Do t-stat Hurdles Need to be Raised? Identification of Publication Bias in the Cross-Section of Stock Returns

Andrew Y. Chen

Federal Reserve Board

andrew.y.chen@frb.gov

July 2019*

Abstract

Estimates of statistical thresholds that control for multiple hypothesis testing among academic publications are weakly identified. Identification is weak because of selective reporting: tests that are close to the null are unlikely to be published. Thus, the proportion of nulls must be extrapolated, and different extrapolations lead to the same observable data. In contrast, adjustments to estimated magnitudes are strongly identified because they are determined by the dispersion of magnitudes, which requires less extrapolation. I demonstrate these results in a dataset of 155 published cross-sectional stock return predictors by studying the bootstrapped distribution of multiple testing statistics under a wide variety of modeling assumptions. t-stat hurdles that control the false discovery rate have large standard errors and are uninformative. Adjustments to expected returns are strongly identified and imply that publication bias for cross-sectional predictors is modest, consistent with McLean and Pontiff (2016).

* First posted to SSRN: September 25, 2018. I thank Preston Harry and Jack McCoy for excellent research assistance, Rebecca Wasyk for excellent scientific programming, and Dino Palazzo and Fabian Winkler for many valuable discussions. I also thank Christine Dobridge, Cam Harvey, Laura Liu, Alan Moreira (WFA discussant), Mihail Velikov, and seminar participants at the Federal Reserve Board and 2019 Western Finance Association Meetings for helpful comments. 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.

1. Introduction

Any empirical literature, when taken as a whole, runs into the multiple testing problem. For example, consider the literature on cross-sectional stock return predictability. More than 100 papers have been published on this topic, and more than 150 hypothesis tests appear in these papers (Chen and Zimmermann 2018). When considered together, this multitude of tests implies that the traditional p-value hurdle of 0.05 (or the t-stat hurdle of 1.96) is invalid, as this hurdle uses the assumption of a single test (Fisher 1925, Harvey, Liu, and Zhu 2016, Benjamin et al. 2018). Estimates of the magnitude of predictability are also invalid, as individual estimates do not account for information in the 150+ estimates examined using the same data source. Indeed, the published tests are just a subset of all tests, as insignificant tests are unlikely to be reported (Fanelli 2010). I use the term "publication bias" to refer to these related problems of multiple testing and selective reporting.

Statistical inference under multiple testing is a well-studied topic that has lead to multiple textbooks (Hsu 1996, Efron 2012). Textbook multiple testing adjustments, however, assume an unbiased sample of hypothesis tests (the Benjamini and Hochberg 1995 algorithm, for example), and do account for the selective reporting found in academic publications. Thus, to apply textbook adjustments to published findings, unreported tests must be modeled, leading to questions of identification: Do the reported tests provide enough information to pin down the adjustments? Or do alternative models, with wildly different implications, lead to the same reported data? How large are the standard errors? Which adjustments are strongly identified, and what do they tell us?[1]

This paper studies the identification of multiple testing statistics in a large collection of hypothesis tests published in academic journals. I estimate models of publication bias (à la Harvey, Liu, and Zhu 2016 and Andrews and Kasy Forthcoming) on test statistics built from a dataset of 155 published cross-sectional stock return predictors (from Chen and Zimmermann 2018). The estimated models allow me to calculate textbook adjustments to t-stat hurdles and predictability magnitudes that account for publication bias. To study identification, I bootstrap my estimates, examine a wide variety of modeling assumptions, and illustrate identification with simple examples. Though my data consists of published stock return predictors, my results come from basic properties of

[1] Andrews and Kasy (Forthcoming) provide a formal theory of identification, but do not address these practical questions.

the data that are likely to be found in other literatures.

I find two main results. The first is that adjustments to t-stat hurdles that account for publication bias are weakly identified and uninformative. To calculate t-stat hurdles, I estimate the Benjamini and Hochberg (1995) false discovery rate (the expected proportion of discoveries that are false) as a function of the t-stat hurdle, and then solve for the hurdle that produces a false discovery rate of 5% or 1% (following Harvey, Liu, and Zhu 2016). Standard errors on the adjusted hurdles are large, at about 1.0. Moreover, the classical hurdles of 1.96 and 2.58 (for two-sided 5% and 1% significance) are well-within one standard error of their multiple testing counterparts. Smaller standard errors are found using alternative models, but this result is highly sensitive to assumptions, further illustrating weak identification. Indeed, of the 10 alternative assumptions I examine, only 2 produce standard errors smaller than 0.70. Overall, the data simply say little about the question of whether t-stat hurdles should be raised to account for publication bias.

My second main result is that adjustments to estimated magnitudes are strongly identified. I measure the magnitude of a predictor using the expected return of a long-short portfolio formed on the predictor. Publication-bias adjusted expected returns are only 16% smaller than simple sample mean returns. More importantly, the standard error is small, at only 3 percentage points, and I find a similar point estimate and standard error using the 10 alternative modeling assumptions.[2] These results show that the data is quite informative about how much estimated magnitudes need to be adjusted, despite the fact that t-stat hurdles are weakly identified.

Though my main results use data from publications on cross-sectional predictability, they come from the fact that the literature shows a strong preference for statistically significant results. This preference is seen in most literatures across economics, psychology, and the social sciences (Fanelli 2010), and thus my main results are likely to be found elsewhere.

A strong preference for significant results implies that published data are uninformative about adjustments to null hypothesis testing. To understand this, it helps to know that textbook adjustments to hypothesis testing center around the proportion of tests that are null. This null proportion, in turn, is estimated using the density of tests near the null (Benjamini and Hochberg 2000, Storey and Tibshirani 2003). Intuitively, a high density near the null implies that a large proportion of tests are null. But since academic

[2]These parametric estimates are also consistent with the upper bound of 26% found by McLean and Pontiff (2016) using a model-free out-of-sample test.

literatures strongly prefer significant results, the density near the null is unobserved. Thus, this density needs to be extrapolated, and uncertainty about this extrapolation leads to uncertainty about adjusted t-stat hurdles. Indeed, as publication selects for t-stats that are more than 2 standard errors from the null, uncertainty about this extrapolation is considerable.

The preference for significant results also implies that adjustments to estimated magnitudes are strongly identified. Textbook adjustments like James and Stein (1961) shrinkage use the dispersion of magnitudes to pin down this adjustment.[3] Intuitively, a high dispersion means that the published t-stat contains signal about the unobserved magnitude, as found in a Bayesian updating problem. The dispersion of magnitudes is indirectly observed in the right tail of t-stats, as the t-stat distribution picks up this underlying dispersion. Finally, the right tail of t-stats is well-observed as long as significant results are likely to be published, as appears to be the case in most literatures across the social sciences.

The idea that t-hurdles do not need to be raised in the presence of publication bias is highly counterintuitive. The traditional logic of multiple testing argues that hurdles must be raised: running multiple tests can lead to lucky results, and thus raising the hurdle is required to control for this luck. This logic is embedded in formal statistics such as the seminal Benjamini and Hochberg (1995) algorithm for estimating an upper bound on the false discovery rate.[4]

There is another logic, however, that is also embedded in multiple testing statistics. The many tests run by academics provide more information than a single test. In particular, many tests allow inference about the proportion of tests that are truly null. If the proportion of nulls is close to zero, then multiple testing statistics imply that statistical standards can be *lowered* relative to the classical single test setting. Such a conclusion is impossible to obtain with a single test, but they are possible with data on multiple tests using Benjamini and Hochberg's (2000) followup to their 1995 paper, and many statisticians have continued to develop on their methodology (Storey and Tibshirani 2003, Genovese and Wasserman 2004, Efron 2004, among others). This logic of "learning from

[3]James-Stein shrinkage is the focus of Chapter 1 of Efron's (2012) textbook. The Benjamini and Hochberg (1995) false discovery rate, on which my t-stat hurdles are based (see also Harvey, Liu, and Zhu 2016) is the focus of Chapters 2 and 4.

[4]The Benjamini and Hochberg (1995) paper contains both the definition of the false discovery rate and an algorithm for estimating an upper bound. The false discovery rate in general allows for the t-hurdle to be lowered, but the algorithm does not.

the experience of others" is a common theme in Efron's (2012) textbook on large scale hypothesis testing.

I illustrate this multiple testing intuition with a few simple examples. These examples use exclusively well-established methods from the statistics literature (Benjamini and Hochberg 1995, 2000), or from the influential Harvey, Liu, and Zhu (2016) paper, and should be free of uncertainty about my data, model, and statistical methods.

My findings address only one motivation for raising statistical standards: the application of frequentist multiple testing statistics to published results. Raising standards, however, can also be motivated using Bayesian arguments (Johnson 2013, Harvey 2017, Benjamin et al. 2018).[5] Moreover, there may be compelling reasons to raise economic standards. For example, many papers find that cross-sectional stock return predictability is not robust to transaction costs (Stoll and Whaley 1983; Schultz 1983; Korajczyk and Sadka 2004; Lesmond, Schill, and Zhou 2004; Novy-Marx and Velikov 2016, among others). Indeed, Chen and Velikov (2018) find that the post-publication returns for 120 predictors are negligible after transaction costs.[6] The fact that predictability is not robust to transaction costs and post-publication decay is helpful for understanding the small publication bias estimated by this and other papers (McLean and Pontiff (2016), Chen and Zimmermann 2018): it is not difficult to believe in stock market predictability that is both temporary and costly to implement.

More broadly, the literature on publication bias is large. Christensen and Miguel (2018) provide a recent review. Within this literature, my approach is most similar to Harvey, Liu, and Zhu (2016), Andrews and Kasy (Forthcoming), and Chen and Zimmermann (2018), who also estimate flexible models of publication bias using generalized method of moments or maximum likelihood. This approach allows for direct estimates of publication bias adjustments under selective publication, which is not possible using the more common meta-regression approach (Card and Krueger 1995, Egger, Smith, Schneider, and Minder 1997), or studies of the distribution of p-values (De Long and Lang 1992, Brodeur, Lé, Sangnier, and Zylberberg 2016). Please see Andrews and Kasy (Forthcoming) for a detailed comparison of these methods. Unlike Harvey, Liu, and Zhu (2016) and Andrews and Kasy (Forthcoming), I show how selective publication implies

[5] For a critique of Bayesian and other arguments, please see Amrhein and Greenland (2018) and McShane et al. (2019).

[6] In line with the microstructure studies, many papers find that predictors are not robust to value-weighting or the exclusion of microcaps (Hou, Xue, and Zhang 2017; Green, Hand, and Zhang 2017; Jacobs and Müller 2017a).

that some publication bias adjustments are more strongly identified than others. Indeed, selective publication implies that frequentist multiple testing statistics can say little about the recent debate raising standards of statistical significance (Harvey, Liu, and Zhu 2016, Benjamin et al. 2018).

Concerns about multiple testing in asset pricing go back at least to Jensen and Bennington (1970) and Merton (1987). Formal empirical studies include Sullivan, Timmermann, and White (1999), Yan and Zheng (2017), Chordia, Goyal, and Saretto (2017), and Harvey and Liu (2018). These papers, however, do not use data on published hypothesis tests, and thus only provide indirect evidence about publication bias.

Studies that use published asset pricing tests have yet to come to consensus. Harvey, Liu, and Zhu (2016) and Linnainmaa and Roberts (2018) find that most published results are mostly false, while McLean and Pontiff (2016), Jacobs and Müller (2017b), and Chen and Zimmermann (2018) find the opposite. My results help reconcile this conflict by estimating t-hurdles and adjusted returns in the same framework. Moreover, my paper is the first to demonstrate the identification challenges that come from selective publication.

The Need for Parametric Estimates Much of the multiple testing literature in statistics is centered around non-parametric algorithms like Benjamini and Hochberg (1995). Related non-parametric tests are used in Harvey, Liu, and Zhu (2016) (HLZ), and Harvey and Liu (2018) propose non-parametric bootstrap methods for studying multiple testing in asset pricing. Why don't I avoid the complications of modeling and simply use these non-parametric tests?

I do not use non-parametric tests from statistics because they are uninformative or infeasible when data is selected, as is the case with publication bias.[7] To understand this, it helps to examine the Benjamini and Yekutieli (2001) algorithm:

1. Let $\text{pval}_{(1)} \leq \text{pval}_{(2)} \leq ... \leq \text{pval}_{(N_{\text{all}})}$, be the ordered p-values, where N_{all} is the number of tests in the family of tests under consideration.

[7] McLean and Pontiff (2016), Jacobs and Müller (2017b) and Linnainmaa and Roberts (2018) provide informative and feasible non-parametric tests for publication bias that are specific to asset pricing.

2. Define

$$i^{\star} = \max\left\{ i : p_{(i)} \leq \frac{(i)}{N_{\text{all}}} \left(\sum_{j=1}^{N_{\text{all}}} \frac{1}{j} \right)^{-1} \alpha \right\} \tag{1}$$

where α is the desired upper bound on the FDR.

