^s + \gamma[1-S(z^*)]\delta V^i.$$

Hence the continuation value is

$$V^i = \delta \frac{(1-\gamma)v^i + \gamma S(z^*)v^s}{1-\delta\gamma[1-S(z^*)]}.$$

Then, when a researcher has obtained one insignificant result, it is optimal to continue p-hacking if $V^i > v^i$. After a few steps of algebra, this condition becomes

$$\delta > \frac{v^i}{v^i + \gamma S(z^*)(v^s - v^i)}.$$

Hence, it is optimal to p-hack if the discount factor is above the threshold

$$\delta^p = \frac{v^i}{v^i + \gamma S(z^*)(v^s - v^i)}.$$

Of course, the discounting threshold is lower when significant results are more rewarded relative to insignificant results. If insignificant results are not rewarded at all, then researchers p-hack irrespective of discounting.

Appendix E. Proofs

We provide proofs that are omitted in the main text.

E.1. Proof of proposition 2

We start by computing the probability that the reported test statistic $R(z)$ exceeds a critical value z under the null hypothesis. From the law of total probability:

$$\mathbb{P}(R(z) > z) = \sum_{j \geq 1} \mathbb{P}(R(z) > z \mid N(z) = j)\, \mathbb{P}(N(z) = j), \tag{A3}$$

where, according to Bayes' rule,

$$\mathbb{P}(R(z) > z \mid N(z) = j) = \frac{\mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j - 1)}{\mathbb{P}(N(z) = j \mid N(z) > j - 1)}. \tag{A4}$$

Then we compute the conditional probability given by (A4). The fact that $N(z) > j - 1$ means that the deadline has not occurred during the first $j - 1$ steps, and that the $j - 1$ test statistics collected during these steps have not been significant. Conditional on $N(z) > j - 1$, three events may happen.

First, with probability $1 - \gamma$, the deadline occurs during step j. If $j > 1$, then $N(z) = j$ and the researcher reports an insignificant result: $R(z) \leq z$. If $j = 1$, the researcher does not report any result.

Second, with probability γ, the deadline does not occur during step j. This creates two subcases. With probability $\gamma S(z)$, the test statistic T_j obtained during step j is significant. Then $N(z) = j$ and $R(z) = T_j > z$. With probability $\gamma[1 - S(z)]$, the test statistic T_j obtained during step j is insignificant. In that case, $N(z) > j$.

From this case-by-case description, we conclude that the probability that the researcher stops at step j given that she has already completed $j - 1$ steps is

$$\mathbb{P}(N(z) = j \mid N(z) > j - 1) = 1 - \gamma + \gamma S(z). \tag{A5}$$

And the probability that the researcher reports a significant result at step j given that she has already completed $j - 1$ steps is

$$\mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j - 1) = \mathbb{P}\Big(T_j > z, K > j \mid N(z) > j - 1\Big) = \gamma S(z). \tag{A6}$$

Combining (A4), (A5), and (A6), we find that the probability to report a significant result given that the researcher stops the research project at step j is

$$\mathbb{P}(R(z) > z \mid N(z) = j) = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}. \tag{A7}$$

The probability (A7) is independent of j, which greatly simplifies (A3):

$$\mathbb{P}(R(z) > z) = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)} \left[\sum_{j \geq 1} \mathbb{P}(N(z) = j)\right] = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}.$$

Finally, we compute the probability to report a significant result given that any result is reported. This conditional probability is given by

$$\mathbb{P}(R(z) > z \mid L > D_1) = \frac{\mathbb{P}(R(z) > z)}{\mathbb{P}(L > D_1)} = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)} \cdot \frac{1}{\gamma}.$$

To compute the ratio, we used that with probability γ, the deadline does not occur before the end of the first step, $L > D_1$, so some results will be reported, either significant or insignificant.

Therefore, when the critical value is set to z, the probability of type 1 error in a reported study is

$$S^*(z) = \mathbb{P}(R(z) > z \mid L > D_1) = \frac{S(z)}{1 - \gamma + \gamma S(z)}.$$

E.2. Proof of proposition 3

First, to compute the ICCV, we rewrite the definition (8) as

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

The inverse of the survival function S is the function Z. Inverting S here, we obtain the explicit expression for the ICCV:

$$z^* = Z\left(\alpha \cdot \frac{1-\gamma}{1-\alpha\gamma}\right). \tag{A8}$$

Equation (A8) indicates that the ICCV always exists. Since $\alpha \in (0,1)$ and $\gamma \in (0,1)$, the ratio $(1-\gamma)/(1-\alpha\gamma)$ is in $(0,1)$. Hence, the argument of the inverse survival function Z in (9) satisfies

$$0 < \alpha \cdot \frac{1-\gamma}{1-\alpha\gamma} < \alpha.$$

Accordingly, the argument is in $(0,1)$. As the domain of the inverse survival function is $(0,1)$, the ICCV exists.

