Do t-stat Hurdles Need to be Raised?

Andrew Y. Chen
Federal Reserve Board
andrew.y.chen@frb.gov

January 2019*

Abstract

Recent studies argue that multiple testing casts doubt on cross-sectional stock return predictors. I show selective reporting casts doubt on certain multiple testing adjustments. t-stat hurdle adjustments require data on both null and non-null predictors, but predictors close to the null are unreported, leading to weak identification. In contrast, adjustments that target published findings focus on reported predictors and are more strongly identified. Accounting for identification problems in a dataset of 155 cross-sectional predictors, I find the data say little about whether t-hurdles should be raised, but at least 80% of published cross-sectional predictors are true (in-sample, before trading costs).

*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, Bjorn Erakor, Cam Harvey, Laura Liu, Alan Moreira (WFA discussant), Ivan Shaliastovich, Mihail Velikov, and seminar participants at the Federal Reserve Board, George Mason University, the University of Wisconsin, and the 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

The zoo of published cross-sectional stock return predictors has made academicians skeptical. This skepticism goes back at least to Jensen and Bennington (1970), who write, “Given enough computer time, we are sure that we can find a mechanical trading rule which ‘works.’” A growing meta-study literature estimates adjustments that account for this multiple testing problem.\(^1\)

Multiple testing is not the only problem, however. Studies of predictability also exhibit selective reporting: Findings of no predictability are unlikely to be published. Indeed, multiple testing by itself is arguably innocuous, as long as it is performed in a transparent and systematic manner.\(^2\) It is the combination of multiple testing and selective reporting that can lead to the existential doubts forcefully expressed by Harvey, Liu, and Zhu (2016); Harvey (2017); Hou, Xue, and Zhang (2017); and Linnainmaa and Roberts (2018).

In this paper, I show that selective reporting leads not only to doubts about predictability, but also to doubts about multiple-testing adjustments—though some adjustments are more affected than others. Textbook adjustments assume access to an unbiased set of hypothesis tests (Efron 2012). Selective reporting, then, implies that the unreported tests need to be extrapolated. This extrapolation leads to questions of identification: Do the reported tests provide enough information to pin down the adjustments? Or do alternative extrapolations, with very different implications, lead to the same observable data? Which adjustments are strongly identified, and what do they tell us? To answer these questions, I apply a variety of multiple testing adjustments to large collections of published cross-sectional stock return patterns, and study their sampling distributions via bootstrap.

I find that adjustments to t-stat hurdles are weakly identified. Intuitively, t-hurdles apply to both published and unpublished results. The unpublished results must be extrapolated, and there are many ways to extrapolate, leading to weak identification. In contrast, adjustments that target only published patterns focus on reported data, and are more strongly identified. Overall, the published data say little about whether t-stat hurdles should be raised, but show confi-

---

\(^1\)Meta-studies of multiple testing adjustments include Harvey, Liu, and Zhu (2016); McLean and Pontiff (2016); Jacobs and Müller (2017b); Linnainmaa and Roberts (2018); Chen and Zimmermann (2018); and Xiong and Pelger (2019).

t-hurdle adjustments come in two forms: non-parametric and parametric. While parametric adjustments are weakly identified, non-parametric adjustments are better described as not identified. Non-parametric adjustments, such as the Benjamini and Yekutieli (2001) algorithm, place an upper bound on the t-hurdle if they are applied to an unbiased set of tests. Since published findings exhibit reporting bias, it is not clear if applying Benjamini and Yekutieli (2001) can identify even an upper bound. While one can argue that assuming all tests are observed leads to a lower bound (Harvey et al. 2016), the resulting estimate is a lower bound on an unknown upper bound, which is uninformative. Any t-hurdle is consistent with any lower bound on an unknown upper bound.

A related logic applies to parametric t-hurdle estimates. I illustrate this by studying Harvey, Liu, and Zhu’s (2016) (HLZ’s) parametric model, which explicitly accounts for selective reporting. HLZ fit their model to moments of hand-collected t-stats, and find that the proportion of nulls among all factors (published and not) is 44%. This null proportion leads to t-stat hurdles of 2.27 and 2.95 for a false discovery rate (FDR) of 5% and 1%, respectively.

I present an alternative parameterization of HLZ’s model that differs primarily in assuming the proportion of nulls is zero. Thus, the alternative model implies that all factors are true, and even a t-hurdle of 0 ensures an FDR less than 1%. With zero nulls, the alternative model has starkly different predictions about the density of t-stats. In particular, the alternative implies that the density of t-stats less than 2.0 is small, while HLZ’s baseline implies a large density. t-stats below 2.0, however, are never observed, and thus the data required for identification is missing. Indeed, both models fit the data well, suggesting that the true t-hurdle may lie anywhere between 0 and 2.27 (for an FDR of 5%) or 0 and 2.95 (for an FDR of 1%).

While this illustration is suggestive, a formal analysis requires both standard errors and a study of alternative model assumptions. I conduct a formal analysis by bootstrapping SMM estimates of variants of HLZ’s model. As SMM targets, I use data from Chen and Zimmermann’s (2018) publically-available replications of 155 cross-sectional predictors. This procedure also allows me to simultane-

---

3Following standard academic writing, I use present tense throughout the paper. But “were” true may be more accurate. I can only say the predictors are true in-sample, and before transaction costs.

4Harvey, Liu, and Zhu (2016) do not provide standard errors.

5This dataset is available at http://sites.google.com/site/chenandrewy/code-and-data.
ously study the identification of other multiple testing statistics such as adjusted expected returns.

The baseline bootstrap finds that standard errors for t-hurdles are large, at about 1.0, and point estimates are well-within one-standard error of their classical 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.

In contrast, adjustments to mean returns of published predictors are strongly identified. Bias-adjusted returns are only 16% smaller than simple sample mean returns, with a small standard error of 3 percentage points. Importantly, I find a similar point estimate and standard error using 10 alternative modeling assumptions. Similarly, the FDR for published predictors is consistently small, with a point estimate of 6% and a standard error of 6 percentage points. These small effects on published predictors are consistent with other estimates in the literature. McLean and Pontiff (2016) finds an upper bound of 26% on the adjustment to sample mean returns. And while HLZ do not report it, I show that their parametric estimates imply the FDR among published factors is about 5%.