3. Reject the hypotheses corresponding to $\text{pval}_{(1)}$, $\text{pval}_{(2)}$, ...,$\text{pval}_{(i^{\star})}$.

Benjamini and Yekuteili show that under this procedure, FDR $\leq \alpha$ for a broad class of tests. Algorithms like Benjamini and Yekutieli (2001) have been widely used in multiple testing studies in genomics, where both (i) and N_{all} are known (Efron 2012).

Publication bias, however, means that neither the ranking of p-values (i) nor the total number of tests N_{all} is observed. Indeed, this is the essence of publication bias: the publication process favors small p-values, and thus tests with large p-values tend to be unobserved.

Adding structure and extrapolating the large p-values is one way to address this selective reporting problem. This approach is used in HLZ's model with correlations, Chen and Zimmermann (2018), and is also the method I use. Extrapolation, however runs into questions of identification. Addressing these questions is precisely the goal of this paper.

In principle, non-parametric identification of (i) and N_{all} is possible, as formally shown in the recent paper by Andrews and Kasy (Forthcoming) (Proposition 3, Nonparametric identification using meta-studies). Sample sizes for published data is limited, however, which makes non-parametric estimation impractical. Thus, even Andrews and Kasy assume parametric models to estimate publication bias using GMM and maximum likelihood, as I do in this paper.

Bootstrapping random empirical tests à la Harvey and Liu (2018) to approximate (i) and N_{all} is not feasible either, because the selection of published data is complex.[8] Publishing a predictor in a respected journal typically requires a large t-stat, but it also requires additional items: supplemental evidence, robustness checks, and economic or psychological motivations, at least most of the time.[9] To properly bootstrap a model of publication, one must have a process for generating all of these additional items from

[8] Harvey and Liu (2018) acknowledge that their non-parametric approach does not address publication bias, and refer the reader to the parametric model of publication in Harvey, Liu, and Zhu (2016).

[9] Harvey (2017) provides a few counterexamples of published predictors that are arguably poorly motivated.

empirical data, as well as a method for estimating this complicated selection process. Such a complicated bootstrap is infeasible, or at least much more difficult than estimating a model like HLZ's model with correlations.

It's worth noting that the Benjamini and Yekutieli (2001) and related adjustments are not free of distributional assumptions. One still needs to assume a null distribution in order to calculate p-values. In cross-sectional asset pricing, the assumed null is almost always normal—the same assumption I use (also the assumption in Harvey, Liu, and Zhu 2016).

Finally, simple non-parametric estimates focus on null hypothesis testing, and thus become mired in its associated controversies. The classic controversy of the appropriate significance level (revived recently by Benjamin et al. 2018) is magnified in a multiple-test setting. Harvey, Liu, and Zhu (2016) suggest an FDR $\leq$ 1%, but a far more generous FDR $\leq$ 10% is popular in the statistics literature (Efron 2012). Indeed, McShane et al. (2019) argue that null hypothesis testing should be abandoned altogether.

These considerations lead me to estimate parametric models of publication bias. The estimations lead to informative inferences about the FDR. They also provide multiple testing adjustments for expected returns, for the readers who wish to move beyond null hypothesis testing. Parametric models come with concerns about assumptions. To alleviate these concerns, I estimate a wide variety of models in Section 4.

2. How Multiple Testing Does Not Necessarily Imply Higher t-hurdles

The idea that multiple testing does not necessarily imply higher t-hurdles is counter-intuitive, perhaps even paradoxical. Thus, before I show my main results, I illustrate this possibility using two simple examples.

2.1. Example 1: Multiple Testing on Height Measurements

The standing height of an adult human is measured with an error of about 1 inch (2.5 cm, Mikula, Hetzel, Binkley, and Anderson 2016). Thus, the true height of any individual cannot be known from a single measurement, and conducting hypothesis tests on many measured heights leads to the multiple testing problem.

To formalize these issues, assume that measurement error is normally distributed, and consider the null hypothesis that an individual is 70 inches (5'10" or 178 cm) tall. 70 inches is the average height of an adult American male. Under these assumptions, the classical 5% test for the null H_0 is

$$\text{Reject } H_0 \text{ if t-stat} \equiv \frac{|\text{height} - 70|}{1} > 1.96 \tag{2}$$

In other words: classical single testing statistics imply that a person is significantly taller than 70 inches if he is 72 inches or taller.

Now suppose we want to test many hypotheses. Basketball-reference.com provides data on many measured heights. Panel A of Figure 1 shows the distribution of 482 height measurements from this website (bars).[10] It also shows the distribution implied by the null of $N(70,1)$ (line).

How do we test the many hypotheses: person i is null (70 inches) for i = 1, 2, ..., 482? The 482 hypotheses imply that the single testing rule (Equation (2)) is invalid. The seminal Benjamini and Hochberg (1995) (BH 1995) paper, however, shows how one can control for multiple tests. BH 1995 presents an algorithm that is represented graphically in Panel B of Figure 1. The algorithm goes as follows: (1) plot p-values as a function of their ranking (circle markers), (2) draw a p-value boundary (dashed line) following the equation

$$\text{p-value boundary} = \frac{5\%}{482}\left[\text{p-value rank}\right], \tag{3}$$

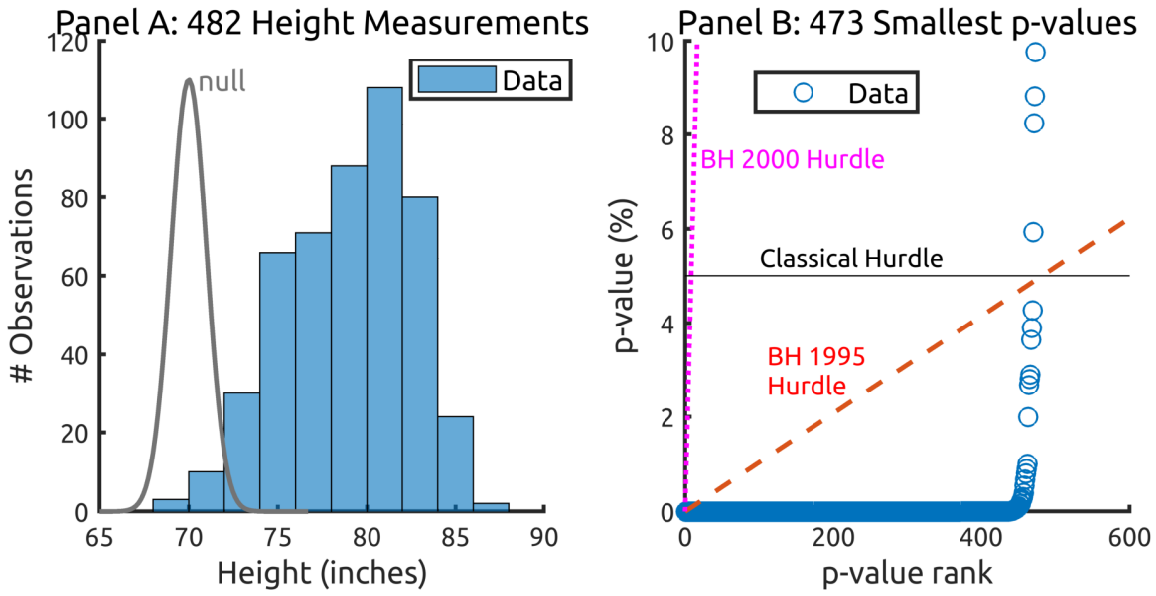
(3) find the last p-value below the boundary, and (4) reject that last p-value and all p-values to the left.[11] Benjamini and Hochberg (1995) prove that this algorithm ensures

[10] The raw data is rounded to the nearest inch, so I add $N(0.0.5)$ noise to remove this discretization.

[11] Formally, the BH 1995 algorithm is:

1. Let $\text{pval}_{(1)} \leq \text{pval}_{(2)} \leq ... \leq \text{pval}_{(N_{\text{all}})}$, be the ordered p-values, where N_{all} is the number of tests in the family of tests under consideration.
2. Define
$$i^\star = \max\left\{ i : p_{(i)} \leq \frac{(i)}{N_{\text{all}}}\alpha \right\} \tag{4}$$
where α is the desired upper bound on the FDR and N_{all} is the number of tests
3. Reject the hypotheses corresponding to $\text{pval}_{(1)}$, $\text{pval}_{(2)}$, ...,$\text{pval}_{(i^\star)}$.

Figure 1: Multiple Testing Does Not Necessarily Imply Higher t-hurdles: Height Measurement Data. Panel A shows the distribution of 482 measured heights from basketball-reference.com. The line shows the distribution implied by the null height of 70 inches (the average American male height), assuming normal errors and a standard error of 1 inch (Mikula et al, 2016). Panel B examines many hypotheses of the form: person i is 70 inches. The solid line corresponds to the naive single testing rule (5% significance, Equation (2)), the dashed line to the Benjamini-Hochberg 1995 rule (FDR ≤ 5%, Equation (3)) and the dotted line to the Benjamini-Hochberg 2000 rule (FDR ≤ 5%, Equation (6)). Each rule rejects p-values below the corresponding line.



that the false discovery rate (FDR) satisfies

$$\text{FDR} \leq p_0\, 5\% \tag{5}$$

where p_0 is the proportion of tests that are null (see Theorem 2 in Benjamini and Hochberg 1995). As $p_0 \leq 1$, this rule also ensures that FDR ≤ 5%.

Panel B shows that, according to BH-1995, no adjustment is needed. The set of tests rejected by BH 1995 (dashed line) is exactly the same as the set rejected by the classical hurdle (solid line). Equivalently, the t-hurdle of 1.96 is sufficient to ensure the FDR ≤ 5%.

Indeed, there is a clear argument for *lowering* the t-hurdle for this dataset. Panel A shows that the measured heights in this dataset are very tall. Only 3% of measurements are within 2 inches of the null of 70 inches. This suggests that the proportion of nulls p_0

is quite small, and thus the BH 1995 adjustment is too strict.

The original draft of Benjamini and Hochberg (1995), written in 1989, recommends *estimating* p_0 (Benjamini 2010). This idea was eventually published in Benjamini and Hochberg (2000). Since then, many statisticians have since followed up on the idea (Storey and Tibshirani 2003; Efron 2004; Genovese and Wasserman 2004; for example).

Specifically, the Benjamini and Hochberg (2000) (BH 2000) algorithm recommends using the following p-value boundary:

$$\text{p-value boundary} = \frac{5\%}{\hat{p}_0 482}\left[\text{p-value rank}\right] \tag{6}$$

where $\hat{p}_0$ is a simple regression estimate of p_0. As proved in BH 1995, the BH 2000 rule results in an FDR $\leq 5\%$. Using the simple regression estimate, I obtain $\hat{p}_0 = 1.7\%$, similar to the 1.5% of measurements that are within 1.5 inches of the null. Details of this estimate are found in Appendix A.1

The dotted line of Panel B (Figure 1) illustrates the BH 2000 algorithm. As $\hat{p}_0$ is very small, the algorithm implies a very large adjustment to the BH 1995 slope. In fact, the slope is so large that *all* hypotheses are rejected. In other words, BH 2000 says that the t-hurdle can be lowered, all the way to 0.

Hypothesis testing with a t-hurdle of 0 may sound absurd. This hurdle implies that no height measurement is needed to reject the null (with an FDR of 5%). How can no measurement be needed?

No measurement is needed because of the nature of the data and null hypothesis. The height measurements at basketball-reference.com come from NBA (professional basketball) players. NBA players are extremely tall, as seen in the 482 hypothesis tests in Figure 1. Indeed, these players are so tall, that one can be 95% sure that a random player is not the null of 70 inches (the average American height), as long as that player is in the NBA.

This example illustrates how the adjusted statistical hurdle depends on of the properties of the dataset and the null under consideration. Indeed, if the data is very far from the null, the adjusted hurdle may be less strict than the single-testing benchmark. This problem of an implausible null hypothesis is one reason why McShane et al. (2019) argue for abandoning null hypothesis testing.

2.2. Example 2: Harvey, Liu, and Zhu's (2016) Asset Pricing Factors

The height measurement data illustrate how t-stat hurdles do not necessarily need to be raised, but they cannot address selective reporting and identification. Thus, to illustrate these more subtle issues, I study Harvey, Liu, and Zhu's (2016) (HLZ's) factor data and their model with correlations.

Unlike the NBA height data, the HLZ data exhibits selective reporting. Selective reporting is found in, for example, venture capital returns that are only observed if new financing is obtained (Cochrane 2005). Published hypothesis tests also exhibit selective reporting, as insignificant tests are likely to be unreported (Fanelli 2010). These reporting biases are not present in the NBA player data: all NBA players are reported at basketball-reference.com, and their heights are typically measured after the players are already selected to the NBA.[12]

Table 1 reprints moments of reported t-stats from asset pricing factor tests in HLZ (in the row "adjusted data"). HLZ adjust their factor data to account for the fact that marginal t-stats are less likely to be reported (see Table caption). This adjustment implies a slight leftward shift for the left shoulder of the t-stat distribution, but the adjustment is small overall.

