The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. , NO. 0 • XXXX 2021

The Limits of *p*-Hacking: Some Thought Experiments

ANDREW Y. CHEN

ABSTRACT

Suppose that the 300+ published asset pricing factors are all spurious. How much *p*-hacking is required to produce these factors? If 10,000 researchers generate eight factors every day, it takes hundreds of years. This is because dozens of published *t*-statistics exceed 6.0, while the corresponding *p*-value is infinitesimal, implying an astronomical amount of *p*-hacking in a general model. More structure implies that *p*-hacking cannot address ≈100 published *t*-statistics that exceed 4.0, as they require an implausibly nonlinear preference for *t*-statistics or even more *p*-hacking. These results imply that mispricing, risk, and/or frictions have a key role in stock returns.

There is a well-known solution to every human problem—neat, plausible, and wrong.

— H.L. Mencken (1920), Prejudices: Second Series.

ACADEMICS HAVE DOCUMENTED MORE THAN 300 factors that help explain stock returns.[1] The documented factors are characterized by large *t*-statistics: statistics that are unlikely to occur by pure chance under the null of no effect. This enormous set of statistically significant factors begs for an economic explanation, yet there is little consensus on their origin.[2]

Andrew Y. Chen is with the Federal Reserve Board. I thank Rebecca John and Preston Harry for excellent research assistance. I thank Dino Palazzo and Fabian Winkler for many valuable discussions. I also thank Bastian von Beschwitz, Bjorn Eraker, Stefan Nagel (the editor), Dmitry Orlov, Emilio Osambela, Valery Polkovnichenko, Ivan Shaliastovich, Steve Sharpe, seminar participants at the Federal Reserve Board, and participants at the Marginal Revolution blog for helpful comments. Lastly, I am grateful to Francisco Palomino and Gustavo Suarez for their unwavering support of my research agenda. The views expressed herein are those of the authors and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System. The author does not have any potential conflicts of interest, as identified in *The Journal of Finance* Disclosure Policy.

[Correction added on 10 June 2021, after first online publication: Minor textual correction has been made at the end of Subsection B of Section I.]

Correspondence: Andrew Y. Chen, Federal Reserve Board; e-mail: andrew.y.chen@frb.gov.

[1] I use the term "factor" to refer to any variable that helps explain returns, following Harvey, Liu, and Zhu (2016).

[2] Cochrane (2017) provides a macrofinance perspective on predictability. Barberis (2018) provides a psychological perspective. Recent explicit factor models based on q-theory, the present value relation, and mispricing are given by Hou, Xue, and Zhang (2015), Fama and French (2015),

DOI: 10.1111/jofi.13036

p-Hacking (also known as data-snooping or data-mining) offers a neat and plausible solution (see Harvey, Liu, and Zhu (2016), Linnainmaa and Roberts (2018), and Chordia, Goyal, and Saretto (2020), among others). In its purest form, this cynical explanation argues that neither risk, mispricing, nor trading costs plays a role in the large *t*-statistics discovered by academics. Instead, the discovered *t*-statistics come from our collective search for the most "notable" results from a massive number of meaningless factors.

In this paper, I rigorously examine this pure *p*-hacking explanation. I write down a model in which there are many *p*-hacking attempts, each of which draws a *t*-statistic with a standard normal marginal distribution. Despite the thin tails of the standard normal, many observed *t*-statistics may be large, because only a selected subset makes it into circulated working papers. The model leads to the following lower bound:

$$\mathbb{E}(\text{Number of } p\text{-hacking attempts}) \geq \frac{\mathbb{E}[\text{Number of observed } (|t\text{-stats}| > \bar{t})]}{\Pr(|Z| > \bar{t})}, \tag{1}$$

where $\bar{t}$ is any *t*-statistic threshold and Z is a standard normal random variable. Intuitively, if one makes N hacking attempts, one should expect to find $N\Pr(|Z| > \bar{t})t$-statistics that exceed $\bar{t}$.

Equation (1) shows that the *p*-hacking story requires an absurd amount of hacking attempts. The cross-sectional asset pricing literature contains dozens of *t*-statistics that exceed 6.0 (Harvey, Liu, and Zhu (2016), Chen and Zimmermann (2020a)). Yet, for $\bar{t} = 6.0$, the denominator of equation (1) is infinitesimal ($\Pr(|Z| > 6.0) = 2.0 \times 10^{-9}$). Thus, the *p*-hacking story implies tens of billions of *p*-hacking attempts. To get a grip on this number, suppose that 10,000 researchers mine the data for 8 hours per day, 365 days per year, and complete on average one hacking attempt per hour. Even with this intense and dedicated *p*-hacking, it would take an expected 451 years to generate the literature.

Equation (1) is very general. It holds regardless of the correlation structure among *t*-statistics, and says relatively little about the *conditional* distributions of *t*-statistics. As a result, it holds under a wide variety of nonrandom searches through false factors, including repeated searches using the same sample, and even optimal directed search. I illustrate this generality by presenting several formal models that lead to equation (1), including a model of optimal *p*-hacking with learning that follows Adam's (2001) extension of Weitzman's (1979) "Pandora's boxes" model. These models also show how equation (1) applies to concepts like data-snooping (Lo and MacKinlay (1990)), multiple testing (Harvey, Liu, and Zhu (2016)), and publication bias (Chen and Zimmermann (2020b)).

The critical assumption is that the *marginal* distribution of *t*-statistics is standard normal under the null hypothesis that all factors are false. This concern leads me to focus on the Chen and Zimmermann (2020a) (CZ) data

and Stambaugh and Yuan (2016), respectively. Rigorous statistical explanations for cross-sectional predictability are proposed by Kozak, Nagel, and Santosh (2018), Kelly, Pruitt, and Su (2019), and Lettau and Pelger (2020).

set, which can be used to verify this assumption using a bootstrap test.[3] The bootstrap shows that the tails of the t-statistics in the CZ data closely match those of a standard normal distribution. This result is intuitive, as the CZ t-statistics test the hypothesis that the mean return on a long-short portfolio is zero, and this simple test is not subject to the many pitfalls that affect other tests like weak factors (Kan and Zhang (1999), Bryzgalova (2015)) or a strong factor structure in test assets (Lewellen, Nagel, and Shanken (2010)).

As a consequence of its generality, equation (1) only has bite for the roughly 30 observed factors with t-statistics that exceed 6.0. For factors with smaller t-statistics, the bound implies "only" millions of p-hacking attempts, which is possible if one is willing to entertain an extremely intense p-hacking effort.

More structured models, however, imply that p-hacking cannot address the roughly 100 factors with t-statistics that exceed 4.0. I demonstrate this using a few special cases of the general model.

The first special case examines the preference for large t-statistics implied by adding a simple discrete-choice problem: for each hacking attempt, a researcher writes the attempt into a working paper only if the payoff from writing exceeds a random disutility cost, where the payoff depends on the t-statistic. I specify a flexible functional form for this payoff, and estimate it using the distribution of observed t-statistics. The resulting preference is extremely nonlinear—t-statistics of 6.5 are 4 million times preferred to a t-statistic of 2.5, while t-statistics of 5.5 are 20,000 times preferred. Even t-statistics of 4.5 are 400 times preferred to a t-statistic of 2.5. Supposing that the disutility cost is just due to time, and that researchers are willing to spend 1 month writing a paper on a t-statistic of 2.5, this preference implies that researchers are willing to spend an implausibly long 33 years writing a paper with a t-statistic of 4.5.

The second special case imposes a set of plausible preferences and reexamines the implied amount of p-hacking. In this special case, the p-hacking attempts originate from a set of p-hacking "investigations." Each investigation begins by drawing an initial t-statistic, after which the reseacher chooses whether to continue the investigation. On continuing, the researcher draws an additional fixed number of variant t-statistics that each have correlation ρ with the initial t-statistic, and writes a paper about the variant with the largest t-statistic. Given a utility function over the largest t-statistic and a cost of drawing the variants, the model implies an expression for the expected number of p-hacking attempts that follows the same logic as equation (1). I find that, given plausible parameter values, even addressing t-statistics that exceed 4.0 requires an implausibly long 65 years of profession-wide devotion to p-hacking. This result is reminiscent of Lo and MacKinlay (1990), who find much smaller p-hacking effects than equation (1) in a plausibly structured model.

The last special case adds correlation information. I incorporate this information by deriving a variant of equation (1) that conditions on the realized t-statistics of a set of benchmark factors (e.g., the Fama-French factors). This

[3] I thank Bjorn Eraker for suggesting this test.

variant adjusts t-statistics for the predicted distribution implied by correlations with the benchmark factors. Using size, B/M, momentum, profitability, and investment as benchmarks, I find that roughly 40 of the CZ factors cannot be accounted for by p-hacking, when merely incorporating correlation information.

These thought experiments provide a framework for understanding the meaning, as well as the limits, of p-hacking and related concepts. p-Hacking, data-mining, data-snooping, and publication bias all boil down to multiple testing combined with selection bias. While selection bias can in principle lead to arbitrarily large t-statistics, not all kinds of selection bias are plausible. As a result, fundamental statistical results like the central limit theorem still have power in a world in which p-hacking is endemic.

Perhaps surprisingly, my findings are consistent with the results of Harvey, Liu, and Zhu (2016, HLZ) and Chordia, Goyal, and Saretto (2020, CGS). These papers are often cited for recommending a t-statistic hurdle of 3.0 or more. Less acknowledged is the fact that their hurdles imply a very small false discovery rate (FDR), and more than 100 published factors have t-statistics in excess of 3.0. Indeed, HLZ's simulated method of moments (SMM) estimates and CGS's calibrations imply that factors with t-statistics above 4.0 are almost guaranteed to be true discoveries, essentially the same result found in my more structured thought experiments.[4] This commonality likely stems from the use of a similar statistical framework. My models can be thought of as variants of HLZ and CGS, which assume that risk premiums and alphas are all zero, and that the empirical tail is generated entirely through selection bias. Other papers that are based on this framework include Chen (2019) and Chen and Zimmermann (2020b), who find quantitatively similar point estimates.

Other papers use out-of-sample tests to estimate the effects of p-hacking. These papers are often cited for finding a decline in predictability out-of-sample, but they also consistently show that the decline is far from 100%—sign that p-hacking is not the only force at work. In a study of 97 published cross-sectional predictors, McLean and Pontiff (2016) find that predictability declines by only 26% in the first few years postsample. Chen and Zimmermann (2020b) find a decline of 23% for their 156 predictors, and Jacobs and Müller (2020) find a decline of 38% for a set of 241 predictors. While data far from the in-sample period consistently produce larger declines, even these tests consistently find statistically significant predictability if all predictors are accounted for. This result is seen in the aforementioned studies as well as Linnainmaa and Robert's (2018) study of 36 predictors using data going back to 1926.

My paper can be thought of as a follow-up to Lo and MacKinlay's (1990) foundational study of data-snooping bias. In the decades since the publication

[4] More formally, a t-hurdle of 4.0 implies an expected false discovery proportion less than 1% in HLZ, and a greater than 95% probability that the false discovery proportion is less than 5% in CGZ. These results regard the probability of false discoveries conditional on the data, while my results regard the probability of the data conditional on all factors being false, but Bayes' rule implies that these probabilities are equal if either probability is zero.

of their classic paper, researchers have produced hundreds of new discoveries regarding the cross-section of returns. I push the limits of their framework to see if it can address these new discoveries. Consistent with their results, I find that data-snooping can significantly inflate the number of discoveries. However, data-snooping in and of itself still leaves many of these relatively recent discoveries unexplained.

The paper is organized as follows. In Section I, I present a general thought experiment. In Section II, I add more structure to obtain sharper inferences. In Section III, I conclude.

The code and data for all exhibits are available in the online article's Supporting Information section.

I. A Simple Thought Experiment

Suppose that the factors proposed in prior academic papers are all false, and that their large t-statistics are simply the result of p-hacking "attempts." We can boil this story down into just two equations.

