Critical Values Robust to P-hacking: Online Appendices

Adam McCloskey, Pascal Michaillat

April 2024

Published in *Review of Economics and Statistics* • Available at https://doi.org/10.1162/rest_a_01456

Appendix A. Proofs

This appendix provides proofs that are omitted in the main text.

A.1. Proof of proposition 2

Defining the probability of type 1 error. We aim to compute the probability of type 1 error $S^*(z)$ when the critical value is set to z. This is the probability that the reported test statistic $R(z)$ exceeds z under the null hypothesis, given that any result is reported:

$$S^*(z) = \mathbb{P}(R(z) > z \mid L > D_1),$$

where $\mathbb{P}$ denotes the probability measure under H_0. Because the scientist can only report a significant result if the first experiment is completed, $\mathbb{P}(R(z) > z, L > D_1) = \mathbb{P}(R(z) > z)$, so

$$S^*(z) = \frac{\mathbb{P}(R(z) > z)}{\mathbb{P}(L > D_1)}.$$

By definition, $\mathbb{P}(L > D_1) = \gamma$. Accordingly, the probability of type 1 error is

$$S^*(z) = \frac{\mathbb{P}(R(z) > z)}{\gamma}. \tag{A1}$$

Total probability of reporting a significant result. To apply formula (A1), we need to compute $\mathbb{P}(R(z) > z)$. To do that, we use the law of total probability:

$$\mathbb{P}(R(z) > z) = \sum_{j=1}^{\infty} \mathbb{P}(R(z) > z, N(z) = j). \tag{A2}$$

Because the scientist can only stop at experiment j if she has already completed $j-1$ experiments, $\mathbb{P}(R(z) > z, N(z) = j) = \mathbb{P}(R(z) > z, N(z) = j, N(z) > j-1)$, so

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

Using this result, we rewrite equation (A2) as

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

Probability of reporting a significant result at experiment j. To apply formula (A3), we must compute $\mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j-1)$. The fact that $N(z) > j-1$ means that the project resources have not been exhausted during the first $j-1$ experiments, but that the $j-1$ test statistics collected have not been significant. Then the event that $R(z) > z$

and $N(z) = j$ is realized if experiment j can be completed, which occurs with probability γ, and if the test statistic obtained from experiment j is significant, which occurs with probability $S(z)$. We therefore find that

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

Computing the probability of type 1 error. The probability (A4) is independent of j, which greatly simplifies (A3):

$$\mathbb{P}(R(z) > z) = \gamma S(z) \cdot \sum_{j=1}^{\infty} \mathbb{P}(N(z) > j - 1) = \gamma S(z) \cdot \sum_{j=1}^{\infty} \mathbb{P}(N(z) \geq j). \tag{A5}$$

Since the optimal stopping time $N(z)$ is a nonnegative, integer-valued random variable, we know from Ross (2014, p. 292) that

$$\sum_{j=1}^{\infty} \mathbb{P}(N(z) \geq j) = \mathbb{E}(N(z)).$$

Moreover, proposition 1 establishes that the expected value of the optimal stopping time is

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

Accordingly, the probability of reporting a significant result is

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

Combining this equation with (A1), we find that the probability of type 1 error is

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

A.2. Proof of proposition 3

Expression of the robust critical value. We begin by rewriting the implicit definition of the robust critical value, given by (8). Since the cumulative distribution function F and survival function S are related by $F = 1 - S$, we rewrite (8) as

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

This equation allows us to express $S(z^*)$ as a function of the parameters α and γ:

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

Inverting S, we obtain an explicit expression for the robust critical value:

$$z^* = Z\left(\alpha \cdot \frac{1-\gamma}{1-\alpha\gamma}\right), \tag{A6}$$

where the function Z is the inverse of the survival function S.

Existence of the robust critical value. 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 (A6), $\alpha(1-\gamma)/(1-\alpha\gamma)$, is in $(0,\alpha) \subset (0,1)$. As the domain of the inverse survival function is $(0,1)$, the robust critical value exists for any $\alpha \in (0,1)$ and $\gamma \in (0,1)$.

Comparing robust and classical critical values. The classical critical value is given by $Z(\alpha)$, while the robust critical value is defined by (A6). Since the argument of the inverse survival function Z in (A6) is strictly less than α, and since the inverse survival function is strictly decreasing, the robust critical value is strictly larger than the classical critical value: $z^* > Z(\alpha)$.

Relation between robust critical value and significance level. The argument of the inverse survival function Z in (A6) is strictly increasing in the significance level $\alpha \in (0,1)$. Since the inverse survival function is strictly decreasing, the robust critical value is strictly decreasing in the significance level.

Appendix B. Prevalence of p-hacking, and reasons for it

This appendix develops the argument made in the introduction that p-hacking is prevalent in science. It also discusses the reasons behind p-hacking. The first is that p-hacking is rewarded because statistically significant results have greater payoffs than insignificant ones. The second is that p-hacking is not very costly because scientists have a lot of flexibility in their empirical work.

B.1. Prevalence of p-hacking

P-hacking is prevalent in many sciences.

Survey of scientists. A survey of 5964 psychologists at major US universities conducted by John, Loewenstein, and Prelec (2012, table 1) 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 data collection earlier than planned after obtaining the desired results.

