Abstract

We propose an estimate of expected returns that accounts for selective anomaly submission and publication, and apply our adjustment to a broad cross-section of anomaly hedge portfolio returns. Selection bias accounts for a modest 10 to 15% of the typical in-sample return. This small bias is due to fact that the dispersion of in-sample returns is nearly twice as large as the typical standard error, indicating a significant amount of variation in true returns. The bias is much smaller than the out-of-sample decline in anomaly returns, showing that that investors learn about mispricing from academic research. Estimations on simulated data show that our bias adjustment is robust.
1. Introduction

The nature of academia leads to an extremely “thorough” investigation of stock return data. Some argue that, subject to this much questioning, the data will tell you whatever you want to hear. Indeed, the data have informed us of hundreds of portfolios with high returns and low market risk, a result which leads many to be suspicious of information obtained in this manner (for example, Sullivan, Timmermann, and White (1999), Harvey, Y. Liu, and Zhu (2015), and Linnainmaa and Roberts (2016)).

Our enhanced interrogation of the data is subject to controls, however. Though a skilled investigator may be able to coerce her desired answer, the confession is only published if editors and referees deem it trustworthy. Indeed, In reflecting on his years as the editor of the Journal of Finance, Harvey (2014) recommends that empirical submissions should “convince the reader that there has been minimal data mining.” These controls are evidenced in Lewellen (2014), McLean and Pontiff (2016), and Jacobs and Müller (2016), who find that anomaly portfolios perform reasonably well out-of-sample.

Data mining and publication controls amount to two kinds of selection bias. These biases have opposing effects, and the net result remains an open question. In this paper, we propose an estimator which measures the net effect and apply our estimator to a broad cross-section of anomaly returns.

We find that the enhanced but controlled interrogation of the CRSP tapes is surprisingly effective at uncovering true cross-sectional variation in returns. We estimate that a modest 15% of the typical in-sample return is due to selection bias—that is, while the typical value-weighted decile return is about 6% per year historically, the bias-adjusted return is roughly 5% ≈ 6% × (1 − 0.15).

How does our estimator find such a small amount of selection bias? The intuition can be found in a cocktail napkin version of our selection-bias adjustment. The bias-adjusted return is just the in-sample return, scaled by a shrinkage coefficient

\[
\text{Bias-Adjusted Return} \equiv (1 - \text{Shrinkage}) \times \text{In-Sample Return} \quad (1)
\]

1 Throughout this paper, “return” refers to “mean return.” We also omit the word “mean” in “in-sample mean return,” “true mean return,” etc.
where the shrinkage is

$$\text{Shrinkage} \equiv \frac{\text{Time-Series Variance of Measurement Error}}{\text{Cross-Sectional Variance of All In-Sample Returns}}. \quad (2)$$

Intuitively, if data mining overwhelms publication controls, then the cross-sectional dispersion of in-sample returns is entirely due to measurement error, the denominator and numerator are equal, and shrinkage is 100%. Conversely, effective publication controls imply that in-sample returns are due to diverse signals. In this case, the dispersion of in-sample returns (denominator) is much larger than that which comes from measurement error alone (the numerator), and shrinkage is small.

For magnitudes, one needs two facts and one assumption. The two facts are: (1) the typical standard error of an anomaly return is about 30 basis points per month and (2) the standard deviation of published in-sample returns is about 40 basis points (McLean and Pontiff (2016), Novy-Marx and Velikov (2016)). The assumption that’s needed is some model of how unpublished in-sample returns relate to the published ones. A simple model is that there are as many unpublished anomalies with negative returns as there are published anomalies (which all have positive returns). Thus, the standard deviation of all in-sample returns is roughly double the standard deviation of published in-sample returns. With these numbers in hand, we have

$$\text{Shrinkage} \approx \frac{30^2}{(2 \times 40)^2} = 14\%, \quad (3)$$

that is, bias-adjusted returns are just 14% smaller than in-sample returns.

Our headline 15% shrinkage comes from a more rigorous calculation. We write down a simple model of selected anomaly publication and estimate the model with simulated method of moments (Gourieroux and Monfort (1997)). The model implies a Bayesian formula for adjusting for selection bias similar to equations (1) and (2), and results in magnitudes similar to the cocktail napkin calculation (3).

Our bias adjustment is robust. It produces unbiased returns out of data from a wide variety of simulated publication processes. This accuracy holds if selection bias is large or small, if standard errors are uniform or dispersed, and even holds in the presence of significant misspecification problems. Though we rely on simulations, theoretical proofs of the effectiveness of similar estimators have
been demonstrated using frequentist (Stein (1956), James and Stein (1961)) as well as empirical Bayes methods (Efron (2009), Efron (2011)).