First, each hacking attempt i results in a t-statistic with *marginal* distribution

$$t_i \sim N(0, 1), \quad i = 1, 2, \ldots, N, \tag{2}$$

where N is the total number of hacking attempts. One p-hacking attempt is one effort at creating a t-statistic that could potentially be shared with the academic community in a working paper. I use the term "hacking attempt" rather than "factor" to be clear that multiple hacking attempts may uncover the same factor or minor variations of the same factor. Further discussion of the hacking attempt concept is provided in Sections I.A and I.B below.

Second, attempt i is observed based on a function

$$\text{obs}(t_i, \theta_i) \in \{0, 1\}, \tag{3}$$

where θ_i is a set of other variables relevant to attempt i and attempt i is observed only if $\text{obs}(t_i, \theta_i) = 1$. Both θ_i and the function $\text{obs}(t_i, \theta_i)$ are general and abstract. The set θ_i may contain attributes of attempt i (e.g., a backstory, supporting results), t-statistics of other hacking attempts, as well as random noise. Similarly, $\text{obs}(t_i, \theta_i)$ can be a simple threshold that depends on $|t_i|$, a probabilistic threshold (as in HLZ), or a result derived from an optimization problem (as in Adam (2001)).

This model implies the following lower bound on the amount of p-hacking.

PROPOSITION 1: *Assuming that N and $t_1, t_2, \ldots$ satisfy technical conditions given in Appendix A, then for any threshold $\bar{t} \in \mathbb{R}$, we have the lower bound*

$$\mathbb{E}(N) \geq \frac{\mathbb{E}[\textit{Number observed } (|t_i| > \bar{t})]}{Pr(|Z| \geq \bar{t})}, \tag{4}$$

where Z is a standard normal random variable.

The proof is found in Appendix A. While the proof is technical due to the stochastic N, the constant-N proof is easy. If N is constant, the expected number of observed t-statistics that exceed $\bar{t}$ is

$$\begin{aligned}\mathbb{E}[\text{Number of observed } (|t_i| > \bar{t})] &= \mathbb{E}\left[\sum_{i=1}^{N} I(|t_i| > \bar{t})\text{obs}(t_i, \theta_i)\right] \\ &= \sum_{i=1}^{N} \mathbb{E}[I(|t_i| > \bar{t})\text{obs}(t_i, \theta_i)] \\ &\leq \sum_{i=1}^{N} \mathbb{E}[I(|t_i| > \bar{t})] \\ &= N\Pr(|Z| > \bar{t}),\end{aligned} \tag{5}$$

where $I(\cdot)$ is an indicator function. The key step is the second line, which moves the sum outside of the expectation. When N is constant, this is easily justified by the linearity of the expectations operator; for stochastic N, one needs technical conditions that produce a Wald's equation.[5] Finally, solving equation (5) for N gives the lower bound in Proposition 1.

Equation (5) holds regardless of the correlation structure. To see this, consider the extreme case in which every (t_i, θ_i) is perfectly correlated with (t_1, θ_1). In this case, the number of observed t-statistics that exceed $\bar{t}$ is given by

$$\text{Number of observed } (|t_i| > \bar{t}) = \begin{cases} N & \text{w/Prob } \Pr(\text{obs}(t_1, \theta_1) = 1 || t_1| > \bar{t})\Pr(|t_1| > \bar{t}) \\ 0 & \text{otherwise.} \end{cases}. \tag{6}$$

In other words, either all of the t-statistics exceed $\bar{t}$ or none do. Taking expectations and noting that $\Pr(\text{obs}(t_1, \theta_1) = 1||t_1| > \bar{t}) \leq 1$ therefore leads to the same upper bound as in equation (5).

Equation (6) raises the concern that high correlations can make p-hacking very efficient. If a single t-statistic happens to be large, hackers could draw an arbitrary number of additional large t-statistics by selecting perfectly correlated t-statistics. The academic community, however, is unlikely to accept a large set of highly correlated factors. Indeed, several papers find that the typical correlation across published factors is very close to zero, even after signing factors to have positive mean returns (McLean and Pontiff (2016), CZ).[6] I return to this issue in Section II.C, where I condition the estimation on the realized t-statistics of well-known factors. In addition, in Section II.C, I use a cluster bootstrap to account for correlation effects in standard errors.

[5] Wald's equation is also known as Wald's identity or Wald's lemma. Textbook proofs typically assume independence of t_i and t_j for $i \neq j$ (Cohen (2019)), perhaps due to the resulting elegance, but this independence is not required and can be replaced with weaker ergodicity or other technical conditions (Fuh and Lai (1998), Moustakides (1999), Fuh (2003)).

[6] Despite the near-zero typical pairwise correlation, principle component and related analyses can uncover notable commonality (Kozak, Nagel, and Santosh (2018), Kelly, Pruitt, and Su (2019), Lettau and Pelger (2020)).

The main analysis applies equation (4) to data on observed t-statistics to estimate a minimum amount of p-hacking. This inference can be justified with method of moments if the expected counts are expressed as expected fractions. I focus on equation (4), as this additional degree of formality does not affect the main results.

Readers familiar with the factor literature may already see the limits of p-hacking in Proposition 1. The literature contains dozens of t-statistics that exceed 6.0. Yet, $\Pr(|Z| > 6.0)$ is infinitesimal. Thus, for $\bar{t} = 6.0$, the right-hand side (RHS) of equation (4) divides a finite number by an infinitesimal number, implying an absurd amount of p-hacking on the left-hand side (LHS). One may ask whether a more reasonable amount of p-hacking can explain smaller observed t-statistics. To address this question, in the main analysis (Section I.C), I examine the amount of p-hacking implied by a variety of values of $\bar{t}$.

A. Model Discussion

The model nests many frameworks for multiple testing and data-snooping under the assumption that all factors are false. This section describes this nesting and is helpful for understanding Proposition 1, but is not required for the main results.

A.1. Nesting of Characteristics-Based Portfolios

Academic asset pricing frequently involves selecting some firm-level variables, defining portfolio weights as a simple function of these variables, and then examining the risk-adjusted returns of this portfolio. If risk-adjusted expected returns are zero, then this framework is nested within my model.

To see this nesting, consider the following process for constructing hacking attempts. There are months $t = 1, 2, \ldots, T$, and stocks $j = 1, 2, \ldots, M_t$ in month t. Each stock-month has a risk-adjusted return $r_{j,t}$ and a vector of end-of-month characteristics $\boldsymbol{x}_{j,t}$. Hacking attempt i involves selecting a portfolio weight function $g_i(\cdot)$ that maps characteristics into an M_t-dimensional end-of-month weight vector $\boldsymbol{w}_{i,t} = g_i(\boldsymbol{X}_t)$, where $\boldsymbol{X}_t \equiv [\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}, \ldots, \boldsymbol{x}_{M_t,t}]$. There are no restrictions on the sequence of functions $g_1(\cdot), g_2(\cdot), \ldots$, and thus multiple hacking attempts may produce highly correlated or even identical weights. The risk-adjusted return on attempt i in month t is the weighted average $f_{i,t} = \boldsymbol{w}'_{i,t-1}\boldsymbol{r}_t$, and the t-statistic is $t_i \equiv \frac{\bar{f}_i}{\mathrm{SE}_i}$ where $\bar{f}_i = T^{-1}\sum_{t=1}^{T} f_{i,t}$ and SE_i is a consistent estimator of $\sqrt{\mathrm{Var}(\bar{f}_i)}$.

Now, to formalize the notion that all factors are false, suppose that $\mathbb{E}_t(\boldsymbol{r}_{t+1}) = 0$, where $\mathbb{E}_t$ is the expectation with respect to the econometrician's information set at time t.[7] This restriction implies $\mathbb{E}(f_{i,t}) = \mathbb{E}(\boldsymbol{w}_{i,t}\mathbb{E}_t(\boldsymbol{r}_{t+1})) = 0$, and thus, $\mathbb{E}(\bar{f}_i) = 0$.

[7] This assumption nests the joint assumptions that expected risk-adjusted returns are zero conditional on investors' information and that the econometrician has less information than investors.

Assuming that the portfolio returns $f_{i,t}$ are well behaved, a central limit theorem implies that each $\bar{f}_i$ is asymptotically normal. Here, "well behaved" typically implies that temporally distant observations are independent (i.e., a mixing condition), which is a reasonable assumption as empirical portfolio returns have very little autocorrelation. Given $\bar{f}_i$ is asymptotically normal with zero mean (all factors are false), and SE_i is a consistent estimator of $\sqrt{\text{Var}(\bar{f}_i)}$, the delta method implies that t_i is asymptotically standard normal and equation (2) holds asymptotically.

Importantly, the single-variable central limit theorem applies regardless of the correlation matrix of portfolio returns. The correlations may exhibit a strong factor structure (Lewellen, Nagel, and Shanken (2010)), or high dimensionality (Feng, Giglio, and Xiu (2020)), a single-variable central limit theorem still applies to each portfolio.

This nesting means that Proposition 1 provides an alternative interpretation for Yan and Zheng's (2017) groundbreaking data-mining study. Yan and Zheng (2017) use Compustat accounting variables to construct 18,000 simple functions $g_i(\cdot)$ and find that some of their functions "exhibit superior long-short performance that is not due to sampling variation" using Fama and French's (2010) bootstrap test (see also Sullivan, Timmermann, and White (1999)). To use Proposition 1 instead, note that Yan and Zheng's (2017) table 1 shows that about $10\% \times 18{,}000 \approx 1{,}800$ of their t-statistics exceed 4.0 in absolute value (equal-weighted), and $\Pr(|Z| > 4.0) = 6.3 \times 10^{-5}$. Thus, Proposition 1 implies that

$$\mathbb{E}(N) \geq \frac{1{,}800}{6.3 \times 10^{-5}} \approx 29 \text{ million},$$

which is far larger than the true $N \approx 18,000$, and hence disproves the idea that data-mining alone underlies their data by contradiction.[8] A similar analysis can be applied to the "econometrician's data set" of CGS.

A.2. Relationship with Lo and MacKinlay's (1990) Data-Snooping Bias

My model is closely related to Lo and MacKinlay's (1990) data-snooping exercise. Lo and MacKinlay (1990) suppose that each firm characteristic is selected among a fixed K candidate characteristics based on its sample correlation with firm alphas, despite the fact that all population correlations are actually zero.

The "econometrician" refers to the person conducting the meta-analysis of all factors, not the researcher who hacks a specific factor.

[8] Harvey and Liu (2020) argue that "Yan and Zheng (2017) are interested in finding an unbiased estimate of the fraction of true discoveries," and point out that this estimate cannot be obtained using the Fama and French (2010) bootstrap. This argument, however, does not affect Yan and Zheng's (2017) finding that "the large actual t-statistics at the extreme percentiles cannot be explained by sampling variation," nor does it affect the computation of t-statistics found in Yan and Zheng's (2017, table 1).

They then form q portfolios by sorting on each characteristic, and construct a corresponding χ^2 test stat, Q_i.[9]

Proposition 1 can be applied in Lo and MacKinlay's (1990) exercise if we define each hacking attempt as a candidate characteristic and replace t-tests with χ^2 tests. Since the number of observed χ^2 tests is N/K, this variation of the model implies that

$$\mathbb{E}[\text{Fraction of observed } (Q_i > \bar{Q})] \leq K\Pr\left(\chi_q^2 \geq \bar{Q}\right), \tag{7}$$

where Q_i is the χ^2 test statistic for hacking attempt i, $\bar{Q} \in \mathbb{R}$ is a threshold, and χ_q^2 is a χ^2-distributed random variable with q degrees of freedom. If all tests are reported, proportion $\Pr(\chi_q^2 \geq \bar{Q})$ of the Q_i would exceed $\bar{Q}$, but data-snooping leads to a share as high as $K\Pr(\chi_q^2 \geq \bar{Q})$. In other words, Proposition 1 implies that p-values in Lo and MacKinlay's (1990) exercise are understated by at most a factor of K.

Equation (7) is a much weaker result than that in Lo and MacKinlay (1990). They find that assuming $K = 25$ ($K = 50$) results in p-values being understated by as much as a factor of 5 (7). These much smaller understatements imply that p-hacking is even less plausible, but they are derived assuming a particular data-snooping process. I revisit this theme of how stronger assumptions lead to sharper inferences in Section II.