Meta-analyses of published studies. The effects of p-hacking also appear in meta-analyses of published studies (Hutton and Williamson 2000; Head et al. 2015; Brodeur et al. 2016; Vivalt 2019; Brodeur, Cook, and Heyes 2020; Elliott, Kudrin, and Wuthrich 2022). The distributions of test statistics or p-values across studies in a literature show that scientists tinker with their econometric specifications in order to obtain significant results.

Lifecycle of scientific studies. Franco, Malhotra, and Simonovits (2014, table 3) track a cohort of 221 experimental studies in the social sciences, from experimental design to publication, and find evidence of p-hacking. Indeed, 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. Thus, scientists report results selectively: insignificant results are likely to remain unreported, whereas significant results are almost certain to be reported.

B.2. Rewards from significant results

Scientists hunt for significant results because such results are more rewarded than insignificant results. The reason is twofold. First, a study presenting significant results is

more likely to be published than one presenting insignificant results. Second, a published study yields higher rewards than an unpublished study.

Publication bias. Indeed, scientific journals prefer publishing significant results. Such publication bias was first identified in psychology journals (Sterling 1959; Bozarth and Roberts 1972; Ferguson and Brannick 2012). It has since been observed across the social sciences (Card and Krueger 1995; Ashenfelter, Harmon, and Oosterbeek 1999; Gerber, Green, and Nickerson 2001; Ashenfelter and Greenstone 2004; Rose and Stanley 2005; Christensen, Freese, and Miguel 2019), medical sciences (Simes 1986; Dickersin et al. 1987; 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 (Smaldino and McElreath 2016). Publications lead of course to promotions (Skeels and Fairbanks 1968), but also to a higher 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 rewarded with cash bonuses as high as $30,000 for publications 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).

Rewards from significant results. Accordingly, scientists have the incentive to obtain significant results by p-hacking. Formally, let V be the random variable giving the rewards from a completed study. Randomness comes from several sources: the study may not be published at all; or it may be published in one of many possible journals, from the most prestigious to the most obscure; even when it is published in a journal of a given standing, the study's impact may vary. The expected rewards from a study with a significant result are

$$v^s = \mathbb{E}(V \mid \text{significant}),$$

and those from a study with an insignificant result are

$$v^i = \mathbb{E}(V \mid \text{insignificant}).$$

We assume that conditional on publication status, rewards are independent from statistical significance. Then, using the law of total expectation, we find

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

Since $\mathbb{P}(\text{unpublished} \mid \text{significant}) + \mathbb{P}(\text{published} \mid \text{significant}) = 1$, we obtain

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

Following the same logic, we find

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

Accordingly, by obtaining a significant result, a scientist expects to gain

$$\begin{aligned} v^s - v^i = {} & [\,\mathbb{P}(\text{published} \mid \text{significant}) - \mathbb{P}(\text{published} \mid \text{insignificant})] \\ & \times [\,\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished})]\,. \end{aligned} \tag{A7}$$

Empirically, significant results are more likely to be published than insignificant ones:

$$\mathbb{P}(\text{published} \mid \text{significant}) > \mathbb{P}(\text{published} \mid \text{insignificant}).$$

Moreover, a published study yields higher rewards than an unpublished one:

$$\mathbb{E}(V \mid \text{published}) > \mathbb{E}(V \mid \text{unpublished}).$$

These facts together with (A7) imply that a scientist benefits more from a significant result than from an insignificant one:

$$v^s > v^i.$$

B.3. Opportunities for p-hacking

Scientists have a lot of flexibility in data collection and analysis (Huntington-Klein et al. 2021). Such flexibility affords them opportunities to obtain significant results, even when

the null hypothesis is true. Indeed scientists have found that it is easy to obtain significant results when the null hypothesis is true, without violating 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 C. Other p-hacking strategies

In the model of section II, scientists p-hack by repeatedly running experiments until they reach significant results. In this appendix, we adapt the model to describe a wider range of p-hacking strategies. We consider scientists who pool data across experiments, successively remove outliers, successively examine different regression specifications, and successively use different instruments. We find that the robust critical value (9) remains useful under these other p-hacking strategies because it maintains the probability of type 1 error below the significance level. More generally, because the robust critical value (9) is derived with independent test statistics, it controls the probability of type 1 error for any p-hacking strategy that induces positive dependence across test statistics. As such, the robust critical value (9) acts as a least-favorable robust critical value over a range of p-hacking strategies.

C.1. General p-hacking strategy

P-hacking process. We begin by considering a general p-hacking process that produces positively dependent test statistics.

ASSUMPTION A1. *The sequence of test statistics $T_1, T_2, T_3, \ldots$ is positively dependent:*

$$\mathbb{P}\Big(T_j > z \mid T_1, \ldots, T_{j-1} \leq z\Big) \leq \mathbb{P}\Big(T_j > z\Big) = S(z) \tag{A8}$$

for all $j \geq 2$ and all $z \geq 0$.

Type 1 error rate with positively dependent test statistics. We show that the robust critical value (9) maintains the probability of type 1 error below the significance level even when test statistics are positively dependent.