The idea that t-hurdles do not need to be raised in the presence of multiple testing is highly counterintuitive. A simple logic suggests 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.

A more sophisticated logic, however, shows that t-hurdles might be lowered. Multiple tests provide more information than a single test. This information allows for estimation of the proportion of nulls, and if the estimated proportion is small, then t-hurdles should be lowered to achieve the desired rate of false discoveries. Note that if the researcher has access to only one hypothesis test, there is no way to estimate the proportion of nulls. Indeed, the concept does not even make sense in classical single testing. It is only by “borrowing strength” from other tests (to use John Tukey’s colorful term) that one can learn about the proportion of nulls.

---

6 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 1995 algorithm does not, as it assumes that the proportion of nulls = 1. In Benjamini and Hochberg’s (2000) followup to their 1995 paper they recommend estimating the proportion of nulls and find the t-hurdle may be lowered. Many statisticians have continued to develop on the idea of estimation the null proportion (Storey and Tibshirani 2003, Genovese and Wasserman 2004, Efron 2004, among others).
portion of nulls. This idea of learning from other tests is a common theme in Efron’s (2012) textbook on large scale hypothesis testing.

While I find skepticism of cross-sectional predictability is not supported by statistics, there is a compelling economic reason to be skeptical. As my paper is on multiple testing effects in published patterns, I follow the original publications in their weighting methods, and nearly all predictability is documented using equal-weighted portfolios or equal-weighted regressions (McLean and Pontiff 2016; Chen and Zimmermann 2018). Relatedly, 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 et al. 2004; Novy-Marx and Velikov 2016). Indeed, Chen and Velikov (2018) find that post-publication, trading costs account for essentially all of the remaining returns. Put another way, there is little reason to be skeptical of stock market predictability that is both temporary and costly to implement.

Moreover, my formal results directly address only cross-sectional predictors, as the Chen and Zimmermann (2018) data do not contain factors that do not predict returns cross-sectionally, nor does it contain aggregate predictors. The Harvey, Liu, and Zhu (2016) data does represent these other asset pricing patterns, but they are mixed up with cross-sectional predictors, which make up a large share of the data. If an estimator were to “borrow strength” from only linear factor tests or only aggregate predictors, the estimated bias could be quite different.

The literature on multiple testing in asset pricing (a.k.a. data mining, data snooping) is large. 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 suggestive evidence about multiple testing effects in published findings.

Studies that use published asset pricing tests include McLean and Pontiff (2016); Harvey, Liu, and Zhu (2016); Jacobs and Müller (2017b); Chen and Zimmermann (2018); Linnainmaa and Roberts (2018); and Xiong and Pelger (2019). My paper is distinct in that it estimates t-hurdles and adjusted returns in the same framework, and thus helps reconcile the conflicting results from this literature. My paper is also the first to demonstrate the identification challenges that come from selective publication.
2. A Primer on Multiple Testing

As multiple testing statistics are new to many readers, I begin with a brief primer that draws heavily on Chapters 1 and 4 of Efron (2012). The primer aims to provide the background necessary to understand identification (Sections 3 and 4). It also explains how t-stat hurdles may be lowered (Section 2.2).

2.1. How to Adjust Hypothesis Tests for Multiple Testing

Suppose we use a predictor to construct a portfolio, and obtain $T$ monthly returns. If we want to check for "statistical significance," we take the mean return and divide by the standard error to get a t-stat. Assuming that $T$ is large, traditional single-testing stats says that this portfolio is predictor if the t-stat is larger than 1.96.

In finance, we apply this procedure so often that we often forget that there is a model in the background. This model says

\begin{align}
\bar{r}_i &= \mu_i + \epsilon_i \\
\epsilon_i &\approx N(0, SE_i) \\
i &= 1
\end{align}

where $\bar{r}_i$ is the sample mean return, $\mu_i$ is the latent expected return, $\epsilon_i$ is sampling noise, and $SE_i$ is the standard error of the sample mean. We consider only a single predictor ($i = 1$) but will soon allow for many. Under these assumptions, the 1.96 hurdle comes from the

$$Pr(|t| > 1.96|\mu = 0) = 5\%.$$  \hspace{1cm} (4)

where $t \equiv \bar{r}/SE$. In words, a t-stat is significant if it is unlikely assuming the null model $\mu = 0$.

Now suppose we have monthly returns based on 500 predictors. And for now assume away any selection (reporting) bias concerns. We'll return to reporting issues in Section 3.

With 500 predictors, there are multiple ways to ask about significance. For example, we may want to check for significance among predictors with t-stats $> 1.96$. Alternatively, we may want find a subset of predictors that is significant. Re-
Regardless, all questions end up looking in the tails of the distribution. And relative to the single test setting, we need to account for three new issues: luck, selection, and the information in the other tests.

The false discovery rate (FDR) accounts for all three of these issues. The traditional definition of the FDR comes from Benjamini and Hochberg (1995). But the Efron and Tibshirani (2002) FDR is often equivalent, easier to interpret, and maps closely to traditional single testing. Indeed, the Efron-Tibshirani FDR is simply the mirror image of the single-testing criterion Equation (4), that is

$$FDR_{ET}( |t_i| > 1.96) \equiv \Pr(\mu_i = 0 | |t_i| > 1.96) \quad (5)$$

where $FDR_{ET}( |t_i| > 1.96)$ is the Efron-Tibshirani FDR for the subset of predictors with t-stats $> 1.96$. In words, the FDR is the probability predictor $i$ is null, assuming that $|t_i| > 1.96$. Note that classical significance test is often misinterpreted this way. Efron and Tibshirani (2002) show that $FDR_{ET}$ leads to the equivalent FDR control algorithm as Benjamini and Hochberg (1995), but does not require the assumption of independence.