A.3. Nesting of Optimal Directed Search (Adam (2001))

The Lo and MacKinlay (1990) exercise described above may overstate the amount of p-hacking required since it does not impose an intelligent search. That is, Section I.A.2 assumes that characteristics are randomly examined, but if researchers only spend time on more promising factors, then perhaps p-hacking could generate the literature in a reasonable amount of time. This concept is formalized in Adam's (2001) learning extension of Weitzman's (1979) "Pandora's boxes" model, which provides the optimal sequence of hacking attempts given a fixed set of candidate factors. In what follows, I explain the intuition behind Adam's (2001) model and its relationship with Proposition 1; a formal statement can be found in Appendix C.

The optimal strategy involves assigning each unexamined candidate a "reservation price," which captures the trade-off between the costs and benefits of conducting additional hacking attempts. At each stage in the search, the unexamined candidate with the highest reservation price is examined if the search is not ended. The search is ended when the highest payoff among examined candidates exceeds the highest reservation price among unexamined candidates.

Adam (2001) shows that the reservation price for each unexamined candidate takes on correlation information. Thus, if a previously examined

[9] This is examined in Section 4 of Lo and MacKinlay (1990). Sections 1 to 3 effectively assume that a characteristic predicts alphas by pure luck, which is very similar to a model extension I examine in Section II.C.

candidate happened to have a large t-statistic, the hacker may want to focus on candidates that are highly correlated, and in a sense hack "near" this previous attempt.[10] How exactly the reservation price depends on the correlations is determined by the hacker's preferences and other model parameters.

This kind of directed search affects the sequence of t-statistics through the conditional distribution $t_i|t_{i-1}$. Nevertheless, there is no effect on the *marginal* distribution of t_i, and thus equation (2) holds, as does Proposition 1. In other words, my model only imposes that t-statistics are standard normal *before* the hacking begins. Once t-statistic realizations are observed, the conditional distributions may be far from standard normal.

More formally, suppose that a directed search as described above is responsible for p-hacking attempts with indexes in the set $S \subset \{1, 2, \ldots\}$ and that this search results in an observed t-statistic t^*. Then t^* satisfies

$$\begin{aligned} \Pr(|t^*| > \bar{t}|S) &\leq \Pr\left(\max_{i \in S} |t_i| > \bar{t}\right) \\ &= \Pr(\cup_{i \in S}(|t_i| > \bar{t})) \\ &\leq \sum_{i \in S} \Pr(|t_i| > \bar{t}) \\ &= \#(S)\Pr(|Z| > \bar{t}), \end{aligned} \tag{8}$$

where $\bar{t}$ is any t-statistic cutoff and $\#(S)$ is the number of elements in S. The key step is the third line, which uses Boole's inequality, and the last line uses equation (2). Equation (8) closely resembles equation (5). Indeed, taking the expectation over S and summing across directed search problems leads exactly to a stochastic-N analog of equation (5), as I show in Appendix C.

A.4. Nesting of Benjamini and Hochberg's (1995) False Discovery Rate Control

Proposition 1 is implied by the Benjamini and Hochberg (1995) FDR control under the assumption that FDR = 1.

To see this, suppose that the Benjamini and Hochberg (1995) algorithm recommends rejecting the null for all factors with $|t_i| > \bar{t}$. In this case, their Theorem 1 implies[11]

$$\Pr(|Z| > \bar{t}) \geq \frac{\mathbb{E}[\text{Number of } (|t_i| > \bar{t})]}{N}\text{FDR}. \tag{9}$$

In my model, FDR = 1 (all factors are false). Thus, rearranging equation (9) and imposing the fact that $\mathbb{E}[\text{Number of observed } (|t_i| > \bar{t})] \leq$

[10] I thank Dmitry Orlov for providing me with this intuition.

[11] In the Benjamini and Hochberg (1995) model, N is constant. More formally, the Benjamini and Hochberg (1995) algorithm orders p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$ and rejects hypotheses with p-values smaller than p_{i^*}, where $i^* = \max\{i : p_{(i)} \leq iq/N\}$ and q is the desired control level. Equation (9) comes from assigning $\bar{t}$ s.t. $\Pr(|Z| > \bar{t}) = i^*q/N$ and noting that $i^* = \text{Number of } (|t_i| > \bar{t})$ and FDR $\leq q$. The Holm (1979) familywise error rate control implies an analogous lower bound on N.

$\mathbb{E}$[Number of $(|t_i| > \bar{t})$] leads to Proposition 1. Effectively, Proposition 1 flips the Benjamini and Hochberg (1995) algorithm on its head by assuming that the FDR is known to be one and then solving for N.

Similarly, equations (2) and (3) can be thought of as a variant of HLZ's "model with correlations" where all factors are assumed to be false, or a variant of Chen and Zimmermann (2020b) where all true returns are zero (see also Andrews and Kasy (2019)). Thus, while HLZ emphasize the concept of "multiple testing" and Chen and Zimmermann (2020b) emphasize "publication bias," these concepts along with p-hacking are simply the combination of conducting many tests (equation (2)) and reporting a selected subset (equation (3))).

Unlike Benjamini and Hochberg's (1995, Theorem 1, Proposition 1 does *not* assume that t_i are independent. This generality may be surprising given that Benjamini and Yekutieli (2001) derive a more stringent algorithm under general dependence (as emphasized in HLZ). However, many papers in the FDR literature find that equation (9) holds under a variety of types of dependence assumptions (Efron et al. (2001), Efron and Tibshirani (2002), Storey, Taylor, and Siegmund (2004), Reiner-Benaim (2007), Clarke and Hall (2009)). Indeed, Efron (2012) argues that equation (9) is valid as long as the empirical survival function of t-statistics ($\mathbb{E}$[Number of $(|t_i| > \bar{t})$]$/N$) is close to the true cumulative distribution function.[12]

B. Data on t*-Statistics from Academic Papers*

I apply equation (4) to two data sets: (i) CZ's reproductions of 210 cross-sectional stock return predictors from academic papers[13] and (ii) data based on HLZ hand-collected t-statistics for 315 factors from academic papers.

My primary data set is the CZ data. This data set consists of firm-level characteristics that have been shown to predict stock returns cross-sectionally in academic papers. The majority of CZ's characteristics are constructed using accounting data or market prices, but about one-third use diverse data sources that include analyst forecasts, trading-related measures, and corporate events. The CZ predictors cover 98% of the firm-level predictors in McLean and Pontiff (2016), 90% of the predictors from Green, Hand, and Zhang (2017), and 90% of the predictors that use widely available data in HLZ.

CZ restrict their characteristics to those that are observable at the time of portfolio construction. Thus, the CZ data can be modeled using Section I.A.1, and when examining these data, I am implicitly applying the null hypothesis that expected risk-adjusted returns are zero under the econometrician's information set. I use CZ's unadjusted long-short portfolio returns, as the literature generally finds small effects for standard risk adjustments (e.g., McLean and

[12] Benjamini and Yekutieli's (2001) more stringent algorithm is useful for highlighting the difficulties of producing formal results in a general dependence structure (Benjamini (2010), Fan, Han, and Gu (2012)).

[13] I refer to the benchmark CZ data of 210 clearly predictive and likely predictive characteristics as "the CZ data." Their full data set contains many other characteristics that were examined in other metastudies.

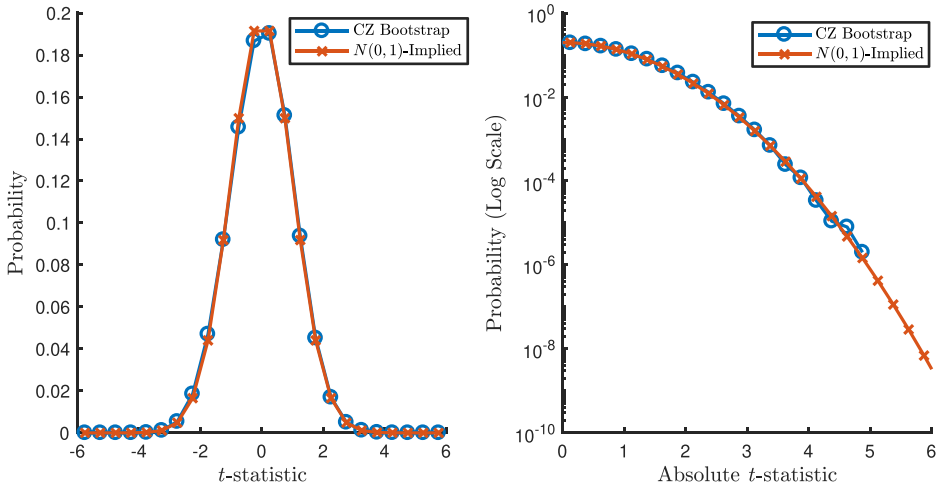

Figure 1. Empirical verification of the standard normal tails assumption. I draw random samples of 240 portfolio-months with replacement from the Chen-Zimmermann (2020a) data set of 210 long-short portfolios. All portfolio-months are pooled together before drawing samples. I calculate t-statistics by taking the sample mean, dividing by the standard deviation, and multiplying by the square root of 240. I subtract the mean t-statistic across samples to center at zero (circles) and compare with the standard-normal distribution (x's). The central limit theorem holds very well (equation (2)). (Color figure can be viewed at wileyonlinelibrary.com)

Pontiff (2016)). Whether standard risk adjustments should be updated to allow for high-dimensional risk (Feng, Giglio, and Xiu (2020)) is an interesting question but is outside of the scope of this paper.

In the CZ data, a p-hacking attempt represents one attempt at creating a firm-level stock return predictor. We observe only a subset of all hacking attempts—those in the CZ data. There are likely many more p-hacking attempts that were made during the creation of these observed attempts.

I focus on the CZ data because they are publicly available and consistent with the assumption that the marginal distribution of t-statistics is standard normal. This consistency can be shown both theoretically and empirically.

Theoretically, t-statistics in the CZ data are constructed by sorting stocks on a lagged predictor, forming long-short portfolios based on extreme quantiles, and dividing the raw sample mean return by its standard error. Thus, the t-statistics correspond to simple predictability tests that are not subject criticisms that affect many other asset pricing tests (Shanken (1992), Kan and Zhang (1999), Lewellen, Nagel, and Shanken (2010), Bryzgalova (2015)). Moreover, the openness of the CZ data allows one to verify that the t-statistics do not involve look-ahead bias, coding errors, or outright fraud. The code is publically available at https://github.com/OpenSourceAP/CrossSection/.

Empirically, I can verify that the CZ t-statistics have standard normal tails using a bootstrap test. Figure 1 shows this bootstrap, which is created by repeatedly drawing random samples of 240 portfolio-months from the CZ

Table I
An Absurd Amount of *p*-Hacking

I compute lower bounds on the amount of p-hacking (equation (4)). The expected number of p-hacking attempts implied by a given t-statistic minimum $\bar{t}$ is the number of observed $|t_i|$ that exceed $\bar{t}$ divided by the implied probability from a standard normal ($\Pr(|Z| > \bar{t})$). "Minimum years of hacking" assumes that 10,000 researchers generate eight t-statistics per day. The Chen and Zimmermann (2020a) data use open-source reproductions of 210 cross-sectional predictors with code and data available at https://sites.google.com/site/chenandrewy/open-source-ap. And 90% confidence intervals use a clustered bootstrap that preserves correlations within each month. Harvey-Liu-Zhu (2013)-implied data simulate their estimated model. Generating the dozens of observed t-statistics that exceed 6.0 would require, at minimum, hundreds of years of p-hacking.

	t-Statistic Minimum $\bar{t}$						
	2	3	4	5	6	7	8
$\Pr(\|Z\| > \bar{t})$	0.0455	2.7E-03	6.3E-05	5.7E-07	2.0E-09	2.6E-12	1.3E-15
Chen-Zimmermann (2020) Data							
Num Obs $\|t_i\| > \bar{t}$	189	130	77	44	26	17	12
90% CI	[156, 184]	[108, 148]	[65, 109]	[37, 72]	[22, 47]	[14, 28]	[8, 19]
Min Attempts $\mathbb{E}(N)$	4.2E+03	4.8E+04	1.2E+06	7.7E+07	1.3E+10	6.6E+12	9.0E+15
Min Years of Hacking	1.4E-04	1.6E-03	0.042	2.6	451	2.27E+05	3.08E+08
Harvey, Liu, and Zhu (2016)-Implied Data							
Num Obs $\|t_i\| > \bar{t}$	311	214	128	77	47	28	17
Min Attempts $\mathbb{E}(N)$	6.8E+03	7.9E+04	2.0E+06	1.3E+08	2.4E+10	1.1E+13	1.3E+16

returns, with all portfolio-months pooled together. For each sample, I calculate t-statistics following the CZ procedure, and subtract the mean t-statistic across samples to center the distribution at zero. Figure 1 plots the resulting distribution. The bootstrapped distribution (circles) is right on top of the standard normal (x's). Equation (2) is also verified by fitting a fat-tailed distribution to monthly returns, and then simulating many sample mean returns (see Appendix D).