PROPOSITION A1. *Under assumption A1, the probability of type 1 error under the robust critical value* (9) *does not exceed the significance level.*

PROOF. The proof proceeds as the proof of proposition 2, with some adjustments. First, we compute (A2) slightly differently:

$$\begin{aligned}
\mathbb{P}\Big(R(z^*) > z^*\Big) &= \sum_{j=1}^{\infty} \mathbb{P}\Big(R(z^*) > z, N(z^*) = j\Big) \\
&= \sum_{j=1}^{\infty} \frac{\mathbb{P}(R(z^*) > z, N(z^*) = j, N(z^*) > j-1)}{\mathbb{P}(N(z^*) = j, N(z^*) > j-1)} \cdot \mathbb{P}\Big(N(z^*) = j, N(z^*) > j-1\Big) \\
&= \sum_{j=1}^{\infty} \frac{\mathbb{P}(R(z^*) > z, N(z^*) = j \mid N(z^*) > j-1)}{\mathbb{P}(N(z^*) = j \mid N(z^*) > j-1)} \cdot \mathbb{P}\Big(N(z^*) = j\Big). \qquad \text{(A9)}
\end{aligned}$$

The term $\mathbb{P}(R(z^*) > z, N(z^*) = j \mid N(z^*) > j-1)$ in (A9) gives the probability that the jth experiment can be completed and the jth test statistic is significant, given that the previous $j-1$ experiments could be completed and the previous $j-1$ test statistics were insignificant. Therefore,

$$\text{(A10)} \qquad \mathbb{P}\Big(R(z^*) > z^*, N(z^*) = j \mid N(z^*) > j-1\Big) = \gamma\, \mathbb{P}\Big(T_j > z^* \mid T_1, \ldots, T_{j-1} \le z^*\Big).$$

The term $\mathbb{P}(N(z^*) = j \mid N(z^*) > j-1)$ in (A9) gives the probability that the scientist stops at the jth experiment, given that the previous $j-1$ experiments could be completed and the previous $j-1$ test statistics were insignificant. This event occurs either if the jth experiment can be completed and the jth test statistic is significant, or if the jth experiment cannot be completed. Therefore,

$$\text{(A11)} \qquad \mathbb{P}\Big(N(z^*) = j \mid N(z^*) > j-1\Big) = 1 - \gamma + \gamma\, \mathbb{P}\Big(T_j > z^* \mid T_1, \ldots, T_{j-1} \le z^*\Big).$$

Combining (A10) and (A11), we obtain

$$\frac{\mathbb{P}(R(z^*) > z^*, N(z^*) = j \mid N(z^*) > j-1)}{\mathbb{P}(N(z^*) = j \mid N(z^*) > j-1)} = \frac{\gamma\, \mathbb{P}\Big(T_j > z^* \mid T_1, \ldots, T_{j-1} \le z^*\Big)}{1 - \gamma + \gamma\, \mathbb{P}\Big(T_j > z^* \mid T_1, \ldots, T_{j-1} \le z^*\Big)}.$$

The function $x \mapsto x/(1-\gamma+x)$ is increasing in $x > 0$ for any $\gamma < 1$, and assumption A1 says that $\mathbb{P}\Big(T_j > z^* \mid T_1, \ldots, T_{j-1} \le z^*\Big) \le S(z^*)$. Thus, we have

$$\text{(A12)} \qquad \frac{\mathbb{P}(R(z^*) > z^*, N(z^*) = j \mid N(z^*) > j-1)}{\mathbb{P}(N(z^*) = j \mid N(z^*) > j-1)} \le \frac{\gamma S(z^*)}{1 - \gamma + \gamma S(z^*)} = \frac{\gamma S(z^*)}{1 - \gamma F(z^*)}.$$

From (A9) and (A12), and given the fact that $\sum_{j=1}^{\infty} \mathbb{P}(N(z^*) = j) = 1$, we infer that

$$\mathbb{P}\Big(R(z^*) > z^*\Big) \le \frac{\gamma S(z^*)}{1 - \gamma F(z^*)} \cdot \sum_{j=1}^{\infty} \mathbb{P}\Big(N(z^*) = j\Big) = \frac{\gamma S(z^*)}{1 - \gamma F(z^*)}.$$

Then using equation (A1), we obtain an upper bound on the probability of type 1 error:

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

But the critical value z^* satisfies (8), so the right-hand side of the inequality is just the significance level α. We conclude that the probability of type 1 error is below the significance level: $S^*(z^*) \le \alpha$. □

Condition ensuring positive dependence of t-statistics. In the common case of sequential z-tests or large-sample t-tests, a simple condition on the covariances between successive

test statistics guarantees that proposition A1 applies:

PROPOSITION A2. *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 $\Omega_{1,1}(n), \ldots, \Omega_{n,n}(n)$ equal 1 and all covariances $\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)$ are nonnegative. Then assumption A1 is satisfied so proposition A1 applies.*

PROOF. We show assumption A1 holds by showing the conditional probability on the left-hand side of (A8) is less than the unconditional probability on the right-hand side of (A8) 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

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

Using the vector $A(n)$, we describe the conditioning event in (A8) as follows:

$$\begin{aligned}
\{T_1, \ldots, T_{n-1} \le z\} &= \big\{[\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n \le z - A(n)\big\} \\
&= \left\{T_n \le \min_{1 \le j \le n-1 : \Omega_{j,n}(n) > 0} \frac{z - A_j(n)}{\Omega_{j,n}(n)}, \max_{1 \le j \le n-1 : \Omega_{j,n}(n) = 0} A_j(n) \le z\right\}.
\end{aligned}$$

Since $A(n)$ and T_n are independent, the conditional distribution of the nth statistic given the conditioning event in (A8) and the realized value of $A(n)$ is a standard normal truncated from above:

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

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

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

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

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

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 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. □

The intuition for the proofs is simple. The optimal p-hacking strategy described by lemma 1 remains identical when the test statistics are dependent. Indeed, the derivation of the optimal stopping time does not rely on the independence of the test statistics, so it remains valid here. The stochastic properties of the optimal stopping time and reported test statistic do change, however. But under assumption A1, we can guarantee that the robust critical value given by (9) keeps the probability of type 1 error below the significance level.