To understand $FDR_{ET}$, it helps to rewrite (5) using Bayes rule:

$$FDR_{ET}( |t_i| > 1.96) = \frac{\Pr(|t_i| > 1.96 | \mu_i = 0) \Pr(\mu_i = 0)}{\Pr(|t_i| > 1.96)} = \frac{5\%}{\Pr(|t_i| > 1.96)} \Pr(\mu_i = 0). \quad (7)$$

Efron and Tibshirani suggest estimating the denominator with the empirical distribution

$$\hat{FDR}_{ET}( |t_i| > 1.96) = \frac{5\%}{\frac{\text{# |t-stats| > 1.96}}{\text{Total # of t-stats}}} \hat{\Pr}(\mu_i = 0). \quad (8)$$

Equation (8) shows how the FDR controls for luck, selection, and the information in other tests. Luck is controlled because $FDR_{ET}$ increases as the total number of tests increases. Selection is controlled explicitly in the definition of $FDR_{ET}$. Last, the $FDR_{ET}$ accounts for information in other tests through the estimated proportion of nulls $\hat{\Pr}(\mu_i = 0)$. This critical parameter scales the entire expression on the RHS, and in principle can range anywhere between 0 and 1.
2.2. Why the t-hurdle May be Lowered Under Multiple Testing

A surprising aspect of Equation (8) is that the FDR among $|t\text{-stats}| > 1.96$ may be less than its traditional single testing counterpart of 5%. Whether the FDR is greater or less than its traditional counterpart depends entirely on whether the $\hat{\Pr}(\mu_i = 0)$ is greater or less than the proportion of $|t\text{-stats}| > 1.96$. Consequently, if one uses the FDR to adjust the traditional t-stat hurdle (e.g. Harvey et al. 2016), the resulting hurdle need not be greater than 1.96.

How can t-hurdles not be raised after adjusting for multiple testing? The answer is that multiple tests provide information that is not available in a single test. This idea is known as “borrowing strength” (Tukey 1969) or “learning from the experience of others,” and is a common theme in multiple testing statistics (Efron 2012).

What one learns from multiple tests depends on the results of the multiple tests. This data-dependence is illustrated in Figure 1. The left panels plot the distribution of t-stats from different model simulations. In the top left, the results are very close to the null distribution. This suggests that most tests are truly false, in other words $\Pr(\mu_i = 0)$ is close to 1.0. Indeed, as the data is simulated we know the true proportion of nulls is 95%. Thus, the traditional t-stat hurdle must be raised to 2.71 to achieve an FDR of 5%.

But the middle left panel shows that the learning can turn out quite differently. In this panel, the t-stats are very far from the null. Indeed, only 5% of t-stats lie between -2 and +2. Correspondingly, only 5% of predictors in this simulation are null ($\Pr(\mu_i = 0) = 5\%$). Thus, even a t-stat hurdle of 0 can ensure an FDR of 5%. In other words, selecting a predictor at random from these tests would already imply that only 5% are false.

This radical result, that statistical thresholds might be lowered in response to multiple testing, was expressed in the original draft of Benjamini and Hochberg (1995), written in 1989. After many years of rejections, the authors shifted to a more conservative formula, recommending assuming $\Pr(\mu_i = 0) = 1$ and thus deriving only an upper bound on the FDR. The resulting implication is that t-stat hurdles can only be raised. This upper bound is what is estimated in the celebrated Benjamini and Hochberg (2000) algorithm, though the original recommendation of estimating $\Pr(\mu_i = 0)$ was finally published in Benjamini and
Figure 1: Learning from Multiple Tests. Each row shows results from a different model simulation. In the top row, 90% of predictors are null (Pr(μ\(_i\) = 0) = 0.90). In the other two rows, 5% are null. The middle row assumes non-null predictors are not very dispersed, while the bottom row assumes a high dispersion. The left panels show the distribution of t-stats in each model (bars), along with the null (t\(_i\) ∼ N(0, 1), solid line) for comparison. The panels on the right illustrate shrinkage estimates of expected returns (E(μ\(_i\)|\(\bar{r}_i\))). The figures illustrate how an econometrician can learn from multiple tests, and what one learns depends on the results of the test.

Hochberg (2000).\(^7\)

\(^7\)This history is described in Benjamini (2010).
2.3. How to Adjust Magnitudes for Multiple Testing

In single testing, we often estimate the “magnitude” of a predictor by the sample mean return $\bar{r}_i$ on a portfolio based on the predictor. Once again, researchers in finance apply this procedure so often that we forget the underlying model (Equations (1)-(3)). Implicitly, the parameter $\mu_i$ is the magnitude, and we estimate it using $\hat{\mu}_i = \bar{r}_i$. $\hat{\mu}_i$ is unbiased because

$$
E(\hat{\mu}_i) = E(\bar{r}_i) \tag{9}
= \mu_i + E(\epsilon_i). \tag{10}
$$

Now consider the 500 predictor case ($i = 1, 2, ..., 500$). Suppose we want to examine only predictors with large t-stats, say $t_i > 1.64$ (corresponding to a 5% one-tailed test). Then the traditional $\hat{\mu}_i = \bar{r}_i$ is biased:

$$
E(\hat{\mu}_i | t_i > 1.64) = E(\bar{r}_i | t_i > 1.64) \tag{11}
= E(\mu_i | t_i > 1.64) + E(\epsilon_i | t_i > 1.64). \tag{12}
$$

Intuitively, selecting for large $t_i$ selects for lucky results.

Note that under multiple testing $\mu_i$ becomes a random variable. In contrast, classical statistics treats $\mu_i$ as a parameter. By treating $\mu_i$ as a random variable, we can make inferences accounting for selection that naturally occurs in multiple testing. That is, we want to estimate the conditional expectation

$$
\hat{\mu}_{shrink}(t_i > 1.64) = E(\mu_i | t_i > 1.64). \tag{13}
$$

I label this estimator with the subscript “shrink” because the conditional expectation typically implies shrinking toward the grand mean.

Evaluating the RHS of (13) typically requires an explicit model. To focus on intuition, consider on a simple case with i.i.d. homoskedastic samples, and normally distributed $\mu_i$ centered around 0. In this case, we have an elegant formula\(^8\)

$$
\hat{\mu}_{shrink}(t_i > 1.64) = \left(1 - \frac{1}{\text{Var}(t_i)}\right) \sum_{i: t_i > 1.64} \frac{\bar{r}_i}{\#[t_i > 1.64]} . \tag{14}
$$

\(^8\)Use standard bivariate normal formulas and the fact that $\text{Var}(r_i) = \text{Var}(\mu_i) + \text{SE}_i^2$. 