One limitation of the CZ data is that they only contain characteristics that were shown to predict returns cross-sectionally. Thus, applying equation (4) to the CZ data produces, in a sense, a lower bound on the lower bound of the amount of p-hacking. To try to get closer to true lower bound, I also examine data based on HLZ's t-statistics from a broader set of factor tests. Results from HLZ's data set, however, may be affected by error-in-variables, useless factor, and other problems that do not affect CZ's simple portfolio sorts. I simulate the HLZ data set based on their SMM estimates (their Table 5).

C. An Absurd Amount of p-Hacking

Table I shows the main result of the paper: estimates of the minimum amount of p-hacking underlying the factor zoo. The table shows the lower

bound on the expected total number of p-hacking attempts $\mathbb{E}(N)$ implied by equation (4) for both the CZ and the HLZ data sets, as well as the inputs behind the calculation.

Both the CZ and the HLZ data sets have dozens of t-statistics that exceed 6.0. The probability that a random draw from $N(0, 1)$ exceeds 6.0, however, is an infinitesimal 2.0×10^{-9} (top row). As a result, the expected number of p-hacking attempts $\mathbb{E}(N)$ that underlie these data sets is at least 10^{10}.

To put 10^{10} p-hacking attempts in perspective, note that Yan and Zheng's (2017) data-mining study uses 240 accounting variables and 76 transformations to produce 18,113 predictors. One would have to repeat this exercise 50,000 times to generate 10^{10} p-hacking attempts. In an even larger data-mining study, CGS use a more flexible predictor-generation process to produce a whopping 2 million factors. This massive data-mining experiment is still 10,000 times smaller, however, than the lower bound on $\mathbb{E}(N)$.

An alternative way to put 10^{10} p-hacking attempts into perspective is to consider the following thought experiment. Suppose that 10,000 economists produce eight factors per day, 365 days per year. For comparison, the Bureau of Labor Statistics reports that there were 12,770 economics professors and 21,300 economists in the United States in 2017. Even with this absurd dedication of resources to p-hacking, it would take hundreds of years to generate 10^{10} attempts.

Data further out in the tail make the p-hacking explanation even more absurd. As seen in Table I, the CZ data contain 17 factors with $|t_i| > 7.0$ and 12 with $|t_i| > 8.0$. The HLZ data contain even more t-statistics that exceed these cutoffs. These t-statistics counts imply a minimum number of attempts on the order of 10^{12} or 10^{16}. Alternatively, they imply hundreds of thousands, or even hundreds of millions, of years of dedicated p-hacking.

One might argue that factors can be mined at a much faster rate than eight per economist-day, given modern computing power. However, factors need to come with supplementary results (monotonic returns, theoretical justifications) that satisfy journal review in order to be published. These additional restrictions are difficult to satisfy using computing power alone. Regardless, the largest t-statistics cannot be reconciled even with purely computing-based factor-mining. Supposing that factors are produced at 10 per economist-*second*, it would still take, in expectation, thousands of years for 10,000 economists to generate the 12 t-statistics larger than 8.0 in the CZ data.

While correlations do not affect the point estimates of the lower bound, they may affect the standard errors (Efron (2010)). To account for this effect, Table I provides 90% confidence intervals using a cluster bootstrap for the CZ histogram counts. This bootstrap draws random samples of months, and builds a resampled panel of portfolio returns using these months. This procedure preserves the correlation structure among portfolio returns within each month while accounting for the uncertainty in histogram counts induced by sampling noise. I set portfolio-months that are outside of the original papers' samples to missing before calculation of t-statistics, but I find that using all available

portfolio-months or subsets of months that tend to have more in-sample observations leads to similar results.

The resulting confidence intervals show that correlation-induced uncertainty has little effect on our lower-bound estimates. The 5^{th} percentile of bootstrapped histogram counts still finds many t-statistics that exceed 6.0 in absolute value. These results are perhaps intuitive, given that predictor portfolios' returns typically have zero correlation even after they are signed to have positive mean returns (McLean and Pontiff (2016), CZ). Indeed, some anomalies with very large t-statistics are known to have negative correlations (e.g., momentum and value), implying that small realized t-statistics for one portfolio implies large realized t-statistics for another portfolio.

Overall, Table I shows that a closer look at p-hacking alone cannot possibly explain the entire zoo of asset pricing factors. Thus, while this explanation is well known, neat, and plausible, it is also wrong.

II. More Structured Models, Sharper Bounds

The lower bounds in Section I are very general. They assume that t-statistics are properly constructed, but otherwise accommodate almost any notion of p-hacking. As a consequence, they only have bite for the roughly 30 factors with t-statistics that exceed 6.0.

In this section, I show how adding more structure produces sharper inferences. Section II.A adds a function describing the preference for large t-statistics, and Section II.B adds a simple and plausible directed p-hacking process. Both of these special cases imply that p-hacking cannot account for the roughly 100 factors with t-statistics that exceed 4.0.

To close this section, Section II.C demonstrates how adding correlation information increases the number of factors that p-hacking cannot explain.

A. An Implausible Preference for Large t*-Statistics*

The p-hacking bounds in Table I are extremely nonlinear: thousands of hacking attempts are required to generate t-statistics that exceed 2.0, but millions are needed to generate t-statistics that exceed 4.0. This nonlinearity implies an implausible preference for large t-statistics. To demonstrate this, I reexamine Proposition 1 through the lens of a simple discrete-choice model (Train (2009)).

Assume that equations (2) and (3) hold. In addition, assume that in hacking attempt i, the hacker chooses whether to write a working paper by maximizing utility function

$$u(\text{write}) = \begin{cases} p(|t_i|, \theta_i) - \xi_i & \text{write = yes} \\ 0 & \text{write = no} \end{cases}, \tag{10}$$

where θ_i is an object that captures attributes of hacking attempt i beyond its t-statistic (e.g., backstory, supporting results), $p(|t_i|, \theta_i)$ is the utility benefit

from writing about attempt i, and $\xi_i \sim U[0, 1]$ is the utility cost of writing the paper (e.g., time, effort, cognitive dissonance). Finally, the econometrician observes the hacking attempt only if the hacker writes about it.

Under these additional assumptions, we have

$$\Pr(\text{obs}(t_i, \theta_i) = 1 \big| |t_i|, \theta_i) = p(|t_i|, \theta_i). \tag{11}$$

Thus, the probability of observing a hacking attempt can be interpreted as the hacker's preference for a large t-statistic, holding constant the backstory and other attributes of the attempt. This preference captures both internal and external rewards to the hacker for writing about a hacking attempt with $|t_i|$ and θ_i. In a model that includes editors and referees, the external rewards will depend on the preferences of the editors and referees, as long as the hacker derives utility from getting a paper published (for example, Ellison (2002)). For simplicity, I focus on $p(|t_i|, \theta_i)$, which aggregates these external and internal preferences.

A.1. Preference Estimates

The complete function $p(|t_i|, \theta_i)$ cannot be estimated for two reasons: (i) θ_i is not observed and (ii) $p(|t_i|, \theta_i)$ is identified only up to a constant factor (Andrews and Kasy (2019)). However, assuming that θ_i and θ_j have identical marginal distributions for any i and any j (e.g., θ_i is stationary), I can estimate the following transformation of $p(|t_i|, \theta_i)$:

$$q(|t_i|) \equiv \frac{\mathbb{E}(p(|t_i|, \theta_i)||t_i|)}{\mathbb{E}(p(|t_i|, \theta_i)||t_i| \in [2, 3))}. \tag{12}$$

In words, $q(|t_i|)$ is the preference for hacking attempts with $|t_i|$ relative to $|t_i| \in [2, 3)$, averaging over the other attributes θ_i.

I parameterize $q(|t_i|)$ as a staircase function,

$$q(|t_i|) = \begin{cases} q_1 & |t_i| \in [e_1, e_2) \\ q_2 & |t_i| \in [e_2, e_3) \\ \dots & \\ q_K & |t_i| \in [e_K, e_{K+1}), \end{cases}$$

where $\{e_1, e_2, \dots, e_{K+1}\}$ is a set of increasing numbers. This flexible form leads to a simple estimator, as captured by the following proposition.

PROPOSITION 2: *Assuming that N, $\theta_1, \theta_2, \dots$, and $t_1, t_2, \dots$ satisfy technical conditions given in Appendix B, it follows that*

$$q_j = \frac{\mathbb{E}[\textit{Number of observed } (|t_i| \in [e_j, e_{j+1}))]}{\mathbb{E}[\textit{Number of observed } (|t_i| \in [2, 3))]} \left[\frac{Pr(|Z| \in [2, 3))}{Pr(|Z| \in [e_j, e_{j+1}))} \right]. \tag{13}$$

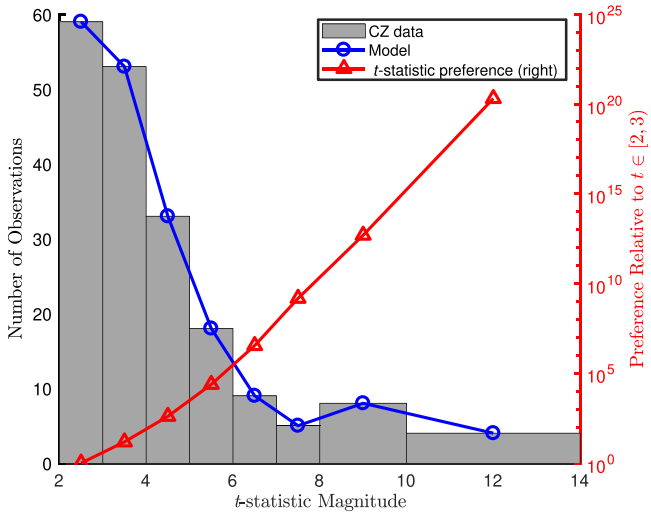

Figure 2. An extreme preference for large *t*-statistics. Bars show t-statistics for 210 replicated long-short portfolios from Chen and Zimmermann (2020a). The t-statistic preference (equation (13), triangles) is fit to CZ data. The p-hacking model (circles) exactly fits the CZ data (bars), but implies an implausibly nonlinear preference for large t-statistics. (Color figure can be viewed at wileyonlinelibrary.com)

The proof is in Appendix B. As with Proposition 1, technical assumptions are required to address a stochastic N. The intuition, however, is straightforward. Without a preference for t-statistics, the empirical distribution would decay like a standard-normal (the fraction on the right of the RHS). Thus, the deviation from standard normal (the fraction on the left of the RHS) identifies the preference (the LHS). Proposition 2 is closely analogous to Proposition 3 of Andrews and Kasy (2019).

Figure 2 shows the result of applying Proposition 2 to the data. In particular, it shows the resulting preference q_j (triangles), as well as the implied distribution of t-statistics (circles). The figure illustrates the powerful logic of p-hacking: one can generate any pattern if the data are selectively published in an arbitrary way. The p-hacking model (circles) exactly fits the empirical distribution (bars) in every histogram bin.

However, the implied preference for large t-statistics is so extreme that it needs to be shown on a log scale. Intuitively, the standard-normal terms in equation (13) increase like an inverse Gaussian, while the empirical t-statistic counts in equation (13) are roughly the same order of magnitude out to $|t| \approx 8$. Thus, the t-statistic preference is roughly quadratic, even on a log scale.