Empirical testing of our bias adjustment is, unfortunately, not straightforward. The discovery of anomalous profits creates powerful market incentives that reduce those profits (Q. Liu, Lu, Sun, and Yan (2015), McLean and Pontiff (2016)). Thus, while the natural empirical counterpart of a bias-adjusted return is the post-sample return, these two returns cannot be compared directly, as the post-sample return is polluted by trading effects.

Nevertheless, our results are consistent with several of McLean and Pontiff (2016)’s findings. They find that the upper bound on selection bias is about 15 basis points per month, and this upper bound increases by 5 basis points for every 10 basis points in additional in-sample return. Our estimates are well within these bounds, and indeed, indicate that the bulk of the decline in returns out-of-sample is due to trading effects. A significant part of this decline occurs before the publication date, indicating that market participants actively learn about mispricing from academics, taking advantage of working papers even before they are published.

An important caveat in our results is that we rely on Novy-Marx and Velikov (2016)’s dataset of 32 portfolios. While the size of this set may appear small, it is representative of the broader literature in several ways.

Most important, the NV data is quite close to McLean and Pontiff (2016) set of 97 anomalies in terms of the key statistics. The two datasets are similar in terms of the mean in-sample return, mean out-of-sample return, and the dispersion of in-sample returns. These similarities imply that the estimated shrinkage, as well as the contribution of trading effects, should be similar across datasets.

Additionally, the NV data is constructed to limit the number of related variables. In contrast, the anomalies literature features many variables related to value, momentum, issuance, etc. For example, McLean and Pontiff (2016)’s set includes 14 variables related to value, 8 related to momentum, and 10 related to equity issuance.

Lastly, the representativeness of 32 portfolios is consistent with Green, Hand, and Zhang (2014) who find that 20-30 variables are required to summarize a longer list of 100. Green, Hand, and Zhang (2014) use a wide variety of approaches, and each leads to a slightly different list of summary variables, but our list contains almost all of the non-trading-volume related variables that they find
are important.

**Related Literature** Concerns about selection bias in published stock market anomalies go back to Jensen and Bennington (1970) (see also Merton (1987) and Black (1993)). Systematic examination of these biases was not possible until the recent proliferation of anomalies, however.

In an influential study, Harvey, Y. Liu, and Zhu (2015) examine more than 300 anomalies and adjust p-values using standard multiple-testing statistics. Harvey and Y. Liu (2015) develop a bootstrap data snooping test in the spirit of White (2000) (see also Sullivan, Timmermann, and White (1999), Sullivan, Timmermann, and White (2001)) and apply them to factor models. Overall, these papers conclude that most published anomalies are false.

Our study differs by taking a more structured approach and explicitly modeling the anomaly publication process. More structure leads to sharper inferences, given that the structure is close enough to the truth (for example, Harvey and Y. Liu (2016)). The fact that our structure leads to a rather different conclusion than the previous papers indicates that the ideal model of anomaly publication is an important question for future research.

The need for sharper inferences is highlighted by the conflicting results from the more empirically focused papers. Anomalies remain positive post-sample in both the United States (McLean and Pontiff (2016)) and in 38 other countries (Jacobs and Müller (2016)), suggesting that selection bias is relatively small. On the other hand, anomaly returns were very poor in the United States before 1963 (Linnainmaa and Roberts (2016)), suggesting that the published in-sample returns were data-mined. As technology and institutions change across decades and countries, there may be multiple explanations for these results. Placing some structure on the problem allows us to measure selection bias without introducing these complications.

Our estimator is a version of James and Stein (1961) shrinkage, which has been studied extensively in the portfolio choice literature (see Brandt (2009) for a review). Papers in this literature only consider a couple of signals (for example, Kan and Zhou (2007)). In contrast, we apply James-Stein to a wide variety of signals, and extend James-Stein to account for the fact that we observe only

---

2 Of the 24 variables in their headline result, our list misses 3 of the non-volume variables, and of the 20 variables that they find describe mid-to-large cap stocks, our list misses just 2. The missing variables are growth in total taxes, convertible debt dummy, and R&D/market cap.
published signals.

Our model complements Q. Liu, Lu, Sun, and Yan (2015)’s model of anomaly discovery. While their model focuses on trading effects and abstracts from selection bias, we do exactly the converse. Thus, the two models capture two distinct components of the decay in returns post-sample.

Section 2 describes our bias adjustment. Readers eager for results may want to skip to Section 3 which applies the bias adjustment to empirical data. Section 4 shows that the bias adjustment is robust.

2. Econometric Method

Our bias adjustment is based off of a simple model of anomaly publication (Section 2.1). The bias adjustment comes from estimating the model, and then using Bayes rule to update on in-sample returns (Section 2.2). We do not prove the unbiasedness of the estimator formally, and instead use simulations to demonstrate its effectiveness (Section 2.3).