In the above expression, we shrink the average (selected) sample mean \( \sum_{i: t_i > 1.64} \frac{\bar{r}_i}{\#(t_i > 1.64)} \) toward zero, where the amount of shrinkage is given by \( \frac{1}{\text{Var}(t_i)} \). Intuitively, if the many t-stats are standard-normal, then \( \frac{1}{\text{Var}(t_i)} = 1 \), shrinkage is 100%, and the multiple-testing adjusted returns are all zero (the mean of \( \mu_i \) is assumed to be zero). In contrast, if t-stats have a very large dispersion, that indicates that \( \mu_i \) is dispersed, and thus \( \bar{r}_i \) is an informative signal about the latent magnitude \( \mu_i \).

The right panels of Figure 1 demonstrate this intuition. The top two panels show multiple tests with very little dispersion. The variance of t-stats is close to 1 (top left), and thus shrinkage is close to 100% (top right, solid line). Note that in this case, the single testing estimate (dotted line) is an extremely upward biased of the truth (x’s), but the shrinkage estimate stays close to the truth.

In contrast, the bottom two panels illustrate a multiple tests with large dispersion. The variance of t-stats is an order of magnitude larger than 1, and thus shrinkage is minimal (right panel).

Shrinkage is more subtle in the middle panels. Here the model deviates strongly from the normal model used to derive the elegant shrinkage Equation (14). Instead, we have a bimodel distribution, and thus the “shrinkage” estimate jumps quickly from one mode to the next as the sample mean return increases.

### 3. Selective Reporting and Identification Problems

In Section 2, we assumed away any selection bias. But when we study tests that are reported in academic papers, we must account for selection. This selection problem implies that (1) simple non-parameteric estimates of multiple testing adjustments are not informative, and (2) parametric estimates may run into identification problems.

#### 3.1. Why Non-Parametric t-hurdles are Not Identified

The multiple testing statistics presented in Section 2 assumed away selective reporting problems. To understand selection, consider two example datasets used in Efron’s (2012) multiple-testing textbook:

1. Expression levels for 6033 genes were obtained for 50 control and 52 cancer patients.
2. The ability of a chemical to bind to 16,822 genes was measured.

And compare to two influential meta-studies on publication effects in asset pricing:

a. “We choose a subset of papers that we suspect are in review at top journals... ...We focus on 313 articles, among which are 250 published articles... ...We obtain t-statistics for each of the 316 factors.” (Harvey et al. 2016)

b. “We limit ourselves to studies... ...in which the null of no return predictability is rejected at the 5% level... ...we examine 97 cross-sectional relations.” (McLean and Pontiff 2016)

In the textbook examples, it’s reasonable to assume the outcomes were not used to select the sample: the genes in examples 1 and 2 were not selected based on their expression levels or binding ability, the factors and predictors in examples a and b were almost certainly selected based on their performance.

Selective reporting creates a problem for non-parametric hypothesis testing adjustments. Non-parametric adjustments Benjamini and Yekutielii (2001) assume the proportion of nulls = 1, and thus put an upper bound on the true multiple testing statistic. But this upper bound requires access to an unbiased set of tests, and thus it is not at all clear that even the upper bound can be identified in published findings.

One can argue that assuming all tests are observed leads to a lower bound (Harvey et al. 2016). But a lower bound on an unknown upper bound is uninformative. Any t-hurdle satisfies their lower bound on an upper bound.

In principle, non-parametric identification with selective reporting is possible, as shown by Andrews and Kasy (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 and estimate using GMM.

Bootstrapping random empirical tests à la Harvey and Liu (2018) is not feasible either, because the selection of published data is complex. 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

---

9Harvey 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).
psychological motivations, at least most of the time.\textsuperscript{10} To properly bootstrap a
model of publication, one must have a process for generating all of these additional items from empirical data, as well as a method for estimating this complicated selection process.

Adding structure and extrapolating the small t-stats is one way to address this
selective reporting problem. This approach is used in HLZ’s model with correla-
tions, Chen and Zimmermann (2018), and is also the method I use. Extrapolation,
however runs into questions of identification, which I discuss shortly.

3.2. Identification of a Parametric Multiple Testing Model

To understand the identification problems under selective reporting, it helps
to first examine how a parametric model of multiple testing is identified. To do
this, consider the selection-free version of Harvey, Liu, and Zhu’s (2016) model
with correlations

\begin{align*}
\text{tests } i = 1, 2, \ldots, N_{\text{all}} \quad (15) \\
\mu_i &\sim \begin{cases} 
\delta(0) & \text{with prob } p_0 \\
\text{Exp}(\lambda) & \text{otherwise}
\end{cases} \\
\bar{r} &\sim N(\mu, \Sigma) \\
\Sigma_{i,j} &= \begin{cases} 
\text{SE}^2 & \text{for } i = j \\
\rho \text{SE}^2 & \text{for } i \neq j
\end{cases} \\
\end{align*}

where $\delta(0)$ is the distribution with a point mass at 0, $\bar{r} = [\bar{r}_1, \ldots, \bar{r}_{N_{\text{all}}}]'$, $\bar{\mu} = [\mu_1, \ldots, \mu_{N_{\text{all}}}]'$, and $\Sigma_{i,j}$ are the elements of the matrix $\Sigma$. HLZ’s model contains
an additional selection restriction (t-stats must be larger than 1.96), but for now
let’s examine the simpler model of pure multiple testing. I’ll return to the selec-
tion issue in Section 3.3.\textsuperscript{11}

In short, the model augments the implicit single testing model (Equations
(1)-(3)) to allow for multiple tests and correlated returns. The parameters gov-
erning $\mu_i$ then allow computation of the multiple testing adjustments described
in Section 2, given valid estimates. This estimate can be done by maximum like-

\textsuperscript{10}Harvey (2017) provides a few counterexamples of published predictors that are arguably poorly motivated.

\textsuperscript{11}HLZ’s is written in terms of monthly returns rather than sample mean returns. To derive my formulation, use $\text{Cov}(\bar{r}_i, \bar{r}_j) = T^{-2} \sum_t \text{Cov}(r_{i,t}, r_{j,t}) = \rho \sigma^2 / T$ and $\text{SE} = \sigma / \sqrt{T}$. 

12
lihood (Chen and Zimmermann 2018) or GMM (Harvey et al. 2016), but either way the identification is the same and comes from properties of the model.