P-hacking strategies generating positively dependent t-statistics. The distributional assumption in proposition A2 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 several common forms of p-hacking. Suppose that the scientist constructs a general estimator of the form

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m_n} X_{nj} W_{nj}}{\sum_{j=1}^{m_n} X_{nj}^2} \tag{A13}$$

at step n, where m_n is equal to the sample size used in step n. In the subsections that follow, we show that several common estimators in applied work take the form of (A13). Under standard moment conditions on two sets of m_n approximately iid data points $(X_{n1}, \ldots, X_{nm_n})$ and $(W_{n1}, \ldots, W_{nm_n})$, a bivariate central limit theorem implies the following distributional approximation for large m_n:

$$\frac{1}{\sqrt{m_n}} \begin{pmatrix} \sum_{j=1}^{m_n} [X_{nj} W_{nj} - \mathbb{E}(X_n W_n)] \\ \sum_{j=1}^{m_n} [X_{nj}^2 - \mathbb{E}(X_n^2)] \end{pmatrix} \sim \mathcal{N}(0, \Sigma_n)$$

with

$$\Sigma_n = \begin{pmatrix} \mathbb{E}(X_n^2 W_n^2) - \mathbb{E}(X_n W_n)^2 & \mathbb{E}(X_n^3 W_n) - \mathbb{E}(X_n^2)\,\mathbb{E}(X_n W_n) \\ \mathbb{E}(X_n^3 W_n) - \mathbb{E}(X_n^2)\,\mathbb{E}(X_n W_n) & \mathbb{E}(X_n^4) - \mathbb{E}(X_n^2)^2 \end{pmatrix}.$$

In turn, the delta method implies that for large m_n,

$$\sqrt{m_n}(\hat{\mu}_n - \mu_n) \sim \mathcal{N}(0, \sigma_n^2) \tag{A14}$$

with

$$\mu_n = \frac{\mathbb{E}(X_n W_n)}{\mathbb{E}\left(X_n^2\right)}$$
$$\sigma_n^2 = \frac{\mathbb{E}\left(X_n^2 W_n^2\right)\mathbb{E}\left(X_n^2\right)^3 - 2\,\mathbb{E}\left(X_n^3 W_n\right)\mathbb{E}\left(X_n^2\right)\mathbb{E}(X_n W_n) + \mathbb{E}\left(X_n^4\right)\mathbb{E}(X_n W_n)^2}{\mathbb{E}\left(X_n^2\right)^4}.$$

By using an estimator of the form (A13), (A14) shows that the scientist is implicitly testing the null hypothesis $H_{0,n} : \mu_n = \mu_{0,n}$ at step n, where the estimand μ_n and its hypothesized value $\mu_{0,n}$ may differ across experiments n, depending upon the context. Under standard moment conditions, the scientist can consistently estimate the large-sample variances σ_n^2, by some estimator $\hat{\sigma}_n^2$. This enables the formation of t-statistics with standard normal distributions under $H_{0,n}$ in large samples:

$$T_n = \frac{\sqrt{m_n}(\hat{\mu}_n - \mu_{0,n})}{\hat{\sigma}_n} \sim \mathcal{N}(0, 1).$$

As $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$ are consistent for σ_i^2 and σ_j^2,

$$\operatorname{cov}(T_i, T_j) \approx \frac{\sqrt{m_i m_j}\operatorname{cov}\left(\hat{\mu}_i, \hat{\mu}_j\right)}{\sigma_i \sigma_j} \geq 0$$

if and only if $\operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$. Thus, for estimators of the form (A13), the conditions of proposition A2 hold in large samples when the standard normal approximation for each T_i holds jointly with the others and $\operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ for each $i, j = 1, \ldots, n$. Sections C.2, C.3, C.4, and C.5 provide common examples of estimators for which these conditions typically hold.

C.2. Pooling data

P-hacking process. The scientist studies a mean parameter $\mu = \mathbb{E}(W)$ for some random variable W. The null hypothesis is $H_0 : \mu = \mu_0$. The alternative hypothesis is $\mu > \mu_0$. After each experiment the scientist adds the newly collected data to the existing dataset. The

new data are independent and collected from the same underlying population. After experiment n the scientist constructs an estimate $\hat{\mu}_n$ of the parameter by taking a mean from the pooled dataset:

$$\hat{\mu}_n = \frac{1}{m_n} \sum_{j=1}^{m_n} W_j, \tag{A15}$$

where m_n is the size of the pooled dataset, and $W_1, \ldots, W_{m_n}$ are iid random variables with mean μ. Using the notation in (A13), we have $X_{nj} = 1$ and $W_{nj} = W_j$ for all n and j.

Verifying the conditions of proposition A2. Since the scientist accumulates data at each step, $m_i > m_j$ for all $i > j$. Hence, using (A15) for $i \geq j$, we obtain

$$\operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j) = \frac{1}{m_i m_j} \sum_{r=1}^{m_j} \sum_{k=1}^{m_i} \operatorname{cov}(W_r, W_k) = \frac{\operatorname{var}(W)}{m_i} \geq 0.$$

Here we used the assumption that $W_1, \ldots, W_{m_n}$ are iid, so $\operatorname{cov}(W_r, W_k) = 0$ for all $r \neq k$ and $\operatorname{cov}(W_r, W_r) = \operatorname{var}(W)$ for all r. Furthermore, any finite set of $\hat{\mu}_i$'s have an approximate joint normal distribution in large samples by a standard multivariate central limit theorem. Therefore, the conditions of proposition A2 are satisfied when the scientist p-hacks by pooling data.