From (A8), we can compare the ICCV to a classical critical value. A classical critical value is defined by $z = Z(\alpha)$, while the ICCV is defined by (A8). Since the inverse survival function is strictly decreasing, and since the argument of the inverse survival function in (9) is strictly less than α, we infer that the ICCV is strictly larger than the classical critical value: $z^* > z$.

Unsurprisingly, the ICCV is strictly decreasing in the significance level α. Indeed, the argument of the inverse survival function in (A8) is strictly increasing in the significance level $\alpha \in (0,1)$. Since the inverse survival function itself is strictly decreasing, we infer that the ICCV is strictly decreasing in the significance level.

E.3. Proof of proposition 4

The proof follows along the same lines as the proof of proposition 2 with some adjustments. In particular, note that (A3) and (A4) continue to hold and the probability that the deadline occurs at any step k continues to equal $1-\gamma$. However conditional on $N(z) > j-1$, the probability the test statistic obtained during step k is significant is now bounded above by $\gamma S(z)$ since

$$\mathbb{P}(T_n > z \mid N(z) > j-1) = \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) \leq S(z). \tag{A9}$$

Therefore, (A5) no longer holds but can be replaced by

(A10) $$\mathbb{P}(N(z) = j \mid N(z) > j-1) = 1 - \gamma + \gamma\,\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z).$$

Similarly, (A6) no longer holds but can be replaced by

(A11) $$\mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j-1) = \gamma\,\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z).$$

Since the function $x \mapsto x/(1-\gamma+x)$ is increasing in $x > 0$ for all $\gamma < 1$, (A9), (A10), (A11) and (A4) imply

$$\mathbb{P}(R(z) > z \mid N(z) = j) \leq \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}$$

so that (A3) implies

$$\mathbb{P}(R(z) > z) \leq \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}.$$

Applying (8), we obtain the statement of the proposition:

$$\mathbb{P}\big(R(z^*) > z \mid L > D_1\big) = \frac{\mathbb{P}(R(z^*) > z^*)}{\mathbb{P}(L > D_1)} \leq \frac{\gamma S(z^*)}{1 - \gamma + \gamma S(z^*)} \cdot \frac{1}{\gamma} = \frac{S(z^*)}{1 - \gamma + \gamma S(z^*)} = \alpha.$$

E.4. Proof of proposition 5

We show (12) holds by showing the conditional probability on the left-hand side is less than the unconditional probability on the right-hand side after further conditioning on any realized value of an additional statistic.

Note that the normally-distributed random vector

$$A(n) = [T_1, \ldots, T_{n-1}] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n$$

is independent of T_n since

$$\begin{aligned}
\operatorname{cov}(A(n), T_n) &= \operatorname{cov}([T_1, \ldots, T_{n-1}] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n, T_n) \\
&= \operatorname{cov}([T_1, \ldots, T_{n-1}], T_n) - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]\operatorname{var}(T_n, T_n) \\
&= [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)] = 0.
\end{aligned}$$

Using the vector $A(n)$, we describe the conditioning event in (12) as follows:

$$\{T_1, \ldots, T_{n-1} \leq z\} = \big\{[\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n \leq z - A(n)\big\}$$

$$= \left\{ T_n \leq \min_{1 \leq j \leq n-1 : \Omega_{j,n}(n) > 0} \frac{z - A_j(n)}{\Omega_{j,n}(n)}, \max_{1 \leq j \leq n-1 : \Omega_{j,n}(n) = 0} A_j(n) \leq z \right\}.$$

Since $A(n)$ and T_n are independent, the conditional distribution of the nth t-statistic given the conditioning event in (12) and the realized value of $A(n)$ is a standard normal truncated from above:

$$T_n \mid \left\{ T_1, \ldots, T_{n-1} \leq z, A(n) = a \right\} \sim \xi \mid \xi \leq \mathcal{U}(a),$$

where $\xi \sim \mathcal{N}(0, 1)$ and

$$\mathcal{U}(a) = \min_{1 \leq j \leq n-1 : \Omega_{j,n}(n) > 0} \frac{z - a_j}{\Omega_{j,n}(n)}.$$

Using the properties of the truncated normal distribution, we characterize the conditional probability of type 1 error for the nth t-statistic given non-rejection by the previous t-statistics in the sequence and the realized value of $A(n)$ as

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) = \begin{cases} 1 - \frac{\Phi(z)}{\Phi(\mathcal{U}(a))} & \text{if } z \leq \mathcal{U}(a), \\ 0 & \text{if } z > \mathcal{U}(a) \end{cases}$$

for all a, where Φ denotes the cumulative distribution function of a standard normal random variable. Therefore for any values of a and z,

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) \leq 1 - \Phi(z).$$

But for $F_A(\cdot)$ equal to the cumulative distribution function of $A(n)$,

$$\begin{aligned} \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) &= \int_{\mathbb{R}^{n-1}} \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) dF_A(a) \\ &\leq 1 - \Phi(z) = \mathbb{P}(T_n > z) \end{aligned}$$

and we obtain the statement of the proposition.