The two key parameters the proportion of nulls $p_0$ and the mean return of non-nulls $\lambda$. These two parameters are closely linked to the FDR and shrinkage adjustments, respectively. Indeed $p_0$ is exactly the FDR among all tests, and $\lambda$ is the shrinkage-adjusted expected return for non-nulls.

The proportion of nulls $p_0$ is identified by the density of t-stats near 0. This identification is illustrated in the top panels of Figure 2, which illustrate how the distribution of t-stats implied by Equations (15)-(18) vary with $p_0$. As $p_0$ increases, the density of model-implied t-stats (solid line) near 0 also increases. In this example, a large $p_0$ of 0.82 is required to fit the data (bars). Intuitively, a large $p_0$ implies many t-stats $\sim N(0,1)$, and thus many t-stats should be between -2 and +2.

**Figure 2: Identification of a Parametric Multiple Testing Model.** I fit the selection-free version of HLZ’s model with correlations (Equations (15)-(18)) to simulated data. In the top row, all parameters are fixed except for $p_0$. In the bottom row, all parameters are fixed except for $\lambda$. $p_0$ is identified by the density of t-stats near 0, $\lambda$ is identified by the right tail of t-stats.
In contrast, the magnitude of true predictors is identified by the right tail of t-stats. This is illustrated in the bottom panels of Figure 2. Like $p_0$, $\lambda$ does affect the density of t-stats near zero, particularly when $\lambda$ is small. But unlike $p_0$, $\lambda$ has large effects on the density of t-stats far away from 0: as $\lambda$ increases, the right tail implied by the model (solid line) gets fatter. Intuitively, only $\lambda$ has an effect on the density of t-stats far from the null.

These results suggest that some multiple testing adjustments may be more strongly identified than others. Adjustments that depend strongly on $p_0$ are weakly identified, as the data required for identifying $p_0$ is unreported. In contrast, adjustments that depend strongly on $\lambda$ have a chance of being strongly identified, as $\lambda$ can be found in the right tail.

3.3. Observationally Equivalent Parameterizations of Harvey, Liu, and Zhu’s Model with Correlations

To make the identification issues concrete, consider the identification of HLZ’s model with correlations. This model consists of the parameteric model (Equations (15)-(18)) and one additional equation:

$$Test_i \text{ is observed iff } t_i > 1.96. \quad (19)$$

This additional equation accounts for selective reporting of factor results with a simple cutoff rule.$^{12}$

HLZ estimate their model on moments of reported t-stats, reprinted the “adjusted data” row in Table 1. The data is adjusted to account for the fact that marginal t-stats are less likely to be reported (see Table caption). This adjustment is small.

Their SMM estimation arrives at the following parameter values: $\rho = 0.2$, $SE = 15/\sqrt{12} \times 240$, $N_{all} = 1378$, $p_0 = 0.444$, 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.

$^{12}$It’s not clear from Harvey et al. (2016) if their selection equation allows for negative t-stats ($t_i < -1.96$). Allowing for only positive t-stats is consistent with their assumption that $\mu_i \geq 0$, as such a restriction could only come if factors are signed based on some kind of theory that implies $\mu_i \geq 0$. If factors are signed by theory, then $t_i < -1.96$ will not be reported. Section 5.6 examines...
Table 1: Moments of t-stats from the Harvey, Liu, and Zhu (2016) Factor Data and Their Parametric Model

“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 $N_{\text{all}}$ to 900, $p_0$ to 0, and $\lambda$ to 0.505%.

<table>
<thead>
<tr>
<th></th>
<th># of t-stats</th>
<th>20</th>
<th>50</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted Data</td>
<td>353</td>
<td>2.39</td>
<td>3.16</td>
<td>6.34</td>
</tr>
<tr>
<td>HLZ baseline</td>
<td>334</td>
<td>2.40</td>
<td>3.37</td>
<td>6.62</td>
</tr>
<tr>
<td>Alternative</td>
<td>349</td>
<td>2.35</td>
<td>3.27</td>
<td>6.14</td>
</tr>
</tbody>
</table>

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%.

Now consider an alternative parameterization. Begin with HLZ’s baseline estimate, but change $N_{\text{all}}$ to 900, $p_0$ to 0, and $\lambda$ 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.

Can the data reject this idea that all academic factors are all extraordinary? Figure 3 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.

**Consistent Implications Across the Two Models** Despite their different t-hurdles, the models have similar implications for adjustments that target only observed factors. Figure 3 illustrates this similarity by plotting the distribution of the expected returns $\mu_i$, but for only observed factors.

Both models imply a very similar distribution of expected returns. The HLZ features a spike near 0 that is missing in the alternative model, but both models models which allow $\mu_i < 0$. 
Figure 3: 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 $\mu_i$ for observed factors.

Panel A: All Factors’ t-stats

Panel B: Observed Factors’ Expected Returns

imply that the bulk of the distribution of $\mu_i$ resides between 0.50% and 1.00% per month. Indeed, summary statistics for observed $\mu_i$ are very similar across models. The mean $\mu_i$ is 0.93% in the HLZ baseline, just a touch above the 0.91% in the alternative model. Similarly, the median observed $\mu_i$ are 0.82% and 0.81% in the two models, only a bit smaller than the median observed sample mean return of 0.88%.$^{13}$

Moreover Panel B shows that both models imply a similar FDR among observed factors. Without the plot, we already know the alternative model implies this FDR is 0, as $p_0 = 0$. But even though the HLZ model implies the FDR is 44%

$^{13}$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 $15/\sqrt{(12)(240)}\%$. 

16
among all factors, among observed factors their model implies an FDR of only 5%. This is seen in Panel B by the fact that the spike in the HLZ distribution around 0 rises to about 5%.