Indeed, a thought experiment shows that the implied t-statistic preference is absurd, even for t-statistics as "small" as 4.5. Suppose that the utility cost of writing a paper ξ_i is just time, and that researchers are willing to spend at most 1 month to write the typical paper with a t-statistic of 2.5. Then Figure 2

implies that researchers are willing to spend 400 months (33 years) to write a paper with a t-statistic of 4.5.

It is possible that this utility cost comes from something other than time. But this argument suggests that papers with small t-statistics have unusually small nontime costs, which is counterintuitive. It is usually harder on the soul to write about unconvincing statistical results.

Larger t-statistics lead to a simply ridiculous willingness-to-pay. Continuing this thought experiment, Figure 2 implies that researchers are willing to spend 2,000 years on a paper with a t-statistic of 5.5, and 30,000 years on a paper with a t-statistic of 6.5.

Even without mapping the utility cost to further details, one struggles to find any supporting evidence for the extreme preferences in Figure 2. If t-statistics > 4.0 were so desirable, certainly some papers would emphasize this fact. But in the academic world, such an emphasis is difficult to find. As a counterexample, papers that show stars to indicate the level of significance go up to at most ***, implying significance at the 1% level or a t-statistic of "only" 2.58. One could potentially obtain survey evidence regarding this t-statistic preference, but I expect such surveys are unnecessary. Asking oneself questions like "can you recall whether earnings surprise or B/M has a larger t-statistic?"or "can you name 10 anomalies with t-statistics above 4.0?" should be enough.

Importantly, unlike the absurd amount of p-hacking (Section I.C), this absurd preference result does not rely on the larger t-statistics among the empirical distribution. Rather, t-statistics that exceed 4.0 are all that is required, and about 40% of the hundreds of t-statistics in the asset pricing literature satisfy this threshold.

B. Plausible Preferences Imply Even More p-Hacking

Section II.A shows that the preference structure implied by Section I's estimates of the amount of p-hacking is implausible. Here, I impose plausible preferences and reexamine the amount of p-hacking. The model is a simple one in which p-hacking involves selecting the best among several "variants." It is intended to be similar to the real-world p-hacking that researchers are likely to be familiar with. This analysis is similar to Lo and MacKinlay's (1990) Section 4, which lays out an explicit p-hacking process and then examines the implied distortion in test statistics.

Suppose that the N p-hacking attempts originate from a sequence of p-hacking "investigations." Investigation j begins with the drawing of a random t-statistic $t_{j,1} \sim N(0, 1)$. After observing $t_{j,1}$, the researcher decides whether to pay a cost and draw $\bar{K}$ additional t-statistics $\{t_{j,2}, \ldots, t_{j,\bar{K}+1}\}$, where the marginal distributions $t_{j,k} \sim N(0, 1)$ but $t_{j,k}|t_{j,1} \not\sim N(0, 1)$ because[14]

$$\text{Corr}(t_{j,1}, t_{j,k}) = \rho > 0 \quad \forall k > 1. \tag{14}$$

[14] Some readers may find it more natural to think about correlation in characteristics. Equation (14) accommodates this alternative, as correlation in characteristics should lead to correlated t-statistics, for example, through correlated portfolio returns (e.g., footnote 16).

After drawing these variations, the reseacher writes a paper about the largest t-statistic in absolute value and receives a payoff $u(|t|)$ that depends on the resulting t-statistic. The econometrician observes only the chosen t-statistic—and only if the reseacher decides to pay the cost and draws the variations. Costs and payoffs are constant across investigations.

The researcher's optimal choice boils down to

$$\text{obs}_{j,k} = \begin{cases} 1 & (|t_{j,1}| > t_{\text{cont}}) \cap (k = \max\{|t_{j,1}|, |t_{j,2}|, \ldots, |t_{j,\bar{K}+1}|\}) \\ 0 & \text{otherwise,} \end{cases}$$

where $\text{obs}_{j,k} = 1$ indicates that investigation j's variant k is written into a working paper, and t_{cont} is t-statistic threshold. Given a functional form for the payoff $u(|t|)$, one could in principle solve for t_{cont} as a function of more "primitive" parameters. Rather than specify $u(|t|)$, I focus on the reduced form t_{cont} because $u(|t|)$ would ultimately be chosen based on intuitive values for t_{cont}.

B.1. Parameter Choices and Model Illustration

In what follows, I focus on $t_{\text{cont}} = 1.80$ and $\bar{K} = 10$. These assumptions can be thought of as a representative researcher who will continue working on an investigation if the initial t-statistic is just below the 5% significance cutoff of 1.96, and for whom generating 10 variations of the initial t-statistic has a small marginal cost. Fixing t_{cont} and $\bar{K}$ allows me to examine several values for ρ: 0.50, 0.70, and 0.90.

Figure 3 illustrates the results of these parameter values. The figure plots the distribution of observed $|t_i|$ implied by the model (bars) and compared to a standard-normal (line). Compared to the standard-normal, this intuitive form of p-hacking leads to a much higher density of t-statistics above 2.0—naturally, only investigations j that begin with $|t_{j,1}| > 1.80$ are observed. Moreover, the density peaks above 2.0.

However, Figure 3 shows that this form of p-hacking fails to produce a notable density of t-statistics above 4.0. All three parameterizations lead to a negligible density in this range. This sharp decay can be understood through the Boole inequality, which implies that

$$\Pr(\max\{|t_{j,1}|, |t_{j,2}|, \ldots, |t_{j,\bar{K}+1}|\} > \bar{t}) \leq (\bar{K}+1)\Pr(|Z| > \bar{t}),$$

for any t-statistic cutoff $\bar{t}$. Thus, maximizing over variations improves over a single random draw by at most a factor of $\bar{K}+1$, and the tails of from this kind of p-hacking are still Gaussian. A similar decay is implied by Gaussian concentration inequalities (Adler and Taylor (2007)).

B.2. Plausible Preference Results

This model is a special case of Section I. To see this, note that the sequence $t_1, t_2, t_3, \ldots$ can be defined by simply stacking the t-statistics from the

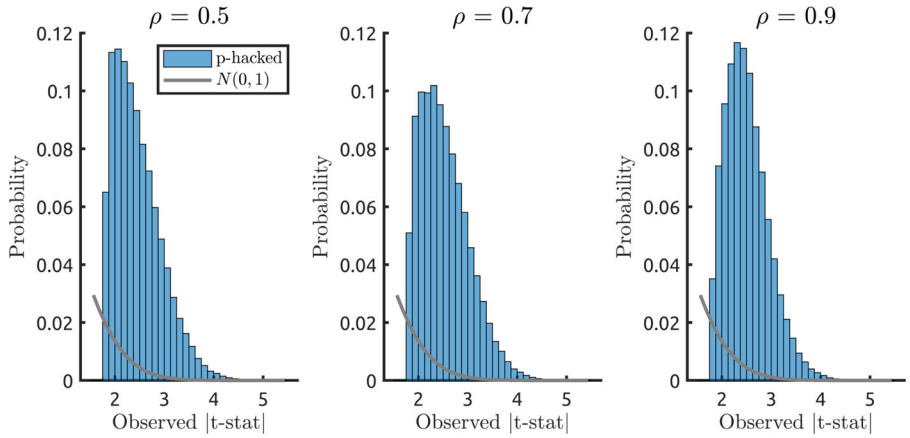

Figure 3. Some plausible preference specifications. I simulate a three variations of the model of Section II.B, which assumes that p-hacking involves drawing an initial t-statistic and continuing to draw 10 variations on the initial t-statistic if the initial t-statistic exceeds 1.80 in absolute value. Researchers prefer to make t-statistics observable only if the initial draw exceeds 1.80 and the t-statistic is the largest among the 10 variations. I simulate three models with different values for ρ, the correlation between the variant t-statistics and the initial t-statistic (bars). I also show the density implied by a standard-normal (line) for comparison. These preference specifications lead to more significant observed t-statistics than a simple standard-normal, but only for t-statistics < 4.0 in absolute value. (Color figure can be viewed at wileyonlinelibrary.com)

investigation variants $\{t_{1,k}\}_{k=1}^{K_1}$, $\{t_{2,k}\}_{k=1}^{K_2}$, ..., where K_j is the number of variants examined in investigation j.

Indeed, the additional structure allows me to derive a refined version of Proposition 1:

$$\mathbb{E}[N] = \frac{\mathbb{E}[\text{Num observed } (|t_{j,k}| > \bar{t})]}{\Pr(|Z| > \bar{t})} \frac{\mathbb{E}[K_j]}{\mathbb{E}\left[\sum_{k=1}^{K_j} \Pr(\text{obs}_{j,k} = 1 | |t_{j,k}| > \bar{t})\right]}, \quad (15)$$

where, as in Proposition 1, $\bar{t}$ is a t-statistic threshold.[15] Note that the first term in equation (15) is simply the lower bound given by equation (4). Thus, the effect of this form of p-hacking is captured by the second term, which can

[15] To see this, note that

$$\begin{aligned}\mathbb{E}[\text{Num observed } (|t_{j,k}| > \bar{t})] &= \mathbb{E}\left[\sum_{j=1}^{J}\sum_{k=1}^{K_j} I(|t_{j,k}| > \bar{t} \cap \text{obs}_{j,k} = 1)\right] \\ &= J\mathbb{E}\left[\sum_{k=1}^{K_j} \Pr(\text{obs}_{j,k} = 1 | |t_{j,k}| > \bar{t}) \Pr(|t_{j,k}| > \bar{t})\right] \\ &= J\mathbb{E}\left[\sum_{k=1}^{K_j} \Pr(\text{obs}_{j,k} = 1 | |t_{j,k}| > \bar{t})\right] \Pr(|Z| > \bar{t})\end{aligned}$$

and $\mathbb{E}(N) = J\mathbb{E}(K_j)$.

Table II

The Amount of *p*-Hacking Implied by a Plausible Preference for Large *t*-statistics

I compute the amount of p-hacking implied by the model in Section II.B. Hacking attempts originate from "investigations" in which initial t-statistics < 1.80 are discarded, only the largest t-statistic among 10 variations of the initial t-statistic are reported, and all t-statistic variations have correlation ρ with the initial t-statistic. The amount of hacking depends on the subset of the data that the econometrician tries to explain and is given by the t-statistic minimum $\bar{t}$ (equation (15)). The years of hacking assumes that 10,000 researchers generate eight t-statistics per day. Equation (15) is computed using 10^6 Monte-Carlo simulations* and indicates that the simulation generates no observations that meet $\bar{t}$ and so the estimate is a lower bound from assuming $\mathbb{E}[\sum_{k=1}^{K_j} \Pr(\text{obs}_{j,k}||t_{j,k}| > \bar{t})] = 10^{-6}$. This preference specification implies that even t-statistics that exceed 4.0 cannot be explained by p-hacking.

	t-Statistic Minimum $\bar{t}$						
	2	3	4	5	6	7	8
Correlation between Variants $\rho = 0.5$							
Attempts $\mathbb{E}(N)$	5.6E+04	3.8E+06	2.5E+09	7.6E+12	1.0E+16*	5.2E+18*	7.6E+21*
Years of Hacking	1.9E-03	1.3E-01	84.7	2.6E+05	3.6E+08*	1.8E+11*	2.6E+14*
Correlation between Variants $\rho = 0.7$							
Attempts $\mathbb{E}(N)$	5.3E+04	3.1E+06	2.0E+09	5.1E+12	1.0E+16*	5.2E+18*	7.6E+21*
Years of Hacking	1.8E-03	1.1E-01	67.9	1.7E+05	3.6E+08*	1.8E+11*	2.6E+14*
Correlation between Variants $\rho = 0.9$							
Attempts $\mathbb{E}(N)$	5.1E+04	4.0E+06	3.2E+09	1.0E+13	1.0E+16*	5.2E+18*	7.6E+21*
Years of Hacking	1.8E-03	1.4E-01	111.1	3.5E+05	3.6E+08*	1.8E+11*	2.6E+14*

be computed using Monte Carlo. As in the main analysis, I examine a variety of values for $\bar{t}$ to see which elements of the data the model is successful in addressing.

Table II shows the result of applying equation (15) to the CZ data set using parameter values from Section II.B.1. Throughout the table, the implied amount of p-hacking is much larger than the lower bounds in Table I. The deviation becomes dramatic when trying to fit the data on t-statistics that exceed 4.0. In this range, variant p-hacking implies on the order of 10^9 hacking attempts, while the general model implies a lower bound of 10^6.