C.3. Removing outliers

P-hacking process. The scientist successively removes outliers from a dataset of size m. At step n, the scientist discards all data points further away than some value c_n from some value χ. She discards more data points at each step so that $c_n < c_q$ for $n > q$. In this scenario, at step n the scientist constructs an estimate $\hat{\mu}_n$ of the parameter by taking a mean from the trimmed sample:

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m} W_j \mathbb{1}(|W_j - \chi| \leq c_n)}{\sum_{j=1}^{m} \mathbb{1}(|W_j - \chi| \leq c_n)}, \tag{A16}$$

where $\mathbb{1}$ denotes the indicator function, and $W_1, \ldots, W_m$ are iid random variables. The scientist is implicitly testing a different null hypothesis $H_{0,n} : \mu_n = \mu_{0,n}$ at each step n in this example, where

$$\mu_n = \frac{\mathbb{E}(W \mathbb{1}(|W - \chi| \leq c_n))}{\mathbb{P}(|W - \chi| \leq c_n)}.$$

Using the notation in (A13), we have $X_{nj} = \mathbb{1}(|W_j - \chi| \leq c_n)$, $W_{nj} = W_j$ and $m_n = m$ for all n and j.

Verifying the conditions of proposition A2. Any finite set of $\sum_{j=1}^{m} W_j \mathbb{1}(|W_j - \chi| \le c_i)$'s and $\sum_{j=1}^{m} \mathbb{1}(|W_j - \chi| \le c_i)$'s have an approximate joint normal distribution in large samples so that the delta method implies the same for any finite set of $\hat{\mu}_i$'s in this example. In addition, the joint normality of the $\hat{\mu}_i$'s and the delta method provide the approximate covariance between any $\hat{\mu}_i$ and $\hat{\mu}_j$ in large samples, as the following proposition shows:

PROPOSITION A3. *For $\hat{\mu}_n$ defined by* (A16) *and a sequence $W_1, W_2, \ldots$ of iid random variables, for any $i \ge j$, $m\,\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_j)$ converges to*

$$\frac{\mathrm{var}(W \mid |W - \chi| \le c_i) + \mathbb{E}(W \mid |W - \chi| \le c_i)\,\mathbb{E}\Big(W \mid |W - \chi| \le c_j\Big)\,\mathbb{P}(|W - \chi| > c_i)\,\mathbb{P}\Big(|W - \chi| > c_j\Big)}{\mathbb{P}\Big(|W - \chi| \le c_j\Big)}$$

as $m \to \infty$.

PROOF. A multivariate central limit theorem and the delta method imply

$$\begin{aligned} m\,\mathrm{cov}\Big(\hat{\mu}_i, \hat{\mu}_j\Big) \to{}& \frac{\mathrm{cov}\Big(W\,\mathbb{1}(|W - \chi| \le c_i), W\,\mathbb{1}(|W - \chi| \le c_j)\Big)}{\mathbb{P}(|W - \chi| \le c_i)\,\mathbb{P}\Big(|W - \chi| \le c_j\Big)} \\ &- \frac{\mathbb{E}\Big(W\,\mathbb{1}(|W - \chi| \le c_j)\Big)\,\mathrm{cov}\Big(W\,\mathbb{1}(|W - \chi| \le c_i), \mathbb{1}(|W - \chi| \le c_j)\Big)}{\mathbb{P}(|W - \chi| \le c_i)\,\mathbb{P}\Big(|W - \chi| \le c_j\Big)^2} \\ &- \frac{\mathbb{E}(W\,\mathbb{1}(|W - \chi| \le c_i))\,\mathrm{cov}\Big(W\,\mathbb{1}(|W - \chi| \le c_j), \mathbb{1}(|W - \chi| \le c_i)\Big)}{\mathbb{P}\Big(|W - \chi| \le c_j\Big)\,\mathbb{P}(|W - \chi| \le c_i)^2} \\ &+ \frac{\mathbb{E}(W\,\mathbb{1}(|W - \chi| \le c_i))\,\mathbb{E}\Big(W\,\mathbb{1}(|W - \chi| \le c_j)\Big)\,\mathrm{cov}\Big(\mathbb{1}(|W - \chi| \le c_j), \mathbb{1}(|W - \chi| \le c_i)\Big)}{\mathbb{P}\Big(|W - \chi| \le c_j\Big)^2\,\mathbb{P}(|W - \chi| \le c_i)^2} \end{aligned}$$

as $m \to \infty$. Next we use the definition of covariance, the fact that for $f(w) = w$ or $f(w) = w^2$,

$$\mathbb{E}(f(W) \mid |W - \chi| \le c_i) = \frac{\mathbb{E}(f(W)\,\mathbb{1}(|W - \chi| \le c_i))}{\mathbb{P}(|W - \chi| \le c_i)},$$

and the result that since $c_i < c_j$,

$$\mathbb{1}(|W - \chi| \le c_i)\,\mathbb{1}(|W - \chi| \le c_j) = \mathbb{1}(|W - \chi| \le c_i).$$

From these and standard algebra, we obtain the result of the proposition. □

The proposition shows when the conditions of proposition A2 should hold in large samples. For example, the conditions hold if $\mathbb{E}(W \mid |W - \chi| \le c_i)$ and $\mathbb{E}(W \mid |W - \chi| \le c_j)$ have the same sign. This latter condition should hold for reasonable removals of outliers—that is, for reasonable choices of χ and $c_1, c_2, c_3, \ldots$. For example, suppose that outliers are

considered based on deviations from the mean, so $\mathbb{E}(W) = \chi$. Then if W is symmetrically distributed, this condition holds for any choice of c_n since $\mathbb{E}(W \mid |W - \chi| \leq c_n) = \chi$.