2.1. A Simple Model of Anomaly Publication

The model is summarized in Table 1. It is a simple statistical description of the portfolio publication process. We introduce characters like “academics” and “journals” for clear interpretation. There are no dynamics, trading, or strategic behavior. Thus the ‘true return’ in the model is best understood as the selection bias adjusted return, or the return in a world in which the anomaly remains untouched by traders.

In search of tenure or other glory, academics mine financial market and related data for anything publishable. Collectively, academics submit every portfolio that has a remote possibility of being published.

Journals only publish portfolios that meet two requirements. The first requirement is that the portfolio must contain a “narrative” that meets the journals’ standards. Portfolios with a narrative are consistent with a sound economic or psychological theory. For example, momentum’s narrative is that investors overreact to the past year’s returns. Thus, a narrative implicitly places a sign on the portfolio (long past winners and short past losers).
Table 1: Model Summary

<table>
<thead>
<tr>
<th>Properties of the Portfolio Based on Narrative i</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>True return</td>
<td>$\mu_{r,i} \sim N(\mu_{\mu}, \sigma_{\mu})$</td>
</tr>
<tr>
<td>In-sample return</td>
<td>$\bar{r}<em>i = \mu</em>{r,i} + \epsilon_{r,i}$ $\epsilon_{r,i} \sim N(0, \sigma_{r,i})$</td>
</tr>
<tr>
<td>Log standard error</td>
<td>$\log(\sigma_{r,i}) \sim N(\mu_{\sigma}, \sigma_{\sigma})$</td>
</tr>
</tbody>
</table>

Publication Cutoffs

- minimum in-sample return $\bar{r}_i > \bar{r}_{\text{min}}$
- minimum t-stat $\frac{\bar{r}_i}{\sigma_{r,i}} > t_{\text{min}}$

This table summarizes the model (Section 2.1). All portfolios which have a remote possibility of being published are submitted. Only portfolios based on narratives and meet cutoffs are published. All distributions are independent.

The quality of narrative $i$ is captured by its true return $\mu_{r,i}$, and the quality of all narratives is described by a normal distribution

$$\mu_{r,i} \sim N(\mu_{\mu}, \sigma_{\mu}). \tag{4}$$

$\mu_{\mu}$, and $\sigma_{\mu}$ are the net result of authors’ data mining and the journals’ narrative screening, and they ultimately describe whether the journals are publishing true returns. Clearly, if $\mu_{\mu} \gg 0$ the net result is that the narrative screen is effective at eliminating spurious portfolios.

The narrative screen is also effective if $\mu_{\mu} = 0$ but $\sigma_{\mu} \gg 0$. In this case, some published portfolios will have truly high expected returns. One can think of the $\sigma_{\mu}$ parameter as a measure of the “openness” of the publication process. We will see that empirically, $\sigma_{\mu}$ is crucial for determining the magnitude of selection bias.

The normal distribution of (4), with its single peak and simple shape, means that we’re examining portfolios of a single “type.” Type is a flexible concept: with enough data, all portfolios may end up falling into the same broad category and have normally distributed true returns. But as our data is limited, equation (4) will lead us to focus on single sort value-weighted decile hedge returns in the empirical results (Section 3).

The narrative screen is not perfect. Equation (4) means that some (perhaps
many) narratives have $\mu_{r,i} < 0$, and the journals can't observe $\mu_{r,i}$. Instead, they observe noisy signals: the in-sample return $\tilde{r}_i$ and standard error $\sigma_{r,i}$.

For a randomly selected narrative $i$, the in-sample return follows:

$$\tilde{r}_i = \mu_{r,i} + \epsilon_{r,i}$$

and

$$\epsilon_{r,i} \sim N(0, \sigma_{r,i}).$$

Narrative portfolios are heterogeneous in standard errors, with a right-skewed distribution

$$\log \sigma_{r,i} \sim N(\mu_\sigma, \sigma_\sigma).$$

This setting leads to the journals’ second requirement for portfolio $i$’s publication: its in-sample return $\tilde{r}_i$ and t-stat $\frac{\tilde{r}_i}{\sigma_{r,i}}$ must meet hurdles

$$\tilde{r}_i > \tilde{r}_{\text{min}}$$

and

$$\frac{\tilde{r}_i}{\sigma_{r,i}} > t_{\text{min}}$$

where $\sigma_{r,i}$ is the standard error of portfolio $i$. This second screen improves the quality of published portfolios, as $\tilde{r}_i$ is a noisy signal of $\mu_{r,i}$. Unfortunately, the cost of this quality control is a bias: equation (5) does not apply to the set of published portfolios, as this is a selected sample of narratives.