Table 1: Moments of t-stats from the Harvey, Liu, and Zhu (2016) Factor Data and their Model with Correlations

"Adjusted Data" reprints the moments found in Harvey, Liu, and Zhu (2016) (HLZ) page 28. HLZ adjust the data by removing t-stats < 1.96 and duplicating the t-stats between 1.96 and 2.57 before calculating moments. "HLZ baseline" is the baseline estimate of Harvey, Liu, and Zhu (2016) (their Table 5, 2nd row). "Alternative" changes N_{all} to 900, p_0 to 0, and λ to 0.505%.

		Observed t-stat Percentiles		
	# of t-stats	20	50	90
Adjusted Data	353	2.39	3.16	6.34
HLZ baseline	334	2.40	3.37	6.62
Alternative	349	2.35	3.27	6.14

The reported t-stats in Table 1 are large. A naive observer may get the impression that the asset pricing factors are similar to NBA heights: 80% of t-stats exceed 2.39, suggesting

[12] There are subtle biases in the NBA data (Herring 2016) but these biases do not lead to unreported data.

that the Benjamini and Hochberg (2000) algorithm would imply a very low t-stat hurdle. But the moments in Table 1 omit factor tests with low t-stats, and thus it is unclear what Benjamini and Hochberg (2000) would imply for the complete data.

Moreover, factor t-stats are likely to be correlated. Many risk factors move with the business cycle and use the same U.S. panel data. In contrast, the height measurements are likely to be independent. This potential for correlated t-stats also implies that Benjamini and Hochberg (2000) cannot be directly applied.

The HLZ Baseline Model To account for these issues, HLZ model both selective reporting and correlation. I summarize their model below:

$$\mu_i \sim \begin{cases} \delta(0) & \text{with prob } p_0 \\ \text{Exp}(\lambda) & \text{otherwise} \end{cases} \tag{7}$$

$$\text{Var}(r_{i,t}) = \sigma^2, \tag{8}$$

$$\text{Cov}(r_{i,t}, r_{j,t}) = \rho\sigma^2, \quad i \neq j, \tag{9}$$

$$\text{Cov}(r_{i,t}, r_{j,s}) = 0, \quad i \neq j. \tag{10}$$

$$t_i \equiv \frac{\sum_{t=1}^{T} r_{i,t}/T}{\sigma/\sqrt{T}} \sim N(\mu_i/(\sigma/\sqrt{T})), 1) \tag{11}$$

$$t_i \text{ is observed if } t_i \geq t_{\min} \tag{12}$$

where

- $i = 1, ..., N_{\text{all}}$ index the factors in the family of tests
- μ_i is the latent expected return for factor i in this family
- $\delta(0)$ is the distribution with a point mass at 0
- $\text{Exp}(\lambda)$ is the exponential distribution with standard deviation λ
- p_0 is the probability of drawing a predictor from $\delta(0)$
- $r_{i,t}$ is the return for factor i in month t
- σ is the volatility of $r_{i,t}$ (constant for all i)
- t_i is the t-stat for factor i
- T is the number of months in each test sample (constant for all i)

- ρ is the pairwise time-series correlation between any two factor returns
- $t_{\min}$ is the minimum t-stat for observation

In short, factor i is false (null) with probability p_0, but we only observe factor i if $t_i \geq t_{\min}$. Moreover, these observed t_i are noisy and correlated proxies for the true expected return μ_i. Explicitly modeling this selection and correlation allows HLZ to compute valid FDRs. As shown by Efron and Tibshirani (2002), this "two groups" framework provides an elegant Bayesian interpretation of the Benjamini and Hochberg (1995) FDR. All of these assumptions can be well-justified, but for conciseness I ask the reader to please see Harvey, Liu, and Zhu (2016) for the justifications.

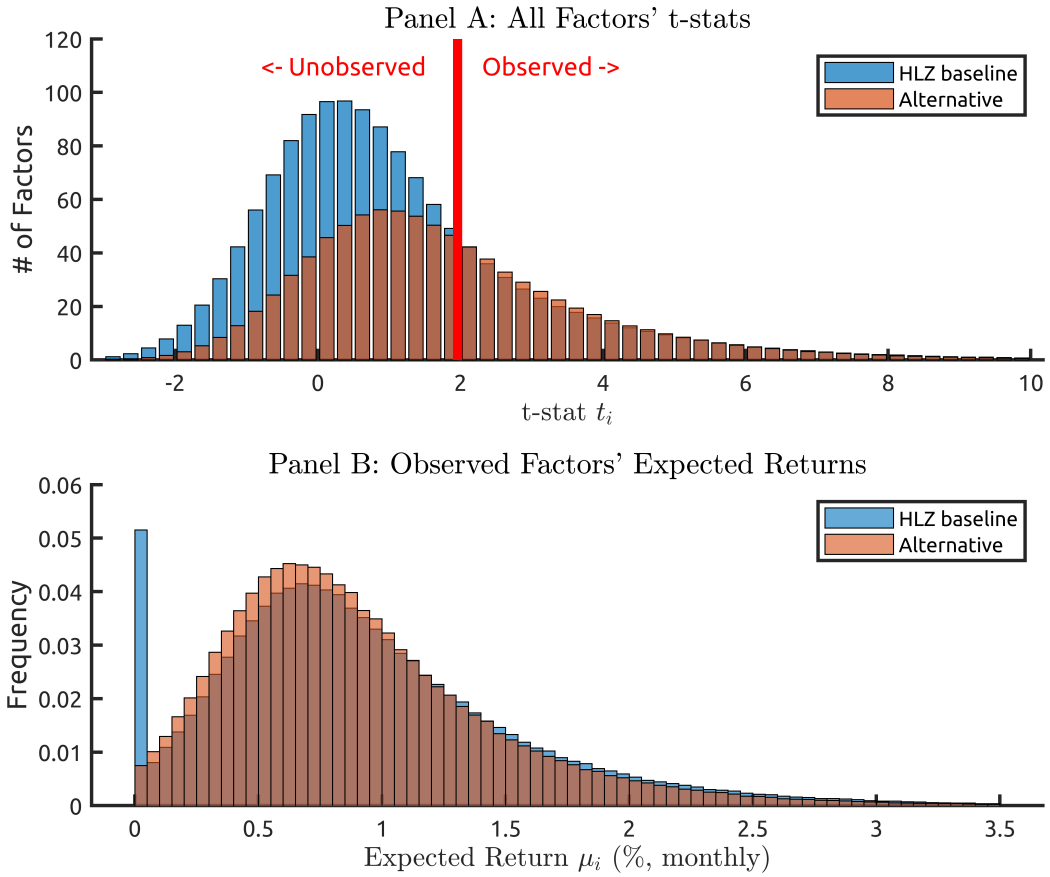
HLZ estimate Equations (7)-(12) using SMM, and arrive at the following parameter values: $\rho = 0.2$, $\sigma = 15/12\%$, $T = 240$ months, $N_{\text{all}} = 1378$, $p_0 = 0.444$, and $\lambda = 0.555\%$. I simulate their estimated model and find their model fits the data well (Table 1). The HLZ baseline predicts the 20th, 50th, and 90th percentiles of published t-stats are 2.40, 3.37, and 6.62, respectively. These values are close to the data values of 2.39, 3.16, and 6.34.

The HLZ estimate implies that raising t-hurdles is necessary. According to HLZ's Table 5, the classical t-hurdle of 1.96 needs to be raised to 2.27 to control the FDR at 5%, and a t-hurdle of 2.95 is required to control the FDR at 1%.

An Alternative, Observationally Equivalent Model Now consider an alternative parameterization. Begin with HLZ's baseline estimate, but change N_{all} to 900, p_0 to 0, and λ to 0.505%. Calculating the FDR for this alternative model is simple: there are no null factors ($p_0 = 0$), and thus the FDR is zero, regardless of the t-hurdle. Thus, the t-hurdle to control the FDR at any level is 0. Effectively, this model assumes that academic factors are like the NBA players: academic factors are all extraordinary, and unlike the null.

Can the data reject this idea that all academic factors are all extraordinary? Figure 2 suggests that the answer is no. Panel A plots the distribution of all t-stats implied by the HLZ estimate and the alternative model. Though the two distributions differ to the left of 1.96 (vertical line), they are nearly identical to the right of the line, and only t-stats to the right of the line are observed. Intuitively, Panel A shows that estimating p_0 requires extrapolating from the observable distribution. There are multiple ways to do this extrapolation, with starkly different implications for t-hurdles.

Figure 2: Multiple Testing Does Not Necessarily Imply Higher t-hurdles: Harvey, Liu, and Zhu's (2016) Model with Correlations. HLZ baseline is the baseline estimate of Harvey, Liu, and Zhu (2016) (their Table 5, 2nd row), which implies that t-hurdles should be raised above 1.96 to ensure FDR ≤ 5%. The alternative model changes three parameter values, implies no false discoveries ($p_0 = 0$), and thus a lowering of the t-hurdle to 0. Panel A shows the distribution of all t-stats, including unobserved. The vertical line is the cutoff for observability. Panel B shows the distribution of expected returns μ_i for observed factors.



Consistent Implications Across the Two Models Despite their different t-hurdles, the models have similar implications for adjusted magnitudes. In this setting magnitudes are described by the expected return of the factor μ_i. This similarity is shown in Panel B of Figure 2, which plots the distribution of the expected returns μ_i among observed factors.

The HLZ baseline features a spike near 0 that is missing in the alternative model, and the HLZ baseline has a slightly fatter right tail. But overall, the distributions are very

similar. Both models imply that the bulk of the distribution of μ_i resides between 0.50% and 1.00% per month.

Indeed, summary statistics for observed μ_i are very similar across models. The mean μ_i is 0.93% in the HLZ baseline, just a touch above the 0.91% in the alternative model. Similarly, the median observed μ_i are 0.82% and 0.81% in the two models, only a bit smaller than the median observed sample mean return of 0.88%.[13]

This similarity originates from the fact that the two parameterizations use similar values for the dispersion of non-null expected returns λ. HLZ's baseline estimate of $\hat{\lambda} = 0.555\%$ is not far from the alternative parameterization of $\lambda = 0.505\%$. This large dispersion is essential, as the right tail of observed t-stats is very fat (Table 1, Panel A). The tail of a t-distribution with many degrees of freedom is too thin to account for this observed fat right tail, and significant dispersion in expected returns (large λ) is required to fit the data. This identification is studied in detail for cross-sectional predictors in Chen and Zimmermann (2018) (see also Chen 2019).

The HLZ factor data illustrates the importance of identification for publication bias. Publication bias is distinct from the textbook multiple testing setting because of selective reporting: small t-stats are unreported. Thus, one must extrapolate in order to calculate adjustments. Some adjustments are more sensitive to this extrapolation than others. A clear understanding of this issue requires a formal estimate of sampling variation. In the next section, I undertake this formal estimate.

3. Main Results: Identification of Publication Bias for Cross-Sectional Predictors

My main analysis follows Harvey, Liu, and Zhu (2016) closely. Like HLZ, I build on their model with correlations, estimate the model using SMM, and calculate multiple testing adjustments using the estimated model.

I differ from HLZ because of my focus on identification. As shown by HLZ, identification of t-hurdles is not possible without information about the correlation between t-stats. The hand-collected t-stats from HLZ, however, do not provide correlation information. Thus, I study the Chen and Zimmermann (2018) (CZ) datset, which provides

[13] To calculate the median sample mean return in HLZ's data, simply multiple the median t-stat of 3.16 with the assumed monthly standard error of $15/\sqrt{(12)(240)}\%$.

the entire time series of using replicated predictors (Section 3.1).

Examination of the CZ data shows that the dispersion of correlations is non-negligible. Such dispersion cannot be accounted for in HLZ's model with a constant correlation, and thus my baseline model differs by allowing for heterogenous correlations (Section 3.2). A model that nests the HLZ model leads to similar results, however (Section 4.1).

I also differ from HLZ by estimating standard errors, examining adjustments to estimated magnitudes, and analyzing many alternative specifications (Sections 3.3, 3.5, and 4). These extensions are critical for understanding identification of different multiple testing statistics.

3.1. Data on Published Cross-Sectional Predictors

My data consists of 155 published cross-sectional predictors from the Chen and Zimmermann (2018) (CZ) dataset. For each predictor, I measure the magnitude of predictability using the sample mean return for a long-short portfolio formed by sorting stocks on the predictor, following McLean and Pontiff (2016). For most predictors, this implies equal-weighting, as the vast majority of papers show only results using equal weighted portfolios or Fama-Macbeth regressions. More than half of the predictors focus on Compustat data, and about 30% use purely price data. Most of the remainder use analyst forecasts, though several focus on institutional ownership data, trading volume, or specialized data (such as Gompers, Ishii, and Metrick's (2003) governance index). The original CZ dataset has 156 predictors, but I drop one predictor with an unusually low t-stat for consistency with HLZ's sharp t-stat cutoff (Equation (12)). Monthly portfolio returns are publically available at http://sites.google.com/site/chenandrewy/code-and-data, as is code for my my estimations. For further details, please see Chen and Zimmermann (2018).