This small FDR is surprising, given Harvey, Liu, and Zhu’s (2016) conclusion that most published findings are likely false. But the small FDR is a clear implication of their baseline estimate. This can be seen by calculating the Efron-Tibshirani FDR$_{ET}$ (Equation (8)) among observed factors using numbers reported in HLZ’s Table 5:

$$\text{FDR}_{ET}(t_i > 1.96) = \frac{0.025(0.444)}{353/1378} = 4.33\%.$$  

(20)

where 0.444 is the proportion of nulls, 353 is the number of observed factors from HLZ’s target moments, and 1378 is HLZ’s estimate of the total number of factors.\footnote{It is not clear from Harvey, Liu, and Zhu (2016) if selection in their model allows for negative t-stats. Allowing for t-stats $<-1.96$ leads to a doubling of the FDR to 8.7%, but still implies that the vast majority of published findings are true.} While the Benjamini-Hochberg FDR cannot be calculated by hand, Efron and Tibshirani (2002) show that the two FDRs are asymptotically equivalent as long as a mixing conditions holds. Indeed, simulations of HLZ’s model find a very similar FDR.\footnote{Code for simulations is available at http://sites.google.com/site/chenandrewy/code-and-data.}

The similarity across models comes from the large expected return of non-nulls $\lambda$ used in both models. This large $\lambda$ is required to fit the fat right tail of observed t-stats (Table 1, Panel A).

4. **Main Results: Identification of Multiple Testing Statistics in Published Cross-Sectional Predictors**

The HLZ factor data illustrates the importance of identification under selective reporting. But a clear understanding of this issue requires a formal estimate of sampling variation. This section provides a formal estimate. I bootstrap SMM estimates of a multiple testing model on a large dataset of published cross-sectional predictors. Sections 4.1-4.4 describe the methods and parameter estimates. Section 4.5 shows the main results.
4.1. 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 et al.’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 (??)).

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.

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
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).

<table>
<thead>
<tr>
<th>Panel A: Measures of Central Tendency</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
</tr>
<tr>
<td>0.038</td>
</tr>
<tr>
<td>(0.009)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Comparison of Mean Correlation with the Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>0.033</td>
</tr>
<tr>
<td>0.050</td>
</tr>
<tr>
<td>0.200</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: Measures of Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>stdev</td>
</tr>
<tr>
<td>0.308</td>
</tr>
<tr>
<td>(0.014)</td>
</tr>
</tbody>
</table>

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.\textsuperscript{16} 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.\textsuperscript{17}

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 et al.’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, 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 multiple testing effects in the original papers, and do not conflate multiple testing with investor learning or other effects.

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.

\textsuperscript{16}The Harvey, Liu, and Zhu (2016) factors include 113 “common” factors and 202 “characteristics,” factors. HLZ define these categories 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.”

\textsuperscript{17}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.
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 Zimmern (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.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>20</th>
<th>50</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-stat</td>
<td>2.12</td>
<td>3.40</td>
<td>8.05</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.19)</td>
<td>(0.91)</td>
</tr>
<tr>
<td>sample mean return (%)</td>
<td>0.38</td>
<td>0.60</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>standard error (%)</td>
<td>0.10</td>
<td>0.17</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
</tbody>
</table>

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 heteroskedasticity in Section 4.2.

4.2. A Model with Heterogeneous Correlations and Heteroskedasticity

To model multiple testing and selective reporting, 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. Indeed their model is a special case of Andrews and Kasy's (ming) non-parametric model. I describe HLZ's model in Section 3.3, Equations (15)-(19). 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 et al. 2009, for example). Nevertheless, Section 5.1 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 et al. 2013), but that correlations display some dispersion. Thus, I replace the constant $\rho$ in Equation (18) with $\rho_{i,j}$, where $\rho_{i,j}$ is constructed from partial correlations that follow a symmetric beta distribution

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

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 (21), the full correlation matrix can be calculated using textbook formulas (Kendall and Yule 1961, Anderson 2003). This approach to generating random correlation matrices follows methods from the multivariate analysis literature (Joe 2006, Lewandowski et al. 2009).

Equation (21) may appear foreign, as random correlation matrices are rarely studied 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 $155 \times 155$ published correlation matrix would imply that portfolios from different blocks are independent. In contrast, Equation (21) implies that all $N_{\text{all}}$ portfolios are correlated, regardless of the size of $N_{\text{all}}$.

Moreover, Equation (21) provides a parsimonious model that can fit the published data well (Figure 4). The main alternative methods for generating a random correlation matrix from the engineering literature are based on random eigenvalues and random orthogonal matrices (Holmes 1991, for example), and are significantly more complex than Equation (21). Simple ad-hoc rejection al-
algorithms, do not work, as they almost always lead to non-PSD matrices (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 SE in Equation (18) with a log-normally distributed standard error

$$\log(SE_i) \sim N(\mu, \sigma), \text{ i.i.d.}$$

The assumption is a simple way to model heterogeneous volatilities and sample lengths while ensuring that standard errors are positive. Moreover, Figure 4 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 5).

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

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

where \(t_{\text{min}}\) is the minimum t-stat for publication, \(t_{\text{good}}\) is the value below which a t-stat is considered marginal, and \(\omega\) 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 \(\omega\). Equation (23) is equivalent to this data modification, and allows me to estimate their model without altering the data.

Formally, Equation (23) 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 (ming) show \(k\) is not identified, despite the fact that the distribution of \(\mu_i\) 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.
4.3. **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 5), 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 Zimmermann (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 4. 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 5).

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

\[ \hat{\theta} = \arg\min_{\theta \in \Theta} \left\{ \sum_{k=1}^{N_{\text{sim}}} \sum_{i=1}^{9} w_{t,i} \left[ Q_{\text{model},t,i}(\theta) - Q_{\text{data},t,i} \right]^2 \\
+ \sum_{k=1}^{N_{\text{sim}}} \sum_{i=1}^{9} w_{r,i} \left[ Q_{\text{model},r,i}(\theta) - Q_{\text{data},r,i} \right]^2 \\
+ \sum_{k=1}^{N_{\text{sim}}} \sum_{i=1}^{9} w_{SE,i} \left[ Q_{\text{model},SE,i}(\theta) - Q_{\text{data},SE,i} \right]^2 \\
+ \sum_{k=1}^{N_{\text{sim}}} w_{F} \left[ F_{\text{model,marg}}(\theta) - F_{\text{data,marg}} \right]^2 \right\} \]  

(24)

where \( N_{\text{sim}} \) is the number of simulations, \( Q_{\text{model},t,i}(\theta) \) is the \( i \)th decile (10\( \times \)ith percentile) t-stat from the model given parameter \( \theta \), 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 \( \omega \). 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 \( \Theta \), I allow \( \lambda, \mu_{\sigma}, \) and \( \sigma_{\sigma} \) to be unrestricted on the
positive portion of the real line, but restrict $p_0$ and $\omega$ to discrete points. Specifically, I only allow

$$p_0 \in \{0, 0.05, 0.10, \ldots, 0.95\} \quad (25)$$

$$\omega \in \{1/3, 1/2, 2/3\}.$$  

Restricting these parameters ensures that the numerical optimizer does not get stuck in a flat region. As illustrated in Section 3.3, the SMM objective can be very flat if $p_0$ and $\lambda$ are altered simultaneously. A similar issue arises with $\omega$ and $\lambda$.