C.4. Examining various regression specifications

P-hacking process. The scientist 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 scientist uses different regression specifications at each p-hacking step, so the parameter of interest differs at each step. Specifically, at step n the scientist uses ordinary least squares to estimate a regression coefficient in a regression of W_n on X_n from two sets of m iid data points $(W_{n1}, \ldots, W_{nm})$ and $(X_{n1}, \ldots, X_{nm})$ so

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m} X_{nj} W_{nj}}{\sum_{j=1}^{m} X_{nj}^2}. \tag{A17}$$

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.

Verifying the conditions of proposition A2. The least squares estimator in (A17) takes the structure of (A13) with $m_n = m$ for all n and therefore satisfies (A14) when, for example, W_n and X_n have finite fourth moments. In this context, the conditions of proposition A2 therefore hold if $\operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ for each $i, j = 1, \ldots, n$, a natural condition for a set of similar regressions. For example, consider two different regressions generating the data

$$W_i = \mu_i X_i + u_i$$
$$W_j = \mu_j X_j + u_j$$

that satisfy standard assumptions such that the least squares estimators of μ_i and μ_j, $\hat{\mu}_i$ and $\hat{\mu}_j$, are jointly asymptotically normally distributed as $m \to \infty$ and centered at μ_i and μ_j with a $\sqrt{m}$ rate of convergence. In this case,

$$m \operatorname{cov}\left(\hat{\mu}_i, \hat{\mu}_j\right) \to \frac{\mathbb{E}\left(u_i u_j X_i X_j\right)}{\mathbb{E}\left(X_i^2\right)\mathbb{E}\left(X_j^2\right)},$$

which is nonnegative if and only if $\mathbb{E}(u_i u_j X_i X_j) \geq 0$. This condition naturally holds when the regressors X_i and X_j and regressands W_i and W_j measure similar quantities. In other words, if the scientist estimates similar population regression coefficients at each p-hacking step, the coefficient estimates should be expected to be positively correlated in large samples.

This is easiest to see when $\mathbb{E}(u_i u_j \mid X_j X_j) = \mathbb{E}(u_i u_j)$ (akin to conditional homoskedasticity) since then $\mathbb{E}(u_i u_j X_i X_j) = \text{cov}(u_i, u_j)\,\text{cov}(X_i, X_j)$ if an intercept is included in the regression. In this case, $\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ in large samples if both u_i and u_j and X_i and X_j are positively correlated.

The condition $\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ is also testable from observed data. The delta method allows one to compute the approximate covariances between any two $\hat{\mu}_i, \hat{\mu}_j$ in large samples for any choices of W_i and X_i. Proposition A3 is an example of such an exercise.

C.5. Examining various instruments

By modifying some of the definitions in the previous example, we can also cover the case in which the scientist uses two-stage least squares to estimate the effect of interest. Assuming that the instruments are both strong and valid, we can 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 scientist 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 equals the true second stage regression coefficient.

Appendix D. P-hacking with cost of doing research

This appendix introduces a cost of doing research into the model of section II. This cost is incurred by the scientist at each new experiment. We find that the robust critical value is not modified by this extension.

D.1. Assumptions

The scientist incurs a cost of doing research $c > 0$ at each experiment. The cost could be monetary or psychological. 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 scientists to engage in research.

D.2. Optimal stopping time and robust critical value

Significant result. Since it is optimal to engage in research, the scientist starts a first experiment. With probability γ, the experiment can be completed, and the scientist obtains a test statistic. If the statistic is significant, the scientist obtains v^s, so she stops immediately. Indeed, she cannot obtain a higher payoff by continuing. The same is true in the future too: any time a scientist obtains a significant result, she immediately stops, since it is impossible to obtain a higher payoff later on.

High research cost. What does the scientist decide if the test statistic is insignificant? It depends on the research cost c. If the cost is high enough, the scientist 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, the scientist does not p-hack: she conducts one experiment and stops, irrespective of the result. The robust critical value is then the classical critical value.

Low research cost. Since p-hacking is prevalent in reality, the most realistic scenario is that the research cost c is low enough so that the scientist runs a new experiment upon obtaining an insignificant result. In that case, because the scientist faces the same situation after each experiment, the scientist continues p-hacking 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 scientist's behavior. It is optimal for the scientist to p-hack until she reaches a significant result. Accordingly, everything remains the same in the model—including the robust critical value.

D.3. Computing the cost boundaries

We now compute the expected payoffs from doing research, 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 depend on the distribution of the test statistic, which in turn depends on which hypothesis is true. We assume that the scientist is conservative and computes the payoff expectations under the null hypothesis.