I use the CZ dataset because it allows for measurement of the correlation between test statistics. As shown by Harvey, Liu, and Zhu (2016) (see also Harvey and Liu 2013), multiple-testing adjustments for t-hurdles cannot be identified without an estimate of these correlations. The CZ data contains monthly time-series of returns for all 156 predictors, allowing for direct estimation of the entire correlation matrix. These correlations are summarized in Table 2.

Table 2: Summary Statistics for Published Pairwise Time-Series Correlations.

This table describes pairwise time-series correlations for the data used in my analysis. Bootstrapped standard errors are shown in parentheses. The data are 155 replicated long-short portfolio returns from Chen and Zimmermann (2018). All portfolios are signed to have positive sample mean returns before calculating correlations. Correlations are calculated using all available data (including post-publication). For comparison, Panel B shows mean correlations from McLean and Pontiff (2016), Green, Hand, and Zhang (2013), and Harvey, Liu, and Zhu (2016). Data is available at http://sites.google.com/site/chenandrewy/code-and-data.

Panel A: Measures of Central Tendency		
mean	median	mode
0.038	0.036	0.050
(0.009)	(0.009)	(0.046)

Panel B: Comparison of Mean Correlation with the Literature		
	mean	Description
McLean-Pontiff 2016	0.033	Replications of 97 published cross-sectional predictors
GHZ 2013	0.050	Replications of 39 published cross-sectional predictors
HLZ 2016	0.200	Calibration of model on 313 factor test t-stats

Panel C: Measures of Dispersion				
	Percentile			
stdev	10	25	75	90
0.308	-0.361	-0.148	0.227	0.432
(0.014)	(0.026)	(0.016)	(0.015)	(0.023)

Panel A of Table 2 shows that the mean pairwise time-series correlation between monthly portfolio returns is tiny, at 0.038. This mean correlation is well-measured, with a bootstrapped standard error of just 0.009. This tiny average correlation does not come from the different signs of predictability: all portfolios are constructed following instructions from the original papers and produce positive in-sample mean returns. I measure correlations using the largest overlapping sample, including post-publication periods, to avoid complications of non-overlapping samples. Correlations tend to rise

post-publication, which may lead to an upward bias in my estimate (McLean and Pontiff 2016, Cho 2017). Regardless, all of my simple estimates of the average correlation are tiny. Panel A also shows that the median correlation is tiny, at 0.036. Similarly, the modal correlation is just 0.050.

As Harvey, Liu, and Zhu (2016) emphasize the importance of correlations, Panel B of Table 2 shows mean pairwise correlations measured elsewhere in the academic literature. Overall, evidence for cross-sectional predictors favors a tiny average correlation. McLean and Pontiff (2016) find an extremely similar correlation of 0.033 among their 97 replicated portfolios. Green, Hand, and Zhang (2013) also find a tiny correlation of 0.050 among their 39 replicated predictors. A low average correlation across portfolio returns is also consistent with firm-level evidence in Jacobs and Müller (2017a). Among their dataset of 250 predictors, the mean absolute correlation among return predictive characteristics is just 0.06.

These findings of a near-zero mean correlation contrast with HLZ's benchmark correlation of 0.20. This contrast is intuitive given the commonality between the near-zero correlation datasets and their differences with the HLZ dataset. The datasets with near-zero correlation consist of only variables that have been shown to predict stock returns cross-sectionally. In contrast, HLZ's factor data includes both long-short returns and t-stats from tests of risk factor models.[14] This difference in composition may intuitively lead to differences in time-series behavior: while risk factors are likely to move together, particularly in bad times, cross-sectional predictors are often designed to demonstrate alpha, and thus by construction tend not to move with aggregate risk.[15]

Though average is close to zero, there is noticeable dispersion in correlations. Panel C of Table 2 shows that the standard deviation of correlations is non-negligible, at 0.308. Similarly, the interquartile range of correlations is 0.376. These findings are consistent with Green, Hand, and Zhang's (2013) mean absolute correlation of 0.29. High correlations are rare, however, as the 90th percentile correlation is only 0.432.

Table 3 summarizes the distributions of t-stats, sample mean returns, and standard errors on the mean returns. Panel A reports selected percentiles. I focus on the 20th,

[14]The Harvey, Liu, and Zhu (2016) factors include 113 "common" factors and 202 "characteristics," factors. HLZ define these catergories as follows: "'Common' means the factor can be viewed as a proxy for a common source of risk... ...'Characteristics' means the factor is specific to the security or portfolio."'

[15]Just two of the CZ predictors are "covariances" in the sense of Daniel and Titman (1997). These two are the CAPM beta and Kelly and Jiang's (2014) tail risk factor beta. The other 154 are "characteristics" in the Daniel-Titman sense.

50th, and 90th percentiles for ease of comparison with HLZ (Table 1). These statistics are calculated using the sample periods from the original papers. Using the original sample periods and portfolio constructions ensures that my estimates measure publication bias, and do not conflate publication bias with investor learning or other effects.

Table 3: Summary Statistics for Published t-stats, Sample Mean Returns, and Standard Errors.

This table summarizes the distribution of sample mean returns, standard errors for the sample mean, and t-stats for the test that the expected return is 0. The data are long-short portfolios based on 155 replicated cross-sectional predictors from Chen and Zimmermann (2018). Statistics for each portfolio are calculated using the sample periods from the original papers. The table shows cross-predictor percentiles. The 20th, 50th, and 90th percentiles are shown for ease of comparison with Harvey, Liu, and Zhu (2016) (Table 1). The t-stats are largely similar across both datasets. The sample mean returns are similar to those replicated by McLean and Pontiff (2016). Bootstrapped standard errors are shown in parentheses.

	Percentile		
	20	50	90
t-stat	2.12	3.40	8.05
	(0.15)	(0.19)	(0.91)
sample mean return	0.38	0.60	1.34
(%, monthly)	(0.03)	(0.04)	(0.12)
standard error	0.10	0.17	0.32
(%, monthly)	(0.01)	(0.01)	(0.01)

Based on t-stats, the CZ dataset is largely similar to the HLZ dataset. The 20th, 50th, and 90th percentiles of t-stats are within two standard errors of the HLZ data (compare Table 3 to Table 1). Both datasets show a sizable median t-stat of about 3.2.

Sample mean returns are large on average. The median sample mean return of 60 bps per month implies an annual return of 7.2%. This median monthly return is very close to the mean sample mean return of 58 bps per month in the McLean and Pontiff (2016) dataset. The data display a long right tail in sample mean returns, with 10% of returns exceeding 134 bps per month.

Table 3 also shows clear evidence of dispersion in standard errors: the 90th percentile

is 32 bps, but the 20th percentile is only 10 bps. These differences are highly statistically significant, leading me to extent the HLZ model to allow for heteroskasticity in Section 3.2.

3.2. A Model with Heterogeneous Correlations and Heteroskedasticity

To model publication bias, I closely follow Harvey, Liu, and Zhu's (2016) model with correlations. HLZ apply their model to asset pricing factors rather than published cross-sectional predictors, but their framework maps into a general setting of publication bias. Indeed their model is a special case of Andrews and Kasy's (Forthcoming) non-parametric model. I describe HLZ's model in Section 2.2, Equations (7)-(12). For brevity I do not repeat the description here, and only explain how my model differs.

My baseline model is not identical to HLZ's because I want to account for the clear evidence of dispersed correlations shown in Table 2. Modeling dispersed correlations is non-trivial, making it difficult to nest HLZ's model within a model of dispersed correlations (Lewandowski, Kurowicka, and Joe 2009, for example). Nevertheless, Section 4.1 examines a constant-correlation model that nests HLZ's and finds similar results.

As shown in Table 2, the CZ data show strong evidence that the typical correlation is very close to zero (consistent with McLean and Pontiff 2016 and Green, Hand, and Zhang 2013), but that correlations display some dispersion. Thus, I replace the constant ρ in Equation (9) with $\rho_{i,j}$, where $\rho_{i,j}$ is constructed from partial correlations that follow a symmetric beta distribution

$$\rho_{i,j|1,\ldots,i-1} \sim \text{beta}(b,b) \text{ on } [-1,1], \quad j = i+1,\ldots,N_{\text{all}} \tag{13}$$

where $\rho_{i,j|1,\ldots,i-1}$ is the partial correlation of $r_{i,t}$ and $r_{j,t}$ holding $r_{1,t},\ldots,r_{i-1,t}$ constant, and beta(b,b) is the beta distribution with both shape parameters equal to b. To map the beta distribution into $[-1,1]$, I subtract 0.5 and then multiply by 2. Given a set of partial correlations following Equation (13), the full correlation matrix can be calculated using textbook formulas (Kendall and Yule 1961, Anderson 2003). This approach to generating random correlation matricies follows methods from the multivariate analysis literature (Joe 2006, Lewandowski, Kurowicka, and Joe 2009).

Equation (13) may appear foreign, as random correlation matricies are rarely stud-

ied in finance or economics. Correlations are more commonly estimated with sample correlations, or built off of a linear factor structure (Brandt 2009). Unfortunately, these more straightforward options are not valid in my setting with selective reporting. For example, generating $N_{\text{all}} \gg 155$ portfolios by repeatedly drawing blocks of portfolios from the 155×155 published correlation matrix would imply that portfolios from different blocks are independent. In contrast, Equation (13) implies that all N_{all} portfolios are correlated, regardless of the size of N_{all}.

Moreover, Equation (13) provides a parsimonious model that can fit the published data well (Figure 3). The main alternative methods for generating a random correlation matrix from the engineering literature are based on random eigenvalues and random orthogonal matricies (Holmes 1991, for example), and are significantly more complex than Equation (13). Simple ad-hoc rejection algorithms, do not work, as they almost always lead to non-PSD matricies (Böhm and Hornik 2014).

Like the correlations in the published data, the standard errors show clear evidence of heterogeneity (Table 3). Thus, I replace the constant σ and constant T in Equation (11) with a log-normally distributed standard error

$$\log(\text{SE}_i) \sim N(\mu_\sigma, \sigma_\sigma), \text{ i.i.d.} \tag{14}$$

where $\text{SE}_i \equiv \frac{\sigma_i}{\sqrt{T_i}}$ is the standard error for portfolio i. The assumption is a simple way to model heterogeneous volatilities and sample lengths while ensuring that standard errors are positive. Moreover, Figure 3 shows that the log-normal assumption provides a good fit to the data. Assuming that log standard errors have a fat tail has little effect on the main results (Section 4).

With the heterogeneous correlations and standard errors of Equations (13) and (14), the covariance matrix of sample mean returns is

$$\text{Cov}(\bar{r}_i, \bar{r}_j) = \rho_{i,j}\text{SE}_i SE_j \tag{15}$$

where $\bar{r}_i$ is the sample mean return $\sum_{t=1}^{T_i} r_{i,t}/T_i$. Equation (15) replaces the constant correlation and homoskedasticity assumptions in Equations (8) and (9).

Last, I formally model the selection process implicit in HLZ's data modification (see

Table 1). That is, I assume that the probability that predictor i is observed follows

$$\pi(t_i|t_{\text{min}}, t_{\text{good}}, \omega) = \begin{cases} 0 & t_i < t_{\text{min}} \\ 1-\omega & t_i \in (t_{\text{min}}, t_{\text{good}}) \\ 1 & t_i > t_{\text{good}}. \end{cases} \tag{16}$$

where t_{min} is the minimum t-stat for publication, t_{good} is the value below which a t-stat is considered marginal, and ω is the fraction of marginal t-stats that go unpublished. HLZ's model assumes that $t_i > t_{\text{min}}$ implies publication with certainty, and thus they alter the data to account for marginal t-stats by duplicating marginal t-stats in accordance with the fraction ω. Equation (16) is equivalent to this data modification, and allows me to estimate their model without altering the data.

Formally, Equation (16) should only be considered a function that is proportional to the probability of publication. That is, the true probability of publication is $\tilde{\pi} = k\pi$ where k is a positive real number. Andrews and Kasy (Forthcoming) show k is not identified, despite the fact that the distribution of μ_i is identified. This identification problem leads me to examine only multiple testing statistics that are stationary as the size of the family of tests increases, such as the FDR and shrinkage for estimated magnitudes.

3.3. SMM Estimation

My estimation method follows Harvey, Liu, and Zhu (2016) closely. Like HLZ, I use SMM and target quantiles of the data. SMM allows for easy analysis of alternative assumptions (Section 4), and the robustness of quantiles leads to a smooth SMM objective. Maximum likelihood leads to similar expected return adjustments (Chen and Zimmermann 2018).

My method deviates from HLZ because of my focus on identification. I use more moment targets than they do, in order to ensure that my weak identification results are not due to inefficient use of the data. For the same reason, I weight the moment targets using the inverse of their bootstrapped variances rather than identity matrix weighting used by HLZ.