Indeed, estimating the amount of hacking attempts required to generate t-statistics above 6.0 runs into numerical problems, as observing such t-statistics in a simulation is exceedingly rare. My simulations of 10^6 hacking investigations result in no t-statistics that exceed 6.0, and so for these estimates, I plug in $\mathbb{E}[\sum_{k=1}^{K_j} \Pr(\text{obs}_{1,k}||t_{1,k}| > \bar{t})] = 10^{-6}$ to produce a lower bound on equation (15) (as indicated by *). Regardless, the amount of p-hacking implied by these larger t-statistics is absurd. If one targets t-statistics that exceed 6.0, variant p-hacking requires at least 10^{16} attempts, compared to the general lower bound of 10^{10}.

To put these numbers into perspective, Table II examines the implied years of hacking, assuming that 10,000 researchers generate eight p-hacking attempts every day (same as in Table I). The table shows that p-hacking even struggles to account for the t-statistics that exceed 4.0, with all correlation specifications implying at least 65 years of devoted p-hacking effort required. This absurd conclusion is consistent with the results from the preference estimation (Section II.A), although the two calculations use very different frameworks. These stronger bounds are also consistent with Lo and MacKinlay (1990), who find that an intuitive model of p-hacking leads to much smaller effects than the limits imposed by Proposition 1 (see Section I.A.2).

Interestingly, Table II shows that the correlation between variants has relatively little effect on the amount of p-hacking required. This result is likely due to the fact that correlations produce offsetting effects in the mean and variance of t-statistics. To see this, note that after observing $t_{j,1}$, the researcher knows

$$t_{j,k}|t_{j,1} \sim N\Big(\rho t_{j,1}, \sqrt{1-\rho^2}\Big). \tag{16}$$

While high correlations increase the expected t-statistic, high correlations also reduce the variance, and thus, limit the ability of the variants to improve on the initial draw.

C. Using Correlation Information Implies More p-Hacking

The point estimates in Section I do not use correlation information. Adding this information can lead to sharper inferences, which, in turn, imply that p-hacking is even less plausible than found in Section I.

To understand how correlation can lead to sharper inferences, note that value and momentum strategies both have very positive mean returns, despite negatively correlated monthly returns. Thus, conditioning on this negative correlation implies even stronger evidence against the null of no predictability than would be obtained by considering only the positive mean returns alone. Similar intuition is seen in multiple testing statistics (Efron (2007), Fan, Han, and Gu (2012)), although the methods from this literature cannot be directly applied due to the selection bias in the published results.

More formally, suppose that the t-statistics in the model of Section I can be divided into two groups: a "benchmark" group and an "evaluation" group. Let the vector $\boldsymbol{t}_b$ represent the t-statistics from the benchmark group, and t_i be a t-statistic from the evaluation group. Assume that all t-statistics are multivariate normal, and denote the joint distribution by

$$\begin{bmatrix} t_i \\ \boldsymbol{t}_b \end{bmatrix} \sim N\left(0, \begin{bmatrix} 1 & \boldsymbol{C}_i \\ \boldsymbol{C}_i' & \boldsymbol{C}_b \end{bmatrix}\right), \tag{17}$$

where $\boldsymbol{C}_b \equiv \text{Cov}(\boldsymbol{t}_b)$ and $\boldsymbol{C}_i \equiv \text{Cov}(t_i, \boldsymbol{t}_b)$. The multivariate normal assumption can be justified by a multivariate central limit theorem.

Conditional on realized values of $\boldsymbol{t}_b$, normal-normal updating formulas imply

$$\frac{t_i - \mu_i}{\sigma_i} \Big| \boldsymbol{t}_b, \boldsymbol{C}_i, \boldsymbol{C}_b \sim N(0, 1), \tag{18}$$

where

$$\mu_i \equiv \boldsymbol{C}_i \boldsymbol{C}_b^{-1} \boldsymbol{t}_b, \tag{19}$$

$$\sigma_i \equiv \sqrt{1 - \boldsymbol{C}_i \boldsymbol{C}_b^{-1} \boldsymbol{C}_i'}. \tag{20}$$

Equations (18) to (20) show how to adjust t-statistics for correlations, if we take as given the realized values of $\boldsymbol{t}_b$. Intuitively, if the elements of $\boldsymbol{t}_b$ are large and positive, high correlations ($\boldsymbol{C}_i$ elements close to one) imply that t_i is expected to be large, and thus needs to be adjusted downward. High correlations also imply precise information about the expected value of t_i (small σ_i), and thus can also lead to an upward adjustment. These formulas and intuitions are similar to those of Propositions 1.1 and 1.2 from Lo and MacKinlay (1990).

This additional structure leads to the following variation of Proposition 1.

COROLLARY 1: *Assuming the technical conditions required for Proposition 1, then for any threshold* $\bar{t} \in \mathbb{R}$*, we have*

$$\mathbb{E}(N|\boldsymbol{t}_b, \boldsymbol{C}_i, \boldsymbol{C}_b) \geq \frac{\mathbb{E}\left[\textit{Number observed}\left(\left|\frac{t_i - \mu_i}{\sigma_i}\right| > \bar{t}\right) \Big| \boldsymbol{t}_b, \boldsymbol{C}_i, \boldsymbol{C}_b\right]}{Pr(|Z| \geq \bar{t})}, \tag{21}$$

where Z *is a standard normal random variable.*

The proof comes from simply replacing $|t_i|$ with $\frac{t_i - \mu_i}{\sigma_i}$ in the proof for Proposition 1.

To take Corollary 1 to the data, I estimate $\boldsymbol{C}_i$ and $\boldsymbol{C}_b$ by using the fact that the correlation between monthly returns is a good estimate of the correlation between t-statistics if monthly returns have little autocorrelation.[16] I then consider two sets of benchmark factors: (i) size, B/M, 12-month momentum, operating profitability, and investment, and (ii) all 19 factors in the CZ data set

[16] Under no autocorrelation in returns, the covariance between t-statistics t_j and t_k is

$$\begin{aligned}
\text{Cov}\left(\frac{\bar{r}_j}{\sigma_j^{(r)}/\sqrt{T}}, \frac{\bar{r}_k}{\sigma_k^{(r)})/\sqrt{T}}\right) &= \frac{T}{\sigma_j^{(r)}\sigma_k^{(r)}} \text{Cov}\left(T^{-1}\sum_t r_{j,t}, T^{-1}\sum_t r_{k,t}\right) \\
&= \frac{1}{\sigma_j^{(r)}\sigma_k^{(r)} T}\sum_t \text{Cov}(r_{j,t}, r_{k,t}) \\
&= \text{Corr}(r_{j,t}, r_{k,t}).
\end{aligned}$$

Table III

The Amount of *p*-Hacking Implied by Correlation Information

I compute lower bounds on the amount of *p*-hacking implied by factors' correlations with benchmark factors (equation (21)). Adjusted *t*-statistics use the mean and standard deviation implied by normal-normal updating (equations (17) to (20)), assuming that all factors are false. Years of hacking assume that 10,000 researchers produce eight hacking attempts per day. Applying correlation information implies that *p*-hacking cannot address the roughly 40 factors in the CZ data with adjusted *t*-statistics that exceed 6.0.

	Adjusted t-Statistic Minimum $\bar{t}$						
	2	3	4	5	6	7	8
Benchmark = Size, B/M, Momentum, Operating Profitability, Investment							
Num Obs Adjusted $\lvert t_i \rvert > \bar{t}$	141	105	74	59	43	27	18
Min Attempts $\mathbb{E}(N)$	3.1E+03	3.9E+04	1.2E+06	1.0E+08	2.2E+10	1.1E+13	1.4E+16
Min Years of Hacking	1.1E-04	1.3E-03	4.0E-02	3.5	746	3.6E+05	4.6E+08
Benchmark = all 19 factors published before 1994							
Num Obs Adjusted $\lvert t_i \rvert > \bar{t}$	142	114	90	58	40	28	21
Min Attempts $\mathbb{E}(N)$	3.1E+03	4.2E+04	1.4E+06	1.0E+08	2.0E+10	1.1E+13	1.6E+16
Min Years of Hacking	1.1E-04	1.4E-03	4.9E-02	3.5	694	3.7E+05	5.4E+08

that were published before 1994 (this includes size, B/M, and 12-month momentum). The results are reported in Table III.

The table shows that roughly 40 of the correlation-adjusted *t*-statistics exceed 6.0. This implies that underlying these 40 *t*-statistics are 2×10^{10} hacking attempts, or roughly 700 years of *p*-hacking. For comparison, without the correlation adjustments, the CZ data contain 26 *t*-statistics that exceed 6.0. Thus, the *p*-hacking story has even more difficulty addressing the data once correlation information is employed.

These results come from the fact that signed predictive portfolios have correlations that are dispersed around zero (McLean and Pontiff (2016), Chen and Zimmermann (2020b)). These average-zero correlations result in relatively small average effects on the conditional mean of *t*-statistics (equation (19)), but simultaneously imply that the conditional standard deviation is smaller than one (equation (20)), as the standard deviation adjustment involves squaring correlations.

III. Conclusion

The idea that all asset pricing factors are due to *p*-hacking is appealing. In one fell swoop, *p*-hacking seems to explain decades of puzzling financial research. A rigorous exploration of this solution, however, shows that it is implausible. The amount of *p*-hacking required is astronomical and the implied preference for *t*-statistics does not stand up to introspection. In particular, while *p*-hacking could potentially explain the observed factors with relatively

small t-statistics, it cannot explain the factors with t-statistics that exceed 4.0, and there are roughly 100 such observed factors.

Although these findings are negative for the p-hacking story, they imply that the many researchers who continue to study asset pricing factors are pursuing meaningful financial market phenomena. For the roughly 100 observed factors that cannot be explained by p-hacking, mispricing, risk, and/or frictions are surely playing a role.

Initial submission: August 7, 2019; Accepted: October 29, 2020
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

Appendix A: Technical Assumptions and Proof of Proposition 1

DEFINITION A1: The technical assumptions for Proposition 1 are

$$\exists M \in \mathbb{N} \quad s.t. \quad N \leq M \tag{A1}$$

$$\forall i, \quad t_{i+1} \quad \text{and} \quad \{N \leq i\} \text{ are independent.} \tag{A2}$$

DISCUSSION OF DEFINITION A1: Equation (A1) is sufficient for moving a sum outside of an expectation. More general conditions also allow for this operation, but the resulting proof is much more technical (e.g., using Lebesgue monotone convergence).

Equation (A2) implies that t_i for $i > N$ has some independence from N. Textbook proofs of Wald's equation typically generate this independence by assuming that N is a stopping time and t_i and t_j are independent for $i \neq j$. Other conditions are sufficient for generating equation (A2), however. For example, it also holds if $t_1, t_2, \ldots$ are created by appending independent $1 \times K$ vectors $\boldsymbol{t}_j$, and N is restricted to be a multiple of K (i.e., if the hacking attempts consist of investigations where t-statistics are uncorrelated across investigations but may be correlated within investigations). Fuh and Lai (1998), Moustakides (1999), and Fuh (2003) provide alternative conditions for Wald's equations that typically involve ergodicity assumptions and large N approximations.

More generally, equation (A2) is implicitly assumed in almost any empirical analysis in finance. Most empirical analyses assume that the sample size is a constant. But strictly speaking, the sample size should be considered a random variable, and thus a Wald equation is necessary for computing any data moment. This insight is a core motivation of Wald's (1945) seminal work. Thus, a condition like equation (A2) is implicitly assumed whenever a finance study calculates a moment condition.

For a deeper understanding of equation (A2), it helps to assume a specific process for the determination of N. Intuitively, N should be determined by the decision of the econometrician to undertake an analysis. Suppose that after observing a constant N_{obs} interesting hacking attempts, the econometrician

decides to run an analysis. Let M be the index of the hacking attempt corresponding to the N_{obs} observations. The let N be an index far enough from M that the corresponding hacking attempts are independent of the first M hacking attempts. Then by construction, N and t_i satisfy equation (A2).