Continuation value of research. We first compute the continuation value of research for a scientist who has already recorded an insignificant result. We denote the value V^i. Because the scientist's situation is invariant in time, the continuation value is the same at each experiment. When a scientist decides to continue p-hacking, three scenarios are possible. With probability $1-\gamma$, the scientist cannot complete the experiment and must submit an insignificant result. She then collects v^i. With probability γ, she can complete the experiment. 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 incurs a cost c to conduct the experiment. Aggregating the 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

$$V^i = \frac{(1-\gamma)v^i + \gamma S(z^*)v^s - c}{1-\gamma[1-S(z^*)]}. \tag{A18}$$

Condition for p-hacking. We now compute the cost below which it is optimal to p-hack. When a scientist has obtained one insignificant result, it is optimal to continue p-hacking if $V^i > v^i$. Using (A18), we rewrite the condition as

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

Hence, it is optimal to p-hack if the cost of each experiment c 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 (A18), we also compute the cost below which it is optimal to engage in research. Given that we have normalized the outside option

of the scientist to 0, it is optimal to engage in research if the expected value from it is positive.

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

Aggregating the 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 (A18), we finally obtain

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

It is optimal to start a research project if $V^r > y_0 = 0$. 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 experiment is below the threshold

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

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 costs between c^p and c^r, scientists engage in research but do not p-hack.

Appendix E. P-hacking with time discounting

This appendix introduces time discounting into the model of section II. When the scientist discounts the future, a result submitted early is more valuable than the same result submitted later. We find that the robust critical value is not modified by this extension.

E.1. Assumptions

The scientist discounts the future with a discount factor $\delta \in (0, 1)$. Time discounting occurs at each new experiment, so the value of a result obtained at experiment n is discounted by δ^n. Because all the possible payoffs from research are positive (either 0 or $v^i > 0$ or $v^s > 0$), the expected present discounted value from research is strictly positive, irrespective of the strength of discounting. Accordingly, it is optimal for the scientist to engage in research for any discount factor.

E.2. Optimal stopping time and robust critical value

Significant result. The scientist stops p-hacking whenever she obtains a significant result. This is because it is impossible to obtain a higher payoff in the future, and furthermore future payoffs are discounted.

Low discount factor. What does the scientist decide if the result is insignificant? It depends on the discount factor. If the discount factor is close enough to 0, the scientist is better off stopping right away. This happens when the possibility of obtaining a significant result in the future does not compensate for time discounting. In that case, the scientist does not p-hack: she conducts one experiment and stops, irrespective of the result. The appropriate critical value is then the classical critical value.

High discount factor. P-hacking is prevalent in reality. Thus, the most realistic scenario is that the discount factor is close enough to 1 that the scientist starts a new experiment upon obtaining an insignificant result. Then the scientist continues p-hacking until she obtains a significant result, because she faces the same situation after each experiment.

Summary. If the discount factor is high enough that p-hacking occurs, the presence of discounting does not modify the scientist's behavior. It is optimal for the scientist to p-hack until she reaches a significant result. Accordingly, everything remains the same in the model—including the robust critical value.

E.3. Computing the discount-factor boundary

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

Continuation value of research. The key step is computing the continuation value of research for a scientist who has already recorded an insignificant result. We denote the value V^i. Because the scientist's situation is invariant in time, the continuation value is the same at each new experiment. When a scientist decides to continue p-hacking, three scenarios are possible. With probability $1-\gamma$, the scientist cannot complete the new experiment and must submit an insignificant result; she then collects δv^i. With probability γ, she can complete the new experiment. Then with probability $S(z^*)$, her result is significant and she collects δv^s. With probability $1-S(z^*)$, her result is insignificant once again and the continuation value is δV^i. Aggregating the 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

$$V^i = \delta \frac{(1-\gamma)v^i + \gamma S(z^*)v^s}{1-\delta\gamma[1-S(z^*)]}. \tag{A19}$$

Condition for p-hacking. When a scientist has obtained an insignificant result, it is optimal to p-hack if $V^i > v^i$. Using (A19), we rewrite the condition as

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

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

$$\delta^p = \frac{v^i}{v^i + \gamma S(z^*)(v^s - v^i)} \in (0,1).$$

If insignificant results are not rewarded at all ($v^i = 0$), then scientists p-hack under any discount factor ($\delta^p = 0$). If insignificant results are rewarded ($v^i > 0$), then scientists p-hack under a broader range of discount factors when significant results are more rewarded relative to insignificant results (δ^p is lower when $v^s - v^i$ is higher).

Appendix F. P-hacking with increasingly difficult experiments

This appendix extends the model of section II by assuming that experiments become successively more difficult to conduct, and therefore less likely to be completed. We show that for any decreasing completion probability, the robust critical value (9) maintains the probability of type 1 error below the significance level.

F.1. Assumptions

The experiments become increasingly difficult to run. Therefore, the resources required for each experiment, $D_1, D_2 - D_1, D_3 - D_2, \ldots$, are independent but not identically distributed. Instead, the amount of resources required for each experiment is increasing, so the probability of completing an experiment before resources are exhausted is decreasing. Formally, the probability of completing the first experiment is

$$\gamma_1 = \mathbb{P}(D_1 < L) = \mathbb{E}(\exp(-\lambda D_1)),$$

and the probability of completing the nth experiment is

$$\gamma_n = \mathbb{P}(D_n < L \mid D_{n-1} < L) = \mathbb{E}(\exp(-\lambda[D_n - D_{n-1}])). \tag{A20}$$

We set $\gamma_1 = \gamma$. To capture the increasing difficulty of running experiments, we assume that the sequence $\gamma_1, \gamma_2, \gamma_3, \ldots$ is decreasing. Accordingly, $\gamma_n \leq \gamma$ for any n.[1]

F.2. Optimal stopping time

Even with increasingly difficult experiments, it is optimal for the scientist to p-hack until she reaches a significant result. First, it remains optimal for the scientist to engage in research because all the possible payoffs from research are positive. Second, it is optimal for the scientist to continue p-hacking when she obtains an insignificant result because she can only obtain equal or higher payoffs in the future. Third, it is optimal for the scientist to stop p-hacking when she obtains a significant result because it is impossible to obtain a higher payoff in the future.