I choose three of the parameter values outside of the main SMM estimation. Reducing the number of estimated parameters helps ensure that the SMM minimization robust. First, I set $t_{\text{min}} = 1.50$ to match the minimum t-stat that Chen and Zimmer-

mann (2018) (CZ) use to define a successful replication. McLean and Pontiff (2016) use the same cutoff in their analysis. McLean and Pontiff (2016) and CZ allow for t-stats lower than 1.96 to account the fact that not all publications used long-short portfolios to demonstrate predictability. Second, I set the cutoff for a marginal t-stat $t_{\text{good}} = 2.57$ for ease of comparison with HLZ. They choose 2.57 to match the t-statistic reported in Fama and MacBeth (1973).

Third, I choose the shape parameter for partial correlations b in a smaller SMM algorithm, outside of the main SMM. Specifically, I choose b such that the deciles of $\rho_{i,j}$ fit the deciles of published pairwise correlations. This optimization results in a $b = 2.8608$ and implies an interquartile range of all correlations of 0.405, slightly larger than the interquartile range of published correlations of 0.376. I estimate b separately because I find that simultaneous estimation results in an SMM objective that is bumpy with respect to b. Moreover, the resulting full SMM produces a good fit for the distribution of published correlations, as seen in Figure 3. Choosing b to fit the standard deviation of published correlations has no effect on the main results, and using a constant positive correlation has no effect either (Section 4).

I estimate the remaining five parameters $\theta \equiv (p_0, \omega, \lambda, \mu_\sigma, \sigma_\sigma)$ via SMM. Specifically, I solve

$$\begin{aligned}\hat{\theta} = \arg\min_{\theta\in\Theta} \Bigg\{ & \sum_{k=1}^{N_{\text{sim}}} \sum_{i=1}^{9} w_{t,i} \left[Q_{\text{model},t,i}(\theta) - Q_{\text{data,t,i}} \right]^2 \\ & + \sum_{k=1}^{N_{\text{sim}}} \sum_{i=1}^{9} w_{r,i} \left[Q_{\text{model},r,i}(\theta) - Q_{\text{data,r,i}} \right]^2 \\ & + \sum_{k=1}^{N_{\text{sim}}} \sum_{i=1}^{9} w_{\text{SE},i} \left[Q_{\text{model,SE},i}(\theta) - Q_{\text{data,SE,i}} \right]^2 \\ & + \sum_{k=1}^{N_{\text{sim}}} w_F \left[F_{\text{model,marg}}(\theta) - F_{\text{data,marg}} \right]^2 \Bigg\}\end{aligned} \tag{17}$$

where N_{sim} is the number of simulations, $Q_{\text{model},t,i}(\theta)$ is the ith decile ($10\times i$th percentile) t-stat from the model given parameter θ, and $Q_{\text{data},t,i}$ is the corresponding data decile. $Q_{\text{model},r,i}(\theta)$, and $Q_{\text{model,SE},i}(\theta)$ represent deciles for sample returns and standard errors, respectively, and $F_{\text{model,marg}}(\theta)$ is the fraction of published t-stats that are between 1.50 and 2.57.

The use of deciles for moment targets leads to a smooth SMM objective, as quantiles

are more robust to sampling variation than other moments that can capture the full distribution. For example, the Poisson regression coefficients recommended by Efron (2011) lead to a very bumpy objective. Deciles are also used in HLZ's SMM estimation, though they target only 3 deciles (the 20th, 50th, and 90th percentiles). I choose to target more deciles than HLZ to help ensure that the identification problems I find are not due to not fully utilizing the data.

In addition to deciles, I also target the fraction of marginal t-stats $F_{\text{data,marg}}$. This moment should be informative about the fraction of unpublished marginal t-stats ω. Indeed, I find that omitting this moment target tends to lead to a poor fit to the left shoulder of the distribution of t-stats. This part of the distribution is important, as the sharp left shoulder is direct evidence of selective publication.

For weights $w_{t,i}$, $w_{r,i}$, $w_{\sigma,i}$, and w_F, I use the inverse of the squared standard error for the related data moment. I calculate standard errors by bootstrap. This weighting reduces emphasis on the poorly measured extreme deciles, and reduces sampling variation. This reduction in sampling variation is important for making statements about weak identification. Fully efficient two-stage SMM would further reduce sampling variation, but would lead to less transparent estimates.

For the parameter choice set Θ, I allow λ, μ_σ, and σ_σ to be unrestricted on the positive portion of the real line, but restrict p_0 and ω to discrete points. Specifically, I only allow

$$\begin{aligned} p_0 &\in \{0, 0.05, 0.10, ..., 0.95\} \\ \omega &\in \{1/3, 1/2, 2/3\}. \end{aligned} \tag{18}$$

Restricting these parameters ensures that the numerical optimizer does not get stuck in a flat region. As illustrated in Section 2.2, the SMM objective can be very flat if p_0 and λ are altered simultaneously. A similar issue arises with ω and λ. The restrictions in Equation (18) allow me to do a two-stage optimization: I first use a quasi-newton method to optimize the other parameters $(\lambda, \mu_\sigma, \sigma_\sigma)$ given p_0 and ω chosen from Equation (18). I then optimize over these 60 parameter sets to arrive at the final parameter estimates. Alternative choice sets have little effect on the main results, as long as they allow for a variety of values of p_0 and ω.

Unlike HLZ, I do not target N_{all} in my SMM objective. As shown by Andrews and Kasy (Forthcoming), the selection function (16) is only identified up to scale, and thus

N_{all} is not identified. Instead, I simply choose a large value for N_{all} and N_{sim}, and make sure that larger values do not change the results. I find that $N_{\text{all}} = 1000$ and $N_{\text{sim}} = 192$ is sufficient, and that larger values dramatically increase computational time due to the $N_{\text{all}} \times N_{\text{all}}$ correlation matrix. Simpler estimations that assume constant correlation and much larger N_{all} values also lead to similar results.

I measure estimation uncertainty using bootstrap. I repeatedly draw 155 portfolios (with replacement) and their related t-stats, sample mean returns, standard errors, and pairwise correlations, and estimate using each bootstrapped dataset. Parameteric bootstrap, which can theoretically do a better job capturing correlations (Efron and Tibshirani 1994) leads very similar results (results available upon request).

3.4. Parameter Estimates and Model Fit

Table 4 shows the resulting parameter estimates and their bootstrapped distribution. The table provides a formal demonstration of the identification issues illustrated in Section 2.2. The proportion of nulls p_0 is weakly identified, while the expected return for non-null predictors λ is pinned down well. These findings lead to the weak identification of t-hurdles and the strong identification of adjusted expected returns in Section 3.5.

Panel B of Table 4 shows that the 90% C.I. for p_0 is huge, at [0, 0.80]. Indeed, even the 50% C.I. is huge, at [0.15, 0.65]. These huge confidence intervals show that one can have little confidence in any point estimate, including Panel B's point estimate of $p_0 = 0.45$.

A similar weak identification is seen for the fraction of marginal t-stats that are unpublished ω. The 50% C.I. cannot rule out any of the three values for ω that are considered in the estimation. $\omega = 1/3$, $\omega = 1/2$, and $\omega = 2/3$ are all quite possible based on the data.

Though p_0 and ω are weakly identified, the other parameters are pinned down quite well. The bulk of the distribution for the standard deviation of expected returns among non-null predictors λ lies between 35 and 40 bps per month. Similarly, the parameters that govern the standard errors are strongly identified. I transform these parameters into the mean standard error $\mathbb{E}(\text{SE}_i)$ and the standard deviation of standard errors $\text{SD}(\text{SE}_i)$ for ease of interpretation.[16]

[16]The lognormal distribution implies that the mean and standard deviation of all standard errors are

Table 4: Parameter Estimates for a Model of Publication Bias

I estimate a model of publication bias (Section 3.2) on 155 long-short portfolios from Chen and Zimmermann (2018) (Tables 2 and 3) using SMM (Section 3.3). Bootstrap parameters are found by repeatedly drawing from the 155 portfolios and re-estimating the model. The probability of drawing a null predictor p_0 is weakly identified, with huge confidence bounds. In contrast, the standard deviation of expected returns for non-null predictors λ is strongly identified. Code is available at http://sites.google.com/site/chenandrewy/code-and-data.

Panel A: Calibrated Parameters			
		Value	Motivation
Minimum t-stat	$t_{\min}$	1.50	Lower limit in CZ and MP
Marginal t-stat Cutoff	$t_{\max}$	2.57	Ease of Comparison with HLZ
Dispersion of Partial Corr	b	2.86	SMM fit of $\rho_{i,j}$ on published corr

Panel B: SMM Estimates							
		Point	Bootstrapped Distribution				
		Estimate	5	25	50	75	95
Probability of Null	p_0	0.45	0.00	0.15	0.40	0.65	0.80
Unpub Marginal t-stats	ω	0.50	0.33	0.33	0.50	0.67	0.67
S.D. $\mathbb{E}(\bar{r}_i)$ for Non-Null	λ	0.38	0.32	0.35	0.37	0.40	0.44
Mean Std Err	$\mathbb{E}(\text{SE}_i)$	0.23	0.21	0.22	0.23	0.24	0.25
S.D. of Std Err	$\text{SD}(\text{SE}_i)$	0.13	0.11	0.12	0.13	0.14	0.16

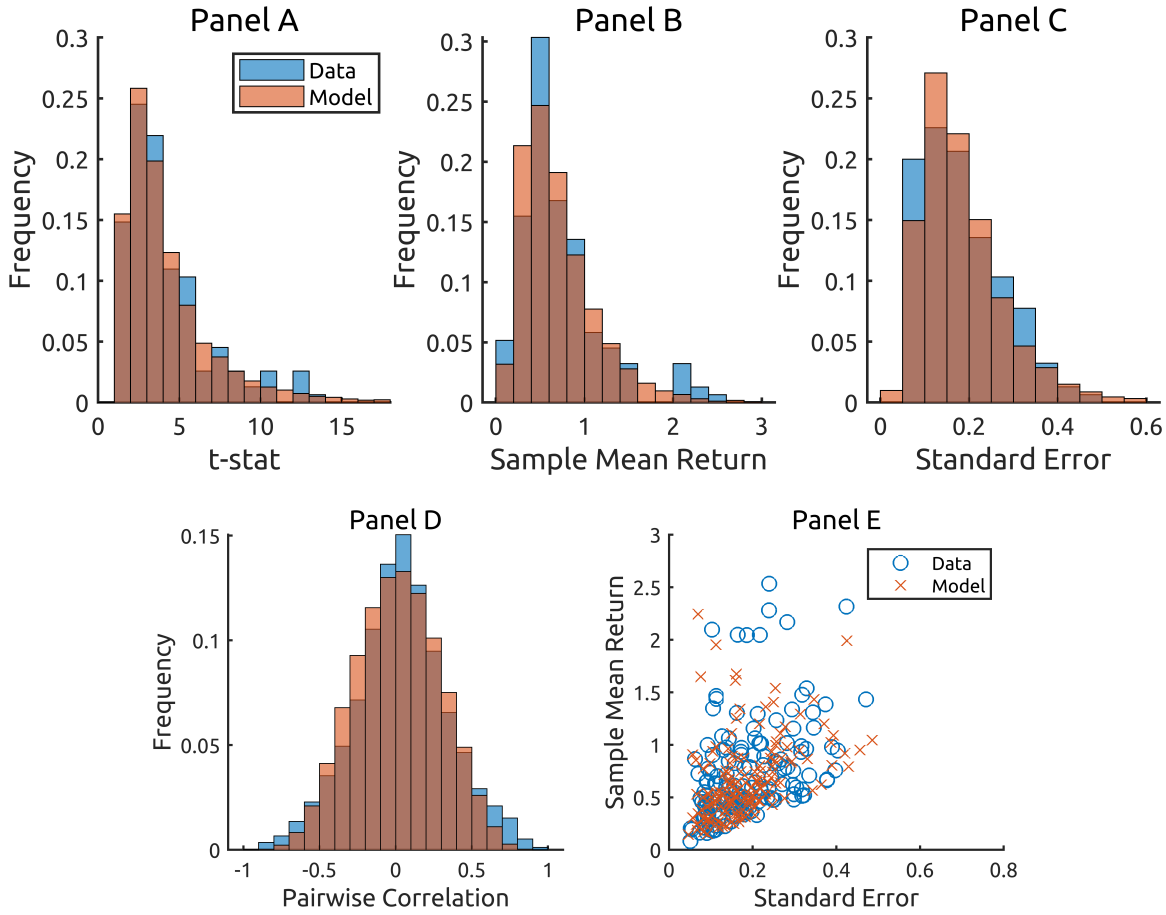
Interestingly, the estimated mean standard error of 0.23% per month is noticeably larger than the mean published standard error of 0.19% per month. Intuitively, the publication process selects for large t-stats, and thus selects for small standard errors, making the mean published standard errors downward biased compared to the mean standard error from the complete data. Similarly, the estimates imply that the complete data show much more dispersion than the published standard errors.

given by

$$\mathbb{E}(\text{SE}_i) = \exp\left(\mu_\sigma + \sigma_\sigma^2/2\right) \tag{19}$$

$$\text{SD}(\text{SE}_i) = \sqrt{\left[\exp(\sigma_\sigma^2) - 1\right]\exp(2\mu_\sigma + \sigma_\sigma^2)}. \tag{20}$$

Figure 3: Model Fit: Point Estimate. I simulate the model (Section 3.2) using the point estimate (Table 4) and compare model-implied published data with empirical published data (see Tables 2-3). Panels A - C show the distributions of published t-stats, sample mean returns, and standard errors across replications of 155 long-short portfolios based on published predictors (data) with the model. Panel D shows the distribution of pairwise correlations between published monthly portfolio returns. Panel E shows a scatterplot of the relationship between standard errors and sample mean returns. The point estimate fits all of these distributions well.