The restrictions in Equation (25) allow me to do a two-stage optimization: I first use a quasi-newton method to optimize the other parameters ($\lambda, \mu, \sigma$) given $p_0$ and $\omega$ chosen from Equation (25). 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 $\omega$.

Unlike HLZ, I do not target $N_{\text{all}}$ in my SMM objective. As shown by Andrews and Kasy (ming), the selection function (23) is only identified up to scale, and thus $N_{\text{all}}$ is not identified. Instead, I simply choose a large value for $N_{\text{all}}$ and $N_{\text{sim}}$, 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 $N_{\text{all}}$ 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).

### 4.4. 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 3.3. The proportion of nulls $p_0$ is weakly identified, while the expected return for non-null predictors $\lambda$ is pinned down well. These find-
ings lead to the weak identification of t-hurdles and the strong identification of adjusted expected returns in Section 4.5.

Table 4: Parameter Estimates for a Model of Multiple Testing and Selective Reporting.

I estimate a model of multiple testing with selective reporting (Section 4.2) on 155 long-short portfolios from Chen and Zimmermann (2018) (Tables 2 and 3) using SMM (Section 4.3). 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 $\lambda$ is strongly identified.

Panel A: Calibrated Parameters

<table>
<thead>
<tr>
<th>Value</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\min}$</td>
<td>1.50</td>
</tr>
<tr>
<td>$t_{\max}$</td>
<td>2.57</td>
</tr>
<tr>
<td>$b$</td>
<td>2.86</td>
</tr>
</tbody>
</table>

Panel B: SMM Estimates

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Point</th>
<th>Bootstrapped Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of Null</td>
<td>$p_0$</td>
<td>0.45</td>
</tr>
<tr>
<td>Unpub Marginal t-stats</td>
<td>$\omega$</td>
<td>0.50</td>
</tr>
<tr>
<td>S.D. $\mathbb{E}(\bar{r}_i)$ for Non-Null</td>
<td>$\lambda$</td>
<td>0.38</td>
</tr>
<tr>
<td>Mean Std Err</td>
<td>$\mathbb{E}(\text{SE}_i)$</td>
<td>0.23</td>
</tr>
<tr>
<td>S.D. of Std Err</td>
<td>SD($\text{SE}_i$)</td>
<td>0.13</td>
</tr>
</tbody>
</table>

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 $\omega$. The 50% C.I. cannot rule out any of the three values for $\omega$ 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 $\omega$ 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 $\lambda$ 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 $E(SE_i)$ and the standard deviation of standard errors $SD(SE_i)$ for ease of interpretation.\(^{18}\)

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.

Panels A - C of Figure 4 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 between mean returns and standard errors among all predictors (Equations (16) and (22)). Thus, selective reporting is enough to generate the observed correlation.

Overall, the strong fit shown in Figure 4 suggests that the model specification is good.

\(^{18}\)The lognormal distribution implies that the mean and standard deviation of all standard errors are given by

\[
E(SE_i) = \exp(\mu_\sigma + \sigma_\sigma^2/2) \tag{26}
\]

\[
SD(SE_i) = \sqrt{\exp(\sigma_\sigma^2) - 1} \exp(2\mu_\sigma + \sigma_\sigma^2). \tag{27}
\]
Figure 4: Model Fit: Point Estimate. I simulate the model (Section 4.2) 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.

4.5. The Bootstrapped Distribution of Multiple Testing Statistics

With parameter estimates in hand, I can finally bootstrap estimates of multiple testing statistics. I first define the statistics, and then present the main result in Figure 5.
**Multiple Testing Statistics Definitions** I examine four kinds of multiple testing statistics: (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 $t_{\text{hurdle}}$ as

$$
FDR(t_{\text{hurdle}}) \equiv \mathbb{E} \left[ \frac{\#\{\text{null}_i : t_i \geq t_{\text{hurdle}}\}}{\max\{\#\{t_i \geq t_{\text{hurdle}}\}, 1\}} \right],
$$

(28)

where $\text{null}_i$ is defined by

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

(29)

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

I calculate Equation (28) by Monte Carlo simulation of the estimated model. For various values of $t_{\text{hurdle}}$, 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 $\alpha$ ($t_{FDR,\alpha}$), I solve

$$
t_{FDR,\alpha} = \min_{t_{\text{hurdle}}} FDR(t_{\text{hurdle}}) \leq \alpha
$$

(30)

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

I also examine the FDR among published t-stats. Analogous to Equation (28), I define this FDR as

$$
FDR_{\text{pub}} \equiv \mathbb{E} \left[ \frac{\#\{\text{null}_i : i \text{ is published}\}}{\max\{\#\{i : i \text{ is published}\}, 1\}} \right].
$$

(31)

Equation (31) 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 (29) 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 (29), even a tiny \( \mu_i \) 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, Gal, Gelman, Robert, and Tackett (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 adjusted expected return. Scaling by the average published mean return in the model provides a “smooth” shrinkage adjustment

\[
s_{\text{smooth}} = 1 - \frac{\mathbb{E} \left[ \sum_{i=1}^{N_{\text{pub}}} \mu_i \right]}{\mathbb{E} \left[ \sum_{i=1}^{N_{\text{pub}}} \bar{r}_i \right]}.
\]

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 et al. 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 and standard
error

\[ \hat{\mu}_i \equiv \mathbb{E} \left[ \mu_i | \hat{\bar{r}}_i, \text{SE}_i \right]. \]  (33)

The fact that Bayes rule is immune to selection (Senn 2008, Dawid 1994) implies that Equation (33) leads to an unbiased estimate of \( \mu_i \). 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 (32)), I express \( \hat{\mu}_i \) in terms of shrinkage

\[ s_i = 1 - \frac{\hat{\mu}_i}{\hat{\bar{r}}_i}. \]  (34)

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).