F.3. Probability of type 1 error

We now compute the probability of type 1 error under the critical value z^* given by (9). Given that the scientist's behavior remains the same as in the basic model, we follow the same steps as in the proof of proposition 2.

[1] To obtain (A20), we note that the resource limit L is exponentially distributed with rate λ, and that the exponential distribution is memoryless, so $\mathbb{P}(L > d_n \mid L > d_{n-1}) = \mathbb{P}(L > d_n - d_{n-1}) = \exp(-\lambda[d_n - d_{n-1}])$ for any $d_n > d_{n-1} > 0$.

Probability of reporting a significant result at experiment j. All the steps of the proof of proposition 2 remain valid until we reach (A4). The probability that experiment j is completed given that $j-1$ experiments have already been completed is $\gamma_j \leq \gamma$. So equation (A4) becomes

$$\mathbb{P}\big(R(z^*) > z^*, N(z^*) = j \mid N(z^*) > j-1\big) = \gamma_j S(z^*) \leq \gamma S(z^*).$$

As a result, equation (A5) is modified:

$$\mathbb{P}\big(R(z^*) > z^*\big) = \sum_{j=1}^{\infty} \gamma_j S(z^*)\, \mathbb{P}(N(z^*) > j-1) \leq \gamma S(z^*) \cdot \sum_{j=0}^{\infty} \mathbb{P}(N(z^*) > j). \tag{A21}$$

Probability of completing more than j experiments. For $j = 0$, we have $\mathbb{P}(N(z^*) > j) = 1$. For $j \geq 1$, the term $\mathbb{P}(N(z^*) > j)$ gives the probability that the scientist conducts strictly more than j experiments. This event happens if the first j experiments could be completed, which occurs with probability $\prod_{k=1}^{j} \gamma_k \leq \gamma^j$, and if the first j test statistics were insignificant, which occurs with probability $F(z^*)^j$. For any $j \geq 0$, we therefore have

$$\mathbb{P}(N(z^*) > j) \leq \gamma^j F(z^*)^j,$$

which implies

$$\sum_{j=0}^{\infty} \mathbb{P}(N(z^*) > j) \leq \sum_{j=0}^{\infty} \gamma^j F(z^*)^j = \frac{1}{1-\gamma F(z^*)}. \tag{A22}$$

Bounding the probability of type 1 error. Combining equations (A21) and (A22), we obtain

$$\mathbb{P}\big(R(z^*) > z^*\big) \leq \frac{\gamma S(z^*)}{1-\gamma F(z^*)}.$$

Then using equation (A1), we bound the probability of type 1 error:

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

But the critical value z^* satisfies (8), so the right-hand side of the inequality is just the significance level α. We conclude that the probability of type 1 error is less than the significance level:

$$S^*(z^*) \leq \alpha.$$

References

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

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.

Ashenfelter, Orley C., and Michael Greenstone. 2004. "Estimating the Value of a Statistical Life: The Importance of Omitted Variables and Publication Bias." *American Economic Review* 94 (2): 454–460.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Labour Economics* 6 (4): 453–470.

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.

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.

Card, David, and Alan B. Krueger. 1995. "Time-Series Minimum-Wage Studies: A Meta-analysis." *American Economic Review* 85 (2): 238–243.

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.

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.

Dickersin, K., S. Chan, T.C. Chalmers, H.S. Sacks, and H. Smith Jr. 1987. "Publication Bias and Clinical Trials." *Controlled Clinical Trials* 8 (4): 343–353.

Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, Carrol Gamble, Davina Ghersi, John P. A. Ioannidis, John Simes, and Paula R. Williamson. 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. 2022. "Detecting P-hacking." *Econometrica* 90 (2): 887–906.

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, Christopher J., and Michael T. Brannick. 2012. "Publication Bias in Psychological Science: Prevalence, Methods for Identifying and Controlling, and Implications for the Use of Meta-

Analyses." *Psychological Methods* 17 (1): 120–128.

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

Gerber, Alan S., Donald P. Green, and David Nickerson. 2001. "Testing for Publication Bias in Political Science." *Political Analysis* 9 (4): 385–392.

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.

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., 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.

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

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.

Rose, Andrew K., and T. D. Stanley. 2005. "A Meta-Analysis of the Effect of Common Currencies on International Trade." *Journal of Economic Surveys* 19 (3): 347–365.

Ross, Sheldon. 2014. *A First Course in Probability*. 9th ed. Boston: Pearson.

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.

Simes, R. J. 1986. “Publication Bias: The Case for an International Registry of Clinical Trials.” *Journal of Clinical Oncology* 4 (10): 1529–1541.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–1366.

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.

Smaldino, Paul E., and Richard McElreath. 2016. “The Natural Selection of Bad Science.” *Royal Society Open Science* 3 (9): 160384.

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.

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.