Panels A - C of Figure 3 show that the point estimate fits the distribution of several statistics from the data. The model fits the distributions of published t-stats (Panel A), published mean returns (Panel B), and published standard errors (Panel C). In all three panels, almost all histogram bins show a close fit between model and data.

The point estimate also fits the distribution of published correlations very well (Panel D). As in the data, the model implies a symmetric distribution centered around zero.

The model captures the dispersion of correlations well too. This close fit is important, as Harvey, Liu, and Zhu (2016) emphasize the importance of fitting correlations for correct multiple testing inference. Panel E shows that the model also fits the correlation between published sample mean returns and published standard errors. This fit is notable because the model assumes no correlation in the full family of tests (Equations (7) and (15)). Thus, publication bias is enough to generate the observed correlation.

Overall, the strong fit shown in Figure 3 suggests that the model specification is good.

3.5. The Bootstrapped Distribution of Publication Bias Adjustments

With parameter estimates in hand, I can finally bootstrap estimates of publication bias adjustments. I first define the adjustments, and then present the main result in Figure 4.

Publication Bias Adjustment Definitions I examine four kinds of publication bias adjustments: (1) t-hurdles that control the FDR, (2) the FDR for published predictors, (3) a simple shrinkage adjustment for expected returns (à la Cochrane 2005), and (4) the James and Stein (1961) shrinkage of Chen and Zimmermann (2018).

Following Benjamini and Hochberg (1995), I define the FDR for a particular t-stat hurdle t_{hurdle} as

$$\text{FDR}(t_{\text{hurdle}}) \equiv \mathbb{E}\left[\frac{\#\{\text{null}_i : t_i \geq t_{\text{hurdle}}\}}{\max\{\#\{t_i \geq t_{\text{hurdle}}\}, 1\}}\middle|\hat{\theta}, b, t_{\text{min}}, t_{\text{good}}\right], \tag{21}$$

where $\hat{\theta}$ are the estimated parameters, b, t_{min}, t_{good} are calibrated parameters, null_i is defined by

$$\text{null}_i \text{ if } \mu_i \sim \delta(0), \tag{22}$$

and $\delta(0)$ is the distribution with a point mass at 0. In words, $\text{FDR}(t_{\text{hurdle}})$ is the expected proportion of null predictors among t-stats that exceed t_{hurdle}.

I calculate Equation (21) by Monte Carlo simulation of the estimated model. For various values of t_{hurdle}, I simulate the model many times, calculate the proportion of discoveries that are null $\#\{\text{null}_i : t_i \geq t_{\text{hurdle}}\}/\#\{t_i \geq t_{\text{hurdle}}\}$, and finally average the across simulations to take the expectation. This calculation can be justified by the law of large

numbers, given standard regularity conditions. The presence of correlated random variables means that other forms of numerical integration are difficult.

Then, to find the t-hurdles that control the FDR at a pre-determined level α ($t_{\text{FDR},\alpha}$), I solve

$$t_{\text{FDR},\alpha} = \min_{\bar{t}_{\text{hurdle}}} \text{FDR}(\bar{t}_{\text{hurdle}}) \leq \alpha \tag{23}$$

where α is either 5% or 1%, following Harvey, Liu, and Zhu (2016).

I also examine the FDR among published t-stats. Analogous to Equation (21), I define this FDR as

$$\text{FDR}_{\text{pub}} \equiv \mathbb{E}\left[\frac{\#\{\text{null}_i : i \text{ is published}\}}{\max\{\#\{i : i \text{ is published}\}, 1\}} \middle| \hat{\theta}, b, t_{\min}, t_{\text{good}} \right]. \tag{24}$$

Equation (24) answers the intuitive question: what is the fraction of published predictors that we expect to be null?

FDRs are a form of null hypothesis testing, and thus they require a definition of a null predictor. I use Equation (22) as it corresponds to the classical hypothesis test in asset pricing. Indeed, testing for $\mu_i \sim \delta(0)$ is exactly where the classical t-hurdle of 1.96 comes from, and $\mu_i \sim \delta(0)$ is the null used by Harvey, Liu, and Zhu (2016).

There are reasons to move away from this null, however. According to Equation (22), even a tiny μ_i of 1 bps per month is a "true discovery." Such a tiny mean return is arguably not notable, particularly when examining so many predictors. Indeed, Efron (2004) and Efron et al. (2007) argue that in many multiple test settings, one should estimate an "empirical null" from the data. Deciding on a new null for asset pricing is a tricky issue, however. Indeed, related issues lead McShane et al. (2019) to argue that null hypothesis testing should be abandoned altogether.

These problems with null hypothesis testing lead me to examine other multiple testing adjustments. In particular, I examine two shrinkage adjustments for estimated magnitudes. In the context of cross-sectional predictability, the magnitude is given by the expected return.

The first expected return adjustment draws directly from Harvey, Liu, and Zhu's (2016) model with correlations and is similar to Cochrane's (2005) selection adjustment for venture capital returns. As the model implies a distribution of expected returns for

published predictors, taking means across this distribution leads to a simple estimate of the average publication bias adjusted return. Scaling by the average published mean return in the model provides a "smooth" shrinkage adjustment

$$s_{\text{smooth}} = 1 - \frac{\mathbb{E}\left[\sum_{i=1}^{N_{\text{pub}}} \mu_i | \text{pub}_i; \hat{\theta}, b, t_{\min}, t_{\text{good}}\right]}{\mathbb{E}\left[\sum_{i=1}^{N_{\text{pub}}} \bar{r}_i | \text{pub}_i; \hat{\theta}, b, t_{\min}, t_{\text{good}}\right]}, \tag{25}$$

where $\hat{\theta} \equiv (\hat{p}_0, \hat{\omega}, \hat{\lambda}, \hat{\mu}_\sigma, \hat{\sigma}_\sigma)$ are the estimated parameters using SMM, and b, $t_{\min}$, and t_{good} are the calibrated parameters. I call this the "smooth" shrinkage because it smooths over noise in sample mean returns before calculating the adjustment. This smooth shrinkage is a direct estimate of McLean and Pontiff's (2016) upper bound on statistical effects. I calculate the expectation by Monte Carlo.

The second expected return adjustment uses James and Stein (1961). Building on Efron (2011) (see also Liu, Moon, and Schorfheide 2016), Chen and Zimmermann use the James-Stein estimator to develop a shrinkage adjustment for expected returns at the predictor level. This adjustment is found by taking the expectation of the expected return, conditional on the observed sample mean return, standard error, and estimated expected return parameters $\hat{p}_0$ and $\hat{\lambda}$

$$\hat{\mu}_i \equiv \mathbb{E}\left[\mu_i | \bar{r}_i^{\text{data}}, \text{SE}_i^{\text{data}}; \hat{p}_0, \hat{\lambda}\right]. \tag{26}$$

The fact that Bayes rule is immune to selection (Senn 2008, Dawid 1994) implies that Equation (26) leads to an unbiased estimate of μ_i. Chen and Zimmermann show that this unbiasedness holds using simulated estimations. I calculate the expectation by numerical integration. For more details please see Chen and Zimmermann (2018).

To ease comparison with McLean and Pontiff (2016) and the smooth shrinkage estimate (Equation (25)), I express $\hat{\mu}_i$ in terms of shrinkage

$$s_i = 1 - \frac{\hat{\mu}_i}{\bar{r}_i^{\text{data}}}. \tag{27}$$

Finally to summarize these predictor-level shrinkage factors, I take the mean and median s_i across published predictors. An in-depth analysis of heterogeneity in shrinkage estimates can be found in Chen and Zimmermann (2018).

Equations (26) and (27) offer several advantages over the smooth shrinkage (Equa-

tion (25)). The first is that it is less dependent on model parameters. Only $\hat{p}_0$ and $\hat{\lambda}$ are used, as $\bar{r}_i^{\text{data}}$ and $\text{SE}_i^{\text{data}}$ are taken directly from the data. Moreover, the James-Stein shrinkage has a closed form solution if μ_i is normally distributed:

$$s_i^{\text{normal}} = \frac{[\text{SE}_i^{\text{data}}]^2}{[\widehat{\text{SD}}(\mu_i)]^2 + [\text{SE}_i^{\text{data}}]^2}. \tag{28}$$

This equation provides a simple interpretation of shrinkage. Shrinkage is a noise-to-signal ratio, where $\text{SE}_i^{\text{data}}$ measures the amount of noise and $\widehat{\text{SD}}(\mu_i)$ captures the amount of signal. Intuitively, if there is no dispersion in expected returns ($\widehat{\text{SD}}(\mu_i) = 0$) then all expected returns should be shrunk to zero.

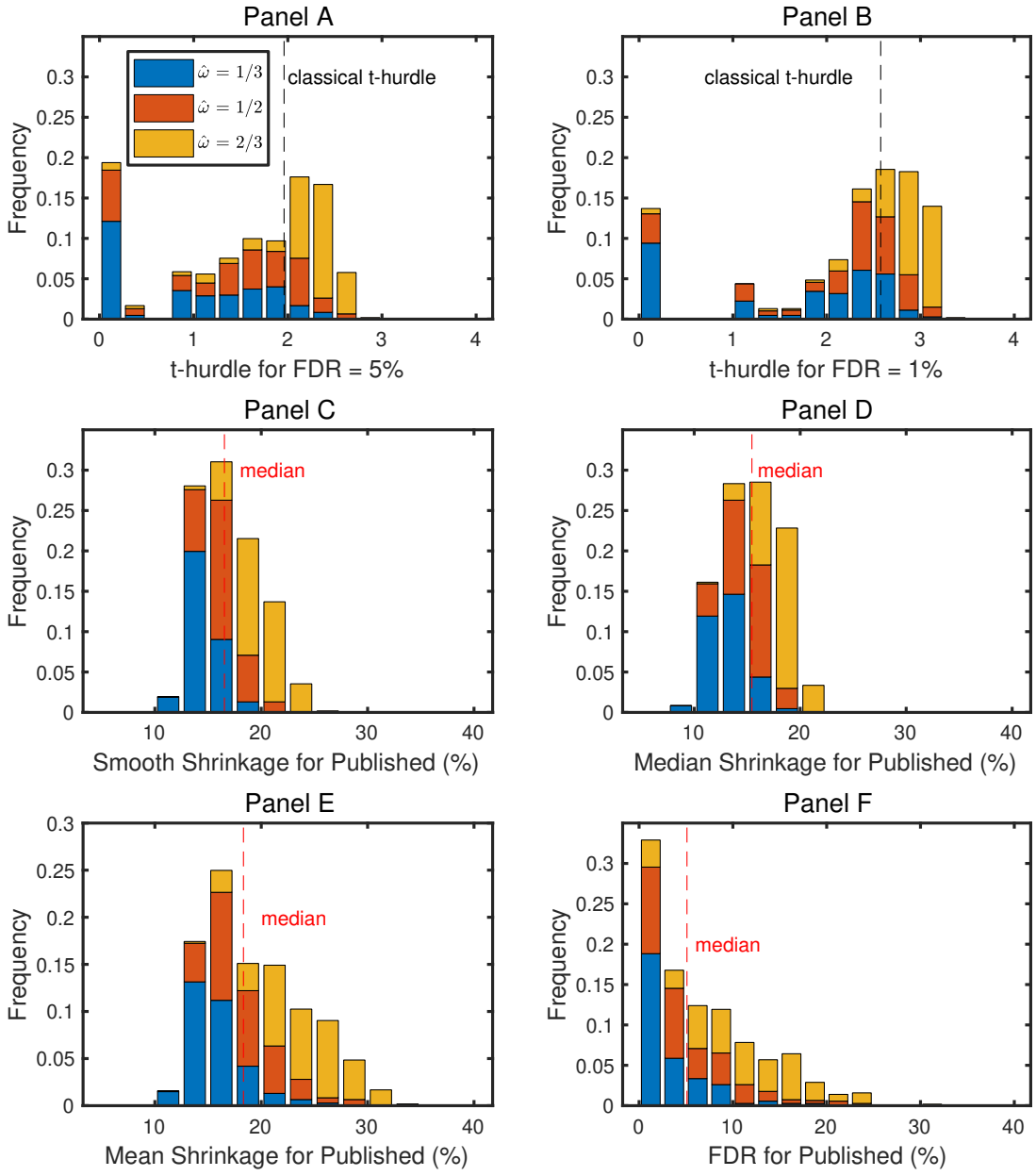
Results Figure 4 plots the main results. I resample from the dataset of 155 predictors, re-estimate the model, and re-calculate publication bias adjustments 1000 times. The figure shows the distribution of t-hurdles, FDRs, and shrinkage factors generated by this bootstrap. I subset the distributions in terms of $\hat{\omega}$, the fraction of marginal t-stats that are estimated to be missing (Equation (16)), to illustrate identification.