PROOF OF PROPOSITION 1: With Definition A1 in hand, I can prove a Wald's equation for sums of functions of t_i. ■

LEMMA A1: *If the technical conditions in Definition A1 hold and, for all* $i = 1, \ldots, M, t_i$ *have identical marginal distributions, then for any real-valued function* $f(\cdot)$*,*

$$\mathbb{E}\left(\sum_{i=1}^{N} f(t_i)\right) = \mathbb{E}[N]\mathbb{E}[f(t_i)].$$

PROOF OF LEMMA A1: Using assumption (A1), and the fact that $I(i \leq N) = 1 - I(i - 1 \geq N)$, I can express the sum as

$$\begin{aligned}\sum_{i=1}^{N} f(t_i) &= \sum_{i=1}^{M} f(t_i) I(i \leq N) \\ &= \sum_{i=1}^{M} f(t_i)[1 - I(i - 1 \geq N)].\end{aligned}$$

Then taking expectations, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N} f(t_i)\right] &= \mathbb{E}\left[\sum_{i=1}^{M} f(t_i)[1 - I(i - 1 \geq N)]\right] \\ &= \sum_{i=1}^{M} \mathbb{E}[f(t_i)[1 - I(i - 1 \geq N)]] \\ &= \sum_{i=1}^{M} \mathbb{E}[f(t_i)]\mathbb{E}[[1 - I(i - 1 \geq N)]] \\ &= \mathbb{E}[f(t_1)] \sum_{i=1}^{M} \Pr(i \leq N),\end{aligned}$$

where the second line uses the linearity of expectations, the third line uses assumption (A2), and the fourth line uses the assumption that all marginal distributions of $t_1, t_2, \ldots, t_N$ are identical. Finally, the tail-sum formula implies that $\sum_{i=1}^{M} \Pr(i \leq N) = \sum_{i=1}^{\infty} \Pr(i \leq N) = \mathbb{E}(N)$.

To finish proving Proposition 1, apply Lemma 1 to the stochastic-N version of equation (5):

$$\begin{aligned}\mathbb{E}[\text{Number observed } |t_i| > \bar{t}] &= \mathbb{E}\left[\sum_{i=1}^{N} I(|t_i| > \bar{t})\text{obs}(t_i, \theta_i)\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{N} I(|t_i| > \bar{t})\right] \\ &= \mathbb{E}[N]\mathbb{E}[I(|t_1| > \bar{t})] \\ &= \mathbb{E}[N]\Pr(|Z| > \bar{t}), \end{aligned} \tag{A3}$$

where the second line uses the fact that $\text{obs}_i \leq 1$. Finally, solving for $\mathbb{E}(N)$ gives equation (4). ■

Appendix B: Technical Assumptions and Proof of Proposition 2

Before I state the technical assumptions and prove the proposition, I first show that under the special case of N constant, the proof is easy. In this case,

$$\begin{aligned}&\mathbb{E}[\text{Num observed } (|t_i| \in [e_j, e_{j+1}))] \\ &= \mathbb{E}\left[\sum_{i=1}^{N} I(t_i \in [e_j, e_{j+1}))\text{obs}(t_i, \theta_i)\right] \\ &= \sum_{i=1}^{N} \mathbb{E}[I(t_i \in [e_j, e_{j+1}))\text{obs}(t_i, \theta_i)] \\ &= \sum_{i=1}^{N} \Pr(\text{obs}(t_i, \theta_i) = 1 || t_i| \in [e_j, e_{j+1}))\Pr(|t_i| \in [e_j, e_{j+1})) \\ &= \sum_{i=1}^{N} \mathbb{E}(p(|t_i|, \theta_i)||t_i| \in [e_j, e_{j+1}))\Pr(|t_i| \in [e_j, e_{j+1})) \\ &= N\mathbb{E}(p(|Z|, \theta_1)||Z| \in [e_j, e_{j+1}))\Pr(|Z| \in [e_j, e_{j+1})), \end{aligned} \tag{B1}$$

where I have used the fact that the expectation is linear, equation (11), and the fact that the marginal distributions of t_i are identical and the same holds for θ_i. Then taking the ratio of the above equation with the same equation for

$[e_j, e_{j+1}) = [2, 3)$, we have

$$\begin{aligned}
&\frac{\mathbb{E}[\text{Number observed } (|t_i| \in [e_j, e_{j+1}))]}{\mathbb{E}[\text{Number observed } (|t_i| \in [2, 3))]} \\
&= \frac{\mathbb{E}(p(|Z|, \theta_1)||Z| \in [e_j, e_{j+1}))\Pr(|Z| \in [e_j, e_{j+1}))}{\mathbb{E}(p(|Z|, \theta_1)||Z| \in [2, 3))\Pr(|Z| \in [2, 3))} \\
&= q_j \frac{\Pr(|Z| \in [e_j, e_{j+1}))}{\Pr(|Z| \in [2, 3))}. \qquad \text{(B2)}
\end{aligned}$$

Finally, solving for q_j yields equation (13).

Now see below for the technical assumptions and general proof.

DEFINITION B1: The technical assumptions for Proposition 2 are

$$\exists M \in \mathbb{N}\, s.t.\, N \leq M \qquad \text{(B3)}$$

$$\forall i, \quad (t_{i+1}, \theta_{i+1}) \quad \text{and} \quad \{N \leq i\} \text{ are independent .} \qquad \text{(B4)}$$

In contrast to equation (A2), equation (B4) requires an assumption on θ_{i+1} that is required to ensure a Wald's equation that can apply to functions of both t_i and θ_i.

PROOF OF PROPOSITION 2: Following the proof of Lemma A1, we have

$$\begin{aligned}
&\mathbb{E}[\text{Number observed } (|t_i| \in [e_j, e_{j+1}))] \\
&= \sum_{i=1}^{M} \mathbb{E}[I(t_i \in [e_j, e_{j+1}))\text{obs}(t_i, \theta_i)1 - I(i - 1 \geq N)] \\
&= \sum_{i=1}^{M} \mathbb{E}[I(t_i \in [e_j, e_{j+1}))\text{obs}(t_i, \theta_i)]\mathbb{E}[1 - I(i - 1 \geq N)],
\end{aligned}$$

where I have used the linearity of expectations under finite sums and assumption (B3), $I(i \leq N) = 1 - I(i - 1 \geq N)$, and assumption (B4). Then applying the logic of equation (B1), we have

$$
\begin{aligned}
&\mathbb{E}[\text{Num observed } (|t_i| \in [e_j, e_{j+1}))] \\
&= \sum_{i=1}^{M} \mathbb{E}[I(t_i \in [e_j, e_{j+1}))\text{obs}(t_i, \theta_i)]\mathbb{E}[1 - I(i-1 \geq N)] \\
&= \mathbb{E}(p(|Z|, \theta_1)||Z| \in [e_j, e_{j+1}))\Pr(|Z| \in [e_j, e_{j+1})) \sum_{i=1}^{M} \mathbb{E}[1 - I(i-1 \geq N)] \\
&= \mathbb{E}(p(|Z|, \theta_1)||Z| \in [e_j, e_{j+1}))\Pr(|Z| \in [e_j, e_{j+1})) \sum_{i=1}^{M} \Pr(i \leq N) \\
&= \mathbb{E}(p(|Z|, \theta_1)||Z| \in [e_j, e_{j+1}))\Pr(|Z| \in [e_j, e_{j+1}))\mathbb{E}(N).
\end{aligned}
$$

Finally, taking the ratio of the above to the same equation for $[e_j, e_{j+1}) = [2, 3)$ and rearranging following equation (B2) yields equation (13). ■

Appendix C: Formal Statement of an Optimal Directed *p*-Hacking Model (Adam (2001))

This section lays out a formal description of an optimal directed p-hacking model (Section I.A.3) and its nesting within the general model (equations (2) to (3)). The model is based on Adam's (2001) generalization of the (Weitzman (1979)) model.[17]

Suppose that the N hacking attempts originate from a sequence of hacking "investigations." Investigation j searches through a set of M factor concepts with t-statistics $\boldsymbol{t}_j \equiv [t_{j,1}, t_{j,2}, \ldots, t_{j,M}]$, which have covariance matrix

$$\text{Cov}(\boldsymbol{t}_j) = \boldsymbol{C}_j. \tag{C1}$$

The M factor concepts also come with a vector $\boldsymbol{\theta}_j \equiv [\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,M}]$, which describe the other qualities of these concepts (e.g., backstory). All factors are false:

$$t_{j,l} \sim N(0, 1), \tag{C2}$$

which also implies that the diagonal terms of $\boldsymbol{C}_j$ are one.

Investigation j proceeds in stages. In the first stage, the hacker has an inherited information set $\mathbb{H}_{j,0} = \{\boldsymbol{C}_j, \boldsymbol{\theta}_j\}$, where for simplicity, I assume that $\boldsymbol{\theta}_j$ is known at the start. At each stage, the hacker selects a factor concept, pays a cost c, and observes the t-statistic. Thus, at stage l, the hacker has information set $\mathbb{H}_{j,l} = (\boldsymbol{C}_j, \boldsymbol{\theta}_j, \{t_{j,k}\}_{k \in S_l})$, where $S_l \subseteq \{1, 2, \ldots, M\}$ are the indexes of the

[17] Most of the directed search literature focuses on equilibrium matching and congestion externalities, and thus assumes perfect information (Montgomery (1991), Burdett, Shi, and Wright (2001)), preventing the modeling of learning. Adam (2001) generalizes the case without learning (Weitzman (1979)) by allowing reservation prices to change along the search process. These models are closely related to multiarmed bandit superprocesses studied by Glazebrook (1979).

factor concepts selected in stages 1 through $l-1$. At any stage, the hacker can stop his search and write one paper based on any of the concepts selected, and receive a corresponding payoff $p_j(|t_{j,k}|, \theta_{j,k})$. For simplicity, assume that each investigation ends with one paper written.

Suppose that in stage l, the best examined payoff is y. Then the expected gain from examining candidate k and ending the search is

$$\begin{aligned} Q(k, y, \mathbb{H}_{j,l}) = y\Pr(p_j(|t_{j,k}|, \theta_{j,k}) \le y | \mathbb{H}_{j,l}) \\ + \mathbb{E}[p_j(|t_{j,k}|, \theta_{j,k}) | p_j(|t_{j,k}|, \theta_{j,k}) > y, \mathbb{H}_{j,l}] - c - y. \end{aligned}$$

Define the reservation price as R_k that satisfies

$$Q(k, R_k, \mathbb{H}_{j,l}) = 0.$$

Then Adam (2001) shows that the optimal decision is to either (i) examine the candidate with the largest R_k if the largest $R_k > y$, or (ii) end the search and receive y. The proof is quite technical, and involves induction on the number of unexamined candidates.

It is very unlikely that finance researchers use optimal search theory to select factor concepts to examine. Fortunately, the analysis that follows depends only on the statistical structure and not the decision process.

First, note that the statistical structure can be nested in equations (2) and (3) by simply stacking the examined t-statistics: $t_1, t_2, t_3, \ldots = \{t_{1,k}\}_{k \in S_1^*}, \{t_{2,k}\}_{k \in S_2^*}$, where S_j^* is the ex-post set of indices corresponding to the candidates examined during investigation j. The observation indicator can be similarly stacked. Thus, Proposition 1 applies, as does the empirical estimation in Section I.C.