**Results** Figure 5 plots the main results. I resample from the dataset of 155 predictors, re-estimate the model, and re-calculate multiple testing 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 (23)), 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).

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 con-
Figure 5: Bootstrapped Distribution of Multiple Testing Statistics. I resample from the dataset of 155 predictors (with replacement), and estimate a multiple testing model (Section 4.3) 1000 times. \( \hat{\omega} \) is the estimated fraction of marginal t-stats that are unpublished. t-hurdles are calculated using Equations (28)-(31) by Monte Carlo. Shrinkage is defined using \([\text{Adjusted Return}] = (1 - \text{Shrinkage})[\text{Sample Mean Return}]\), and is calculated using Equation (32) (Panel C) or Equation (34) (Panels D and E).

trol 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, 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 multiple testing adjustments, however, require less extrapolation. In particular, adjustments for published expected returns have small standard errors, as seen in Panel C of Figure 5. Panel C shows the smoothed shrinkage (Equation (32)), 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 (34)) 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 multiple testing.

Finally, Panel F shows that the FDR among published predictors implies that the vast majority of published results are true, even after accounting for identifi-
cation. The median of the bootstrapped estimates is about 5.1%, with a standard error of 6.0%.

5. Alternative Model Specifications

To further examine identification, this section examines a wide variety of modeling assumptions. Table 5 shows that 10 alternative 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: Alternative Model Assumptions

I bootstrap multiple testing statistics for 10 alternative models (Section 5). This table shows the median (med) and standard deviation (sd) of the bootstrapped distributions. Definitions for the multiple testing statistics are found in Section 4.5. Each row represents a different set of modeling assumptions. The main findings in Figure 5 are reinforced in the bootstraps of these 10 alternative models.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>FDR = 5%</th>
<th>FDR = 1%</th>
<th>FDR for Published</th>
<th>Smooth Shrink for Published</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>med</td>
<td>sd</td>
<td>med</td>
<td>sd</td>
</tr>
<tr>
<td></td>
<td>med</td>
<td>sd</td>
<td>med</td>
<td>sd</td>
</tr>
<tr>
<td>(1) Constant Corr</td>
<td>1.42</td>
<td>0.85</td>
<td>2.51</td>
<td>0.83</td>
</tr>
<tr>
<td>(2) Const Corr = 0.20</td>
<td>1.53</td>
<td>0.91</td>
<td>2.56</td>
<td>1.00</td>
</tr>
<tr>
<td>(3) More SD Corr</td>
<td>1.82</td>
<td>0.98</td>
<td>2.66</td>
<td>1.12</td>
</tr>
<tr>
<td>(4) Risk Prem</td>
<td>2.07</td>
<td>0.69</td>
<td>2.82</td>
<td>0.67</td>
</tr>
<tr>
<td>(5) Logit Select</td>
<td>0.00</td>
<td>0.58</td>
<td>0.00</td>
<td>0.97</td>
</tr>
<tr>
<td>(6) Gamma Shape 2.0</td>
<td>2.09</td>
<td>0.77</td>
<td>2.73</td>
<td>0.86</td>
</tr>
<tr>
<td>(7) Gamma Shape 0.5</td>
<td>1.56</td>
<td>0.87</td>
<td>2.36</td>
<td>1.01</td>
</tr>
<tr>
<td>(8) t d.o.f. 4</td>
<td>1.35</td>
<td>0.95</td>
<td>2.25</td>
<td>1.13</td>
</tr>
<tr>
<td>(9) t d.o.f. 10</td>
<td>2.02</td>
<td>0.84</td>
<td>2.67</td>
<td>0.80</td>
</tr>
<tr>
<td>(10) Fat Log SE</td>
<td>2.44</td>
<td>0.86</td>
<td>3.11</td>
<td>0.90</td>
</tr>
</tbody>
</table>
5.1. **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 4.1. 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 (18) instead of Equation (21) 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.

5.2. **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.

5.3. **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 (21), and a slightly larger standard deviation of correlations. Row (3) of Table 5 show that this assumption has little effect on the main results.

5.4. **Correlation Between Sample Mean Returns and Standard Errors**

The baseline model assumes that expected returns and standard errors are uncorrelated (Equation (16)). This assumption conflicts with the classical as-
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 (16) with

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

(35)

where, as before, $\mu_i$ follows Equation (16) and $\text{SE}_i$ follows Equation (22). Estimating both $\beta$ and $\omega$ 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 $\beta$ 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 $\frac{7}{15} \sqrt{\frac{1}{12}} \sqrt{300} = 2.33$

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

5.5. 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:

$$\pi_{\text{logistic}}(t_i \mid t_{\text{mid}}, t_{\text{slope}}) = \frac{1}{1 + \exp(-t_{\text{slope}}(t_i - t_{\text{mid}}))}. \quad (36)$$

Similar to Equation (23), 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 (36) 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 \( \omega \). As with the low \( \omega \) results in Figure 5, 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.

### 5.6. 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 (16)) 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.

### 5.7. Fat-Tailed Log Standard Errors

Here, I examine replacing the normal assumption for log standard errors (Equation (22)) with a t-distribution with degrees of freedom = 4. This assumption is somewhat extreme, as Figure 4 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 counter-
parts. Similarly, shrinkage is somewhat larger than other estimates but not unusually large considering its standard error.

6. Conclusion

Recent studies argue that multiple testing casts doubt on published cross-sectional stock return predictors. I show that selective reporting casts doubt on multiple testing adjustments—though some adjustments are more affected than others. t-stat hurdles are weakly identified and uninformative because they target both unpublished and published predictors. In contrast, statistics that apply to only published results are more strongly identified. Intuitively, identification is weak for hypothesis tests that are not observed.

Accounting for identification problems, I find that data on published stock return predictors say little about whether t-hurdles should be raised. The data are informative, however, about the veracity of published findings, and show that the vast majority of published predictors are true.

An important caveat is that these results use only in-sample data and focus on equal-weighting, as they are intended to capture published findings. There are compelling reasons to believe that these findings, though they may survive multiple testing adjustments, will not survive investor learning and market frictions, and thus may not be relevant in the future.
References


Harvey, C. and Y. Liu (2013). Multiple testing in economics.


Jacobs, H. and S. Müller (2017b). Anomalies across the globe: Once public, no longer existent?