Panels A and B show t-hurdles that control the FDR at 5% and 1%. These are the same FDR levels examined by Harvey, Liu, and Zhu (2016) (HLZ). The panels show that these t-hurdles are weakly identified—that is, the bootstrapped distributions are very dispersed. The t-hurdle for an FDR of 5% ranges from 0 to about 2.6, and the t-hurdle for an FDR of 1% ranges from 0 to about 3.2. This weak identification is intuitive, given that the proportion of nulls p_0 is weakly identified (Table 4). As illustrated in Section 2.1, p_0 is critical to determining t-hurdles that control the FDR.

The weak identification of p_0, in turn, is due to the nature of the selective reporting in published data. Predictors that are close-to-null are unlikely to be published. As a result, the proportion of nulls p_0 must be extrapolated, and uncertainty about this extrapolation leads to uncertainty about t-hurdles that control the expected proportion of predictors that are null (false), that is, the FDR.

Subsets of the t-hurdle distribution based on the fraction of marginal t-stats that are missing ($\hat{\omega}$) illustrate this extrapolation in detail. For estimates that find $\hat{\omega} = 2/3$ (2/3 of t-stats between 1.5 and 2.57 are missing), t-hurdles tend to be large. Intuitively, if a large share of marginal t-stats is missing, then the proportion of unobserved t-stats near 2.0 is large, and extrapolating to 0 results in a large proportion of nulls. In contrast,

Figure 4: Bootstrapped Distribution of Publication Bias Adjustments. I resample from the dataset of 155 predictors (with replacement), and estimate a model of biased publication (Section 3.3) 1000 times. $\hat{\omega}$ is the estimated fraction of marginal t-stats that are unpublished. t-hurdles are calculated using Equations (21)-(24) by Monte Carlo. Shrinkage is defined using [Adjusted Return] = (1 − Shrinkage)[Sample Mean Return], and is calculated using Equation (25) (Panel C) or Equation (27) (Panels D and E).



for $\hat{\omega} = 1/3$, the proportion of t-stats near 2.0 is small, and extrapolating results in a small proportion of nulls. The fact that $\hat{\omega}$ is about equally distributed between the three possible values 1/3, 1/2, and 2/3 shows that the data do not speak strongly about $\hat{\omega}$. Thus, steep extrapolations (large $\hat{\omega}$)are as likely as shallow ones (small $\hat{\omega}$), leading to significant uncertainty about p_0 and t-hurdles.

Not only are t-hurdle distributions dispersed, but they are centered around their classical counterparts. For a classical two-sided test with a 5% significance level and a large number of degrees of freedom, the corresponding t-hurdle is 1.96 (dashed line, Panel A). The corresponding t-hurdle for the 1% significance level is 2.58 (dashed line, Panel B). These classical significance levels are not FDRs, but they have a close relationship. Indeed, the classical significance levels are an upper bound on the Efron and Tibshirani (2002) FDR (see Efron 2012). Overall, these results show that it is very unclear if t-hurdles should be raised or lowered, once weak identification (indeed, just sampling variability) is accounted for.

Other publication bias adjustments, however, require less extrapolation. In particular, adjustments for published expected returns have small standard errors, as seen in Panel C of Figure 4. Panel C shows the smoothed shrinkage (Equation (25)), which is the direct estimate of McLean and Pontiff's (2016) upper bound on statistical effects for published mean returns. The distribution of smoothed shrinkage is centered around 16%, well within the 26% upper bound estimated by McLean and Pontiff. Moreover, the standard error is small, at just 3 percentage points. This small standard error is nearly an order of magnitude smaller than McLean and Pontiff's standard error of 13 percentage points.

Panels D and E show that the James-Stein adjustment of Chen and Zimmermann (2018) leads to similar results. This adjustment (Equation (27)) applies at the predictor level, and Panels E and F show the median and mean shrinkage across predictors, respectively. Both panels show that the typical shrinkage is strongly identified, and imply that the vast majority of the typical predictor's in-sample return is not due to publication bias.

Shrinkage adjustments are strongly identified because they are determined by both p_0 and λ. Indeed, for published predictors, which are far from the null, λ is the dominant parameter. And as seen in Table 4, standard errors on λ are small. This strong identification comes from the fact that the null distribution has a tail that is far too thin to account

for the right tail in published t-stats. Thus, a large dispersion in expected returns is required to fit this right tail, and this right tail is well-observed, as the model assumes that all t-stats > 2.57 are published (as in Harvey, Liu, and Zhu 2016). This identification is studied in detail in Chen and Zimmermann (2018) and Chen (2019).

Finally, Panel F shows that the FDR among published predictors implies that the vast majority of published results are true, even after accounting for identification. The median of the bootstrapped estimates is about 5.1%, with a standard error of 6.0%. This small FDR is, perhaps surprisingly, consistent with the Harvey, Liu, and Zhu's (2016) baseline parameter estimates.[17]

All together, the bootstrapped publication bias adjustments of Figure 4 formally demonstrate the main argument of this paper for the literature on predicting the cross-section of stock returns. Estimates of statistical standards that control for multiple testing are weakly identified. In contrast, shrinkage adjustments are strongly identified. These results are due to the particular kind of selective reporting that is exhibited by academic publications across many literatures (Fanelli 2010), and are consistent with the fact that multiple testing statistics do not necessarily imply raising statistical standards (Benjamini and Hochberg 2000). Indeed, the strongly identified adjustments imply that publication bias for cross-sectional predictors is modest.

4. Alternative Model Specifications

To further examine identification, this section examines a wide variety of modeling assumptions. Table 5 summarizes these robustness checks. In short, the 10 alterna-

[17] To show this, one can calculate the empirical Bayesian FDR for observed factors in HLZ by hand, following Efron, Tibshirani, Storey, and Tusher (2001):

$$\begin{aligned} \Pr(\text{null}_i|\text{obs}_i) &= \frac{\Pr(\text{obs}_i|\text{null}_i)\Pr(\text{null}_i)}{\Pr(\text{obs}_i)} \\ &= \frac{\Pr(t_i > 1.96|\text{null}_i)\,p_0}{[\#\text{observed}]/N_{\text{all}}}. \end{aligned} \tag{29}$$

Then using baseline estimates from HLZ's Table 5 and the standard normal CDF, we have

$$\Pr(\text{null}_i|\text{obs}_i) = \frac{0.025(0.444)}{353/1378} = 4.33\%, \tag{30}$$

where 353 is the number of observed factors from HLZ's target moments. Simulations of HLZ's baseline model and using the Benjamini and Hochberg (1995) FDR lead to similar results (available upon request).

tive assumptions lead to the same main results: t-stat hurdles that control the FDR are weakly identified, shrinkage adjustments for effect magnitudes are strongly identified, and the FDR among published predictors imply that the vast majority of published predictors are true discoveries.

Table 5: Robustness

I bootstrap multiple testing statistics for 10 alternative models (Section 4). This table shows the median (med) and standard deviation (sd) of the bootstrapped distributions. Definitions for the multiple testing statistics are found in Section 3.5. Each row represents a different set of modeling assumptions. The main findings in Figure 4 are reinforced in the bootstraps of these 10 alternative models.

		t-hurdles				FDR		Smooth Shrink	
		FDR = 5%		FDR = 1%		for Published		for Published	
		med	sd	med	sd	med	sd	med	sd
(1)	Constant Corr	1.42	0.85	2.51	0.83	2.98	7.11	16.02	2.90
(2)	Const Corr = 0.20	1.53	0.91	2.56	1.00	3.64	6.42	16.27	2.84
(3)	More SD Corr	1.82	0.98	2.66	1.12	5.66	7.38	16.81	3.25
(4)	Risk Prem	2.07	0.69	2.82	0.67	9.21	7.23	13.57	3.14
(5)	Logit Select	0.00	0.58	0.00	0.97	0.00	2.50	12.73	1.27
(6)	Gamma Shape 2.0	2.09	0.77	2.73	0.86	11.12	8.19	16.97	3.70
(7)	Gamma Shape 0.5	1.56	0.87	2.36	1.01	3.58	4.06	18.40	1.83
(8)	t d.o.f. 4	1.35	0.95	2.25	1.13	2.50	8.00	15.60	3.46
(9)	t d.o.f. 10	2.02	0.84	2.67	0.80	9.33	8.39	17.41	3.48
(10)	Fat Log SE	2.44	0.86	3.11	0.90	14.32	8.19	20.90	3.77

4.1. Constant Estimated Correlation

The main estimation focuses on the dispersion of correlations because empirical evidence suggests that the average correlation is close to zero (Section 3.1. In this robustness check I examine the assumption of a constant, but on average non-zero, correlation. That is, I use the same assumption in Harvey, Liu, and Zhu (2016) regarding correlations, and use Equation (9) instead of Equation (13) to model correlations. This model differs from HLZ only in that it has heteroskedasticity. Row (1) of Table 5 shows that assuming a constant, but on average non-zero correlation has little effect on the main results.

4.2. Constant Correlation of 0.20

In this robustness check I assume a constant correlation but do not estimate it. Instead, I assume the baseline value assumed in Harvey, Liu, and Zhu (2016). Row (2) of Table 5 shows that this assumption has little effect on the main results.

The finding that a correlation of 0.20 has little effect on t-hurdles is consistent with Table 5 of Harvey, Liu, and Zhu (2016). There HLZ show that increasing the correlation from 0 to 0.20 has little effect on t-hurdles. The t-hurdle for an FDR of 5% from 2.16 to 2.27. Similarly the t-hurdle for an FDR of 1% increases from 2.88 to 2.95.

4.3. More Dispersed Correlations

In this robustness check I assume heterogeneous correlations, as in the baseline model, but target the standard deviation of published correlations rather than the deciles of published correlations. This assumption leads to a slightly larger beta parameter b in Equation (13), and a slightly larger standard deviation of correlations. Row (3) of Table 5 show that this assumption has little effect on the main results.

4.4. Correlation Between Sample Mean Returns and Standard Errors

The baseline model assumes that expected returns and standard errors are uncorrelated (Equation (11)). This assumption conflicts with the classical assumption of a risk-return tradeoff, that is, that expected returns are positively related to volatility.

To examine the alternative assumption of a positive risk-return tradeoff, I replace Equation (11) with

$$\tilde{\mu}_i = \mu_i + \beta \text{SE}_i \tag{31}$$

where, as before, μ_i follows Equation (7) and SE_i follows Equation (14). Estimating both β and ω simultaneously runs into questions of identification, as both of these parameters are related to the slope of the relationship between sample mean returns and standard errors. Thus, to be conservative, I calibrate β before the SMM. Specifically, I run a regression of sample mean returns on standard errors, find a slope of 2.15, and I assume $\beta = 1.08 = 2.15/2$. This assumption allows for half of the relationship between mean

returns and standard errors to be accounted for by risk. The remainder of the relationship, then is accounted for by selective reporting. For comparison, assuming an annual equity premium of 7% and a volatility of 15%, and using the median sample size of 300 months implies a slope of $\frac{7}{15}\frac{1}{\sqrt{12}}\sqrt{300} = 2.33$

Row (4) of Table 5 show that this assumption has little effect on the main results.

4.5. An Alternative Modeling of Selection

The bootstrap results show that there is considerable uncertainty about the selection function. Here I examine an alternative functional form for selection:

$$\pi_{\text{logistic}}(t_i|t_{\text{mid}}, t_{\text{slope}}) = \frac{1}{1+\exp(-t_{\text{slope}}(t_i - t_{\text{mid}}))}. \tag{32}$$

Similar to Equation (16), this function allows for the concepts of marginal and readily publishable t-stats. This function differs in that it is smoother (linear near its midpoint), and has 2 rather than 3 parameters. This is the same function used in Chen and Zimmermann (2018), and is also used in Cochrane's (2005) model of bias in venture capital returns.

Row (5) of Table 5 show that this assumption implies that t-hurdles can be confidently lowered. This lack of robustness highlights the fact that t-hurdles are weakly identified. Indeed, the estimation of t-hurdles is highly sensitive to the modeling of selective reporting. The low t-hurdles implied by Equation (32) are due to the fact that the estimated slope is very strict, and thus estimation leads to a very steep selection function. This steep selection function is similar to a close-to-zero ω. As with the low ω results in Figure 4, the logistic selection function implies a low t-stat threshold.

In contrast, row (5) of Table 5 shows that other results are robust. Under the logistic selection function, the vast majority of published results are true (the FDR is small), and published sample mean returns are only modestly biased upward, even after accounting for standard errors.

4.6. Alternative Distributions for Expected Returns

Few papers attempt to model the distribution of expected returns among published factors or predictors. The only papers, to my knowledge, are Harvey, Liu, and Zhu (2016),

who assume a mixture exponential distribution, and Chen and Zimmermann (2018), who assume a t-distribution.