Next, to demonstrate clearly the relevance of Proposition 1, let k^* be the index of the observed candidate in investigation j, define $|t_j^*| = |t_{j,k^*}|$, and let $\bar{S}_1, \ldots, \bar{S}_L$ represent all sets of indexes that can be chosen. Then this selected t-statistic satisfies

$$\begin{aligned} \Pr(|t_j^*| > \bar{t}) &\le \Pr\left(\max_{k \in S_j^*} |t_{j,k}| > \bar{t}\right) \\ &= \Pr(\cup_{k \in S_j^*}(|t_{j,k}| > \bar{t})) \\ &= \sum_{l=1}^{L} \Pr\left(\cup_{k \in \bar{S}_l}\left(|t_{j,k}| > \bar{t}\right) | S_j^* = \bar{S}_l\right) \Pr(\bar{S}_l) \\ &\le \sum_{l=1}^{L} \sum_{k \in \bar{S}_l} \Pr(|t_{j,k}| > \bar{t}) \Pr(\bar{S}_l) \\ &= \sum_{l=1}^{L} \#(\bar{S}_l) \Pr(\bar{S}_l) \Pr(|Z| > \bar{t}) \\ &= \mathbb{E}\left[\#\left(S_j^*\right)\right] \Pr(|Z| > \bar{t}). \end{aligned} \tag{C3}$$

I can now apply the same argument as equation (5)

$$\begin{aligned}
\mathbb{E}[\text{Number of observed } (|t_j^*| > \bar{t})] &= \mathbb{E}\left[\sum_{j=1}^{J} I(|t_j^*| > \bar{t})\right] \\
&= \sum_{j=1}^{J} \mathbb{E}\left[I\left(|t_j^*| > \bar{t}\right)\right] \\
&\leq \sum_{j=1}^{J} \mathbb{E}\left[I\left(|t_j^*| > \bar{t}\right)\right] \\
&= \sum_{j=1}^{J} \Pr\left(|t_j^*| > \bar{t}\right) \\
&\leq \sum_{j=1}^{J} \mathbb{E}\left[\#\left(S_j^*\right)\right] \Pr\left(|Z| > \bar{t}\right), \qquad \text{(C4)}
\end{aligned}$$

where the second line uses the linearity of expectations and the last line uses equation (C3). Finally, noting that $\sum_{j=1}^{J} \mathbb{E}[\#(S_j^*)] = \mathbb{E}(N)$ and solving for $\mathbb{E}(N)$ in equation (C4) implies Proposition 1.

Appendix D: Additional Empirical Results

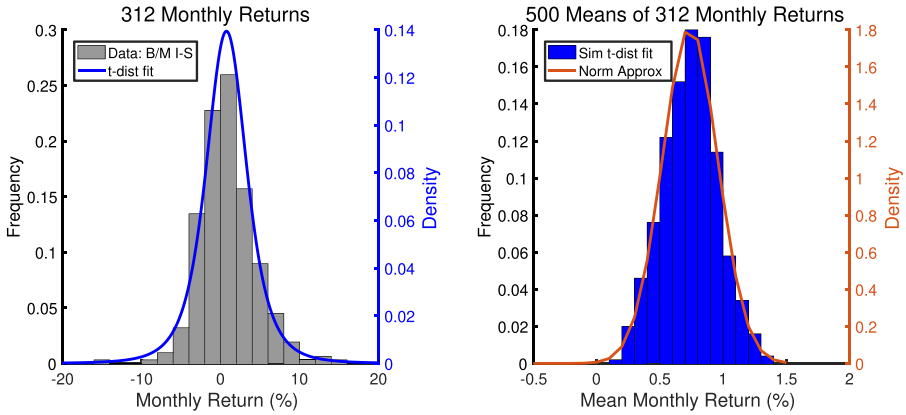

Figure D1. Verifying normal *t*-statistic tails using an estimated model of fat-tailed returns. I fit a scaled t-distribution to monthly returns on the equal-weighted long-short B/M portfolio from the Chen-Zimmermann data (2020a) (left panel). The estimated degrees of freedom parameter is 3.6, consistent with the literature that finds that returns are fat-tailed (Fama (1965)). I simulate 500 samples of 312 monthly returns using the fitted distribution and calculate the sample mean (right panel). Despite fat-tailed returns, the normal distribution fits the distribution of sample mean returns very well. (Color figure can be viewed at wileyonlinelibrary.com)

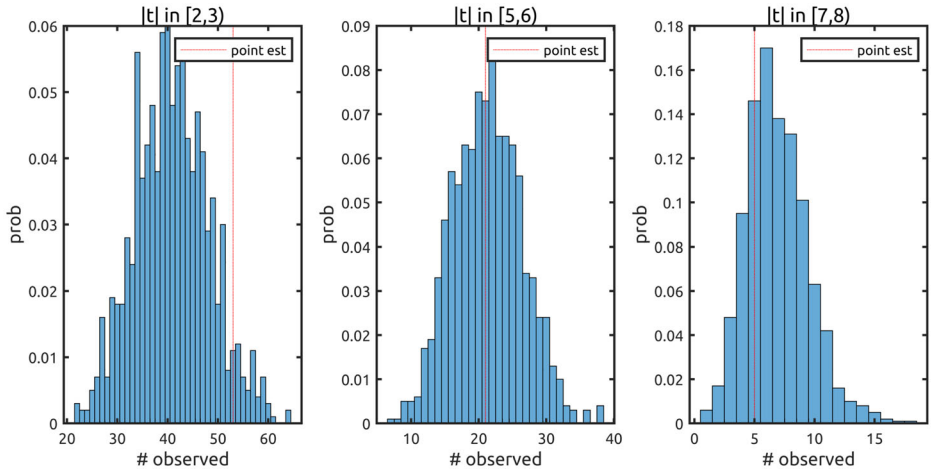

Figure D2. A closer look at the bootstrapped histogram confidence interval. I plot the distribution of bootstrapped histogram counts for selected bins from the data underlying the 90% confidence interval in Table I. (Color figure can be viewed at wileyonlinelibrary.com)

REFERENCES

Adam, Klaus, 2001, Learning while searching for the best alternative, *Journal of Economic Theory* 101, 252–280.

Adler, Robert J., and Jonathan E. Taylor, 2007, Random Fields and Geometry (Springer, New York; London).

Andrews, Isaiah, and Maximilian Kasy, 2019, Identification of and correction for publication bias, *American Economic Review* 109, 2766–2794.

Barberis, Nicholas C., 2018, Psychology-based models of asset prices and trading volume, Working Paper, National Bureau of Economic Research.

Benjamini, Yoav, 2010, Discovering the false discovery rate, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 405–416.

Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57, 289–300.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* 29, 1165–1188.

Bryzgalova, Svetlana, 2015, Spurious factors in linear asset pricing models, Working Paper, London School of Economics.

Burdett, Kenneth, Shouyong Shi, and Randall Wright, 2001, Pricing and matching with frictions, *Journal of Political Economy* 109, 1060–1085.

Chen, Andrew Y., 2019, Do t-stat hurdles need to be raised? Identification of publication bias in the cross-section of stock returns, Working Paper, Federal Reserve Board.

Chen, Andrew Y., and Tom Zimmermann, 2020a, Open source cross-sectional asset pricing, Working Paper, Federal Reserve Board and University of Cologne.

Chen, Andrew Y., and Tom Zimmermann, 2020b, Publication bias and the cross-section of stock returns, *Review of Asset Pricing Studies* 10, 249–289.

Chordia, Tarun, Amit Goyal, and Alessio Saretto, 2020, Anomalies and false rejections, *Review of Financial Studies* 33, 2134–2179.

Clarke, Sandy, and Peter Hall, 2009, Robustness of multiple testing procedures against dependence, *Annals of Statistics* 37, 332–358.

Cochrane, John H., 2017, Macro-finance, *Review of Finance* 21, 945–985.
Cohen, Joel E., 2019, Sum of a random number of correlated random variables that depend on the number of summands, *American Statistician* 73, 56–60.
Efron, Bradley, 2007, Correlation and large-scale simultaneous significance testing, *Journal of the American Statistical Association* 102, 93–103.
Efron, Bradley, 2010, Correlated z-values and the accuracy of large-scale statistical estimates, *Journal of the American Statistical Association* 105, 1042–1055.
Efron, Bradley, 2012, Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Volume 1 (Cambridge University Press, Cambridge).
Efron, Bradley, and Robert Tibshirani, 2002, Empirical Bayes methods and false discovery rates for microarrays, *Genetic Epidemiology* 23, 70–86.
Efron, Bradley, Robert Tibshirani, John D. Storey, and Virginia Tusher, 2001, Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* 96, 1151–1160.
Ellison, Glenn, 2002, Evolving standards for academic publishing: A q-r theory, *Journal of Political Economy* 110, 994–1034.
Fama, Eugene F., 1965, Behavior of stock-market prices, *Journal of Business* 38, 34–105.
Fama, Eugene F., and Kenneth R. French, 2010, Luck versus skill in the cross-section of mutual fund returns, *Journal of Finance* 65, 1915–1947.
Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.
Fan, Jianqing, Xu Han, and Weijie Gu, 2012, Estimating false discovery proportion under arbitrary covariance dependence, *Journal of the American Statistical Association* 107, 1019–1035.
Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *Journal of Finance* 75, 1327–1370.
Fuh, Cheng-Der, 2003, SPRT and CUSUM in hidden Markov models, *Annals of Statistics* 31, 942–977.
Fuh, Cheng-Der, and Tze Leung Lai, 1998, Wald's equations, first passage times and moments of ladder variables in Markov random walks, *Journal of Applied Probability* 35, 566–580.
Glazebrook, Kevin D., 1979, Stoppable families of alternative bandit processes, *Journal of Applied Probability* 16, 843–854.
Green, Jeremiah, John R. M. Hand, and X Frank Zhang, 2017, Characteristics that provide independent information about average U.S. monthly stock returns, *Review of Financial Studies* 4389–4436.
Harvey, Campbell R., and Yan Liu, 2020, False (and missed) discoveries in financial economics, *Journal of Finance* 75, 2503–2553.
Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, …and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.
Holm, Sture, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* 6, 65–70.
Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting anomalies: An investment approach, *Review of Financial Studies* 28, 650–705.
Jacobs, Heiko, and Sebastian Müller, 2020, Anomalies across the globe: Once public, no longer existent?, *Journal of Financial Economics* 135, 213–230.
Kan, Raymond, and Chu Zhang, 1999, Two-pass tests of asset pricing models with useless factors, *Journal of Finance* 54, 203–235.
Kelly, Bryan T., Seth Pruitt, and Yinan Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.
Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Interpreting factor models, *Journal of Finance* 73, 1183–1223.
Lettau, Martin, and Markus Pelger, 2020, Factors that fit the time series and cross-section of stock returns, *Review of Financial Studies* 33, 2274–2325.
Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics* 96, 175–194.

Linnainmaa, Juhani T., and Michael R. Roberts, 2018, History of the cross-section of stock returns, *Review of Financial Studies* 31, 2606–2649.

Lo, Andrew W., and A. Craig MacKinlay, 1990, Data-snooping biases in tests of financial asset pricing models, *Review of Financial Studies* 3, 431–467.

McLean, R. David, and Jeffrey Pontiff, 2016, Does academic research destroy stock return predictability?, *Journal of Finance* 71, 5–32.

Mencken, Henry Louis, 1922, Prejudices: Second Series (Alfred A. Knopf, New York).

Montgomery, James D., 1991, Equilibrium wage dispersion and interindustry wage differentials, *Quarterly Journal of Economics* 106, 163–179.

Moustakides, George V., 1999, Extension of Wald's first lemma to Markov processes, *Journal of Applied Probability* 36, 48–59.

Reiner-Benaim, Anat, 2007, Fdr control by the bh procedure for two-sided correlated tests with implications to gene expression data analysis, *Biometrical Journal* 49, 107–126.

Shanken, Jay, 1992, On the estimation of beta-pricing models, *Review of Financial Studies* 5, 1–33.

Stambaugh, Robert F., and Yu Yuan, 2016, Mispricing factors, *Review of Financial Studies* 30, 1270–1315.

Storey, John D., Jonathan E. Taylor, and David Siegmund, 2004, Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66, 187–205.

Sullivan, Ryan, Allan Timmermann, and Halbert White, 1999, Data-snooping, technical trading rule performance, and the bootstrap, *Journal of Finance* 54, 1647–1691.

Train, Kenneth E., 2009, Discrete Choice Methods with Simulation (Cambridge University Press, Cambridge).

Wald, Abraham, 1945, Sequential tests of statistical hypotheses, *Annals of Mathematical Statistics* 16, 117–186.

Weitzman, Martin L., 1979, Optimal search for the best alternative, *Econometrica: Journal of the Econometric Society* 641–654.

Yan, Xuemin Sterling, and Lingling Zheng, 2017, Fundamental analysis and the cross-section of stock returns: A data-mining approach, *Review of Financial Studies* 30, 1382–1423.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication Code.