In my baseline model, I follow Harvey, Liu, and Zhu (2016) for ease of comparison. Rows (6)-(9) of Table 5 examine alternative distributions. These alternative assumptions have little effect on the main results.

Rows (6) and (7) consider replacing the exponential distribution (Equation (7)) with a gamma distribution. Row (6) assumes a shape parameter of 2.0 and estimates the scale parameter. Row (7) assumes a shape of 0.5. Both gamma distributions have little effect on the main results.

Rows (8) and (9) replace the exponential distribution with a t distribution with d.o.f. parameter equal to 4 and 10, respectively. Both t-distributions have little effect on the main results.

4.7. Fat-Tailed Log Standard Errors

Here, I examine replacing the normal assumption for log standard errors (Equation (14)) with a t-distribution with degrees of freedom = 4. This assumption is somewhat extreme, as Figure 3 shows that the log-normal assumption fits the right tail of the data very well.

Row (10) of Table 5 shows that fat-tailed standard errors lead to higher t-hurdles, but they are still within one standard error of their classical counterparts. Similarly, shrinkage is somewhat larger than other estimates but not unusually large considering its standard error.

5. Conclusion

Recently, academics in social and biomedical sciences have argued that standards of statistical significance need to be raised (Harvey, Liu, and Zhu 2016, Benjamin et al. 2018). I show that this idea is unlikely to find support from multiple testing statistics, as estimates of statistical standards that control for multiple testing among published hypothesis tests tend to be weakly identified. In contrast, shrinkage adjustments for estimated magnitudes tend to be strongly identified. This strong identification favors the idea that statistical significance should be abandoned, and that test statistics should

be treated continuously instead (Amrhein and Greenland 2018, McShane et al. 2019).

I formally demonstrate these arguments for the literature on cross-sectional stock return predictability using a dataset of 155 published predictors. Bootstrapped adjustments to statistical standards are weakly identified and uninformative. In contrast, estimates of shrinkage adjustments for expected returns are strongly identified, and imply that publication bias is modest, consistent with McLean and Pontiff (2016). These results are due to identification issues that are likely to exist in other literatures: test statistics that are close to the null are unlikely to be observed and must be extrapolated, but the right tail of test statistics is well-observed, as highly significant results are likely to be published.

References

Amrhein, Valentin and Sander Greenland. "Remove, rather than redefine, statistical significance". *Nature Human Behaviour* 2.1 (2018), p. 4.

Anderson, TW. "An Introduction to Multivariate Analysis, John Wiley & Sons (New York)" (2003).

Andrews, Isaiah and Maximilian Kasy. "Identification of and correction for publication bias". *the American Economic Review* (Forthcoming).

Benjamin, Daniel J, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. "Redefine statistical significance". *Nature Human Behaviour* 2.1 (2018), p. 6.

Benjamini, Yoav. "Discovering the false discovery rate". *Journal of the Royal Statistical Society: series B (statistical methodology)* 72.4 (2010), pp. 405–416.

Benjamini, Yoav and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

— "On the adaptive control of the false discovery rate in multiple testing with independent statistics". *Journal of educational and Behavioral Statistics* 25.1 (2000), pp. 60–83.

Benjamini, Yoav and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency". *Annals of statistics* (2001), pp. 1165–1188.

Böhm, Walter and Kurt Hornik. "Generating random correlation matrices by the simple rejection method: Why it does not work". *Statistics & Probability Letters* 87 (2014), pp. 27–30.

Brandt, M.W. "Portfolio choice problems". *Handbook of financial econometrics* 1.1 (2009).

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. "Star wars: The empirics strike back". *American Economic Journal: Applied Economics* 8.1 (2016), pp. 1–32.

Card, David and Alan B Krueger. "Time-series minimum-wage studies: a meta-analysis". *The American Economic Review* 85.2 (1995), pp. 238–243.

Chen, Andrew Y. "The Limits of p-Hacking: a Thought Experiment". *Available at SSRN* (2019).

Chen, Andrew Y and Mihail Velikov. "Accounting for the Anomaly Zoo: A Trading Cost Perspective". *Available at SSRN: https://papers.ssrn.com/abstract=3073681* (2018).

Chen, Andrew Y and Tom Zimmermann. "Publication Bias and the Cross-Section of Stock Returns". *Available at SSRN: https://ssrn.com/abstract=2802357* (2018).

Cho, Thummim. *Turning alphas into betas: Arbitrage and endogenous risk*. Tech. rep. Harvard University Mimeo, 2017.

Chordia, Tarun, Amit Goyal, and Alessio Saretto. "p-hacking: Evidence from two mmillion trading strategies" (2017).

Christensen, Garret and Edward Miguel. "Transparency, reproducibility, and the credibility of economics research". *Journal of Economic Literature* 56.3 (2018), pp. 920–80.

Cochrane, John H. "The risk and return of venture capital". *Journal of financial economics* 75.1 (2005), pp. 3–52.

Daniel, K. and S. Titman. "Evidence on the Characteristics of Cross Sectional Variation in Stock Returns". *Journal of Finance* 52.1 (1997), pp. 1–33.

Dawid, AP. "Selection paradoxes of Bayesian inference". *Lecture Notes-Monograph Series* (1994), pp. 211–220.

De Long, J Bradford and Kevin Lang. "Are all economic hypotheses false?" *Journal of Political Economy* 100.6 (1992), pp. 1257–1272.

Efron, Bradley. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press, 2012.

Efron, Bradley. “Large-scale simultaneous hypothesis testing: the choice of a null hypothesis”. *Journal of the American Statistical Association* 99.465 (2004), pp. 96–104.

Efron, Bradley et al. “Size, power and false discovery rates”. *The Annals of Statistics* 35.4 (2007), pp. 1351–1377.

Efron, Bradley. “Tweedie’s formula and selection bias”. *Journal of the American Statistical Association* 106.496 (2011), pp. 1602–1614.

Efron, Bradley and Robert Tibshirani. “Empirical Bayes methods and false discovery rates for microarrays”. *Genetic epidemiology* 23.1 (2002), pp. 70–86.

Efron, Bradley and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Efron, Bradley, Robert Tibshirani, John D Storey, and Virginia Tusher. “Empirical Bayes analysis of a microarray experiment”. *Journal of the American statistical association* 96.456 (2001), pp. 1151–1160.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. “Bias in meta-analysis detected by a simple, graphical test”. *Bmj* 315.7109 (1997), pp. 629–634.

Fama, Eugene F and James D MacBeth. “Risk, return, and equilibrium: Empirical tests”. *The Journal of Political Economy* (1973), pp. 607–636.

Fanelli, Daniele. ““Positive” results increase down the hierarchy of the sciences”. *PloS one* 5.4 (2010), e10068.

Fisher, RA. “Statistical methods for research workers.” (1925).

Genovese, Christopher and Larry Wasserman. “A stochastic process approach to false discovery control”. *The Annals of Statistics* 32.3 (2004), pp. 1035–1061.

Gompers, Paul, Joy Ishii, and Andrew Metrick. “Corporate governance and equity prices”. *The quarterly journal of economics* 118.1 (2003), pp. 107–156.

Green, Jeremiah, John RM Hand, and X Frank Zhang. “The characteristics that provide independent information about average us monthly stock returns”. *The Review of Financial Studies* (2017), hhx019.

— “The supraview of return predictive signals”. *Review of Accounting Studies* 18.3 (2013), pp. 692–730.

Harvey, Campbell R. “Presidential address: The scientific outlook in financial economics”. *The Journal of Finance* 72.4 (2017), pp. 1399–1440.

Harvey, Campbell R and Yan Liu. “False (and missed) discoveries in financial economics”. *Available at SSRN 3073799* (2018).

Harvey, Campbell R, Yan Liu, and Heqing Zhu. "... and the cross-section of expected returns". *The Review of Financial Studies* 29.1 (2016), pp. 5–68.

Harvey, Campbell and Yan Liu. "Multiple testing in economics" (2013).

Herring, Chris. "Why NBA players lie about their height". *The Wall Street Journal* (2016).

Holmes, Richard B. "On random correlation matrices". *SIAM journal on matrix analysis and applications* 12.2 (1991), pp. 239–272.

Hou, Kewei, Chen Xue, and Lu Zhang. *Replicating Anomalies*. Tech. rep. National Bureau of Economic Research, 2017.

Hsu, Jason. *Multiple comparisons: theory and methods*. Chapman and Hall/CRC, 1996.

Jacobs, Heiko and Sebastian Müller. "... And Nothing Else Matters? On the Dimensionality and Predictability of International Stock Returns". *On the Dimensionality and Predictability of International Stock Returns (May 25, 2017)* (2017).

— "Anomalies across the globe: Once public, no longer existent?" (2017).

James, William and Charles Stein. "Estimation with quadratic loss". *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1961. 1961, pp. 361–379.

Jensen, Michael C and George A Bennington. "Random walks and technical theories: Some additional evidence". *Journal of Finance* 25 (1970), pp. 469–482.

Joe, Harry. "Generating random correlation matrices based on partial correlations". *Journal of Multivariate Analysis* 97.10 (2006), pp. 2177–2189.

Johnson, Valen E. "Revised standards for statistical evidence". *Proceedings of the National Academy of Sciences* 110.48 (2013), pp. 19313–19317.

Kelly, Bryan and Hao Jiang. "Tail risk and asset prices". *Review of Financial Studies* (2014), hhu039.

Kendall, Maurice George and G Udny Yule. *An introduction to the theory of statistics*. Griffin & Company, 1961.

Korajczyk, Robert A and Ronnie Sadka. "Are momentum profits robust to trading costs?" *The Journal of Finance* 59.3 (2004), pp. 1039–1082.

Lesmond, David A, Michael J Schill, and Chunsheng Zhou. "The illusory nature of momentum profits". *Journal of financial economics* 71.2 (2004), pp. 349–380.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. "Generating random correlation matrices based on vines and extended onion method". *Journal of multivariate analysis* 100.9 (2009), pp. 1989–2001.

Linnainmaa, Juhani T and Michael R Roberts. “The history of the cross-section of stock returns”. *The Review of Financial Studies* 31.7 (2018), pp. 2606–2649.

Liu, Laura, Hyungsik Roger Moon, and Frank Schorfheide. *Forecasting with Dynamic Panel Data Models*. Tech. rep. Mimeo, 2016.

McLean, R David and Jeffrey Pontiff. “Does academic research destroy stock return predictability?” *The Journal of Finance* 71.1 (2016), pp. 5–32.

McShane, Blakeley B, David Gal, Andrew Gelman, Christian Robert, and Jennifer L Tackett. “Abandon statistical significance”. *The American Statistician* 73.sup1 (2019), pp. 235–245.

Merton, Robert C. *On the current state of the stock market rationality hypothesis*. 1987.

Mikula, AL, SJ Hetzel, N Binkley, and PA Anderson. “Clinical height measurements are unreliable: a call for improvement”. *Osteoporosis International* 27.10 (2016), pp. 3041–3047.

Novy-Marx, Robert and Mihail Velikov. “A taxonomy of anomalies and their trading costs”. *Review of Financial Studies* 29.1 (2016), pp. 104–147.

Schultz, Paul. “Transaction costs and the small firm effect: A comment”. *Journal of Financial Economics* 12.1 (1983), pp. 81–88.

Senn, Stephen. “A note concerning a selection “paradox” of dawid’s”. *The American Statistician* 62.3 (2008), pp. 206–210.

Stoll, Hans R and Robert E Whaley. “Transaction costs and the small firm effect”. *Journal of Financial Economics* 12.1 (1983), pp. 57–79.

Storey, John D and Robert Tibshirani. “Statistical significance for genomewide studies”. *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445.

Sullivan, Ryan, Allan Timmermann, and Halbert White. “Data-snooping, technical trading rule performance, and the bootstrap”. *The journal of Finance* 54.5 (1999), pp. 1647–1691.

Yan, Xuemin Sterling and Lingling Zheng. “Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach”. *The Review of Financial Studies* 30.4 (2017), pp. 1382–1423.

A. Appendix

A.1. Details on Benjamini-Hochberg (2000)

Benjamini and Hochberg (2000) suggest a graphical approach for estimating $\hat{p}_0$. To derive it, note that the null predictors have uniformly distributed p-values, and thus the following equation should hold:

$$\mathbb{E}(\text{pval}_i|\text{null}_i) = \beta\left[\text{ranking of pval}_i|\text{null}_i\right]. \tag{33}$$

where

$$\beta = \left(\frac{0.5}{p_0 N_{\text{all}}}\right) \tag{34}$$

Benjamini and Hochberg (2000) suggest estimating Equation (33) on a subset of the data is close to the null distribution. This subset is subjective. BH recommend "using a suitable set of the largest p-values." I use a p-value cutoff of 0.20. Storey and Tibshirani (2003) suggest a more rigorous estimator based on fitting a cubic spline.

Figure A.1: Estimation of the Proportion of Nulls Using Benjamini-Hochberg 2000.