Equations (5) and (6) incorporate a couple of simplifying assumptions that are justified by data. The first is that $\tilde{r}_i$ are independent of each other and the other draws. This assumption is consistent with McLean and Pontiff (2016)'s finding that anomaly portfolios have small correlations, on average.

The second simplifying assumption is that the standard errors are independent of the true return. One might think that the volatility component of the standard error should be correlated with the true return, as in equilibrium theories based on risk. The anomalies literature, however, is focused on portfolios that survive risk adjustment. Indeed, the literature tends to find a wide variety of in-sample returns with similar volatilities.
2.2. Selection Bias Adjustment

Our bias adjustment proceeds in two steps: (1) we estimate the model using SMM, then (2) apply Bayes rule to in-sample returns through the lens of the estimated model. Bayes rule gives the true, bias-adjusted, expected return.

To make the estimation transparent, we choose as many parameters as we can outside of the SMM algorithm. We set the following parameters according to sample statistics:

\[ \hat{r}_{\text{min}} = p \text{th percentile of published } \bar{r}_i \]  
\[ \hat{t}_{\text{min}} = p \text{th percentile of published } \bar{r}_i / \sigma_{r,i} \]  
\[ \hat{\sigma}_\sigma^2 = \frac{1}{N} \sum_i \left( \log \sigma_{r,i} - \log \sigma_r \right)^2, \]  

where \( \log \sigma_r \) is the mean of \( \log \sigma_{r,i} \). Taking the model literally implies that \( p = 0 \). However, there is some leeway in this parameter in real life, as the long-short portfolios we examine may not match exactly the published data or portfolio formation method.

As the mean of true means \( \mu_{\mu} \) is very difficult to identify, we simply assume its value based off of simple rules of thumb. For example, our baseline rule of thumb is very skeptical: \( \mu_{\mu} = 0 \). This approach can be thought of as having a strict prior about \( \mu_{\mu} \), or simply as having an arbitrary shrinkage target for improving accuracy (as in Stein (1956)). Regardless, We will see that the overall bias adjustment does not depend on the choice of \( \mu_{\mu} \) (Sections 2.3 and 3.2).

Just two parameters, the dispersion of true returns \( \sigma_{\mu} \) and the mean standard error \( E(\sigma_{r,i}) \), are estimated by SMM. The estimation targets the univariate histograms of \( \bar{r}_i \) and \( \sigma_{r,i} \). We use one-stage, identity matrix weighted SMM for simplicity. Specifically,

\[
[\hat{\sigma}_{\mu}, \hat{E}(\sigma_{r,i})] = \arg\min_{\sigma_{\mu}, E(\sigma_{r,i})} \sum_{k=1}^{N_{\text{Hist}}} \left\{ f_{r,k}^{\text{sim}} \left( \sigma_{\mu}, E(\sigma_{r,i}); \hat{r}_{\text{min}}, \hat{t}_{\text{min}}, \hat{\sigma}_\sigma^2, \mu_{\mu} \right) - f_{r,k}^{\text{data}} \right\}^2 + \sum_{i=k}^{N_{\text{Hist}}} \left\{ f_{\sigma,k}^{\text{sim}} \left( \sigma_{\mu}, E(\sigma_{r,i}); \hat{r}_{\text{min}}, \hat{t}_{\text{min}}, \hat{\sigma}_\sigma^2, \mu_{\mu} \right) - f_{\sigma,k}^{\text{data}} \right\}^2.
\]  

(13)

where \( E(\sigma_{r,i}) = \mu_{\sigma} + \frac{1}{2} \sigma_{\sigma}^2 \), \( f_{r,k}^{\text{sim}} \) is the normalized number of simulated portfolios with \( \bar{r}_i \) in bin \( k \), and \( f_{\sigma,k}^{\text{sim}} \) is the analogous number for \( \sigma_{r,i} \). Histogram bins are
chosen algorithmically within each simulation. The histogram binning needs to be tuned, however, we find that the results are fairly insensitive to the binning method.

The selection bias adjustment is based off of the estimated parameters. Suppose we observe a published in-sample return $r_j$ with standard error $\sigma_{r,j}$. The bias adjusted return is

$$\hat{\mu}_{r,j} = (1 - s_j) \bar{r}_j + s_j \mu_{\mu}$$

(14)

where the “shrinkage” $s_j$ is

$$s_j = \frac{\sigma^2_{r,j}}{\hat{\sigma}_{\mu}^2 + \sigma^2_{r,j}}.$$  

(15)

This bias adjustment is a simulated version of the celebrated James and Stein (1961) estimator for a vector of means. These estimators improve on the in-sample mean of a given observation by incorporating information from other observations. There is a rich literature that derives theoretical properties of this kind of estimator using frequentist approaches. Here we give a simple Bayesian interpretation following Senn (2008) (see also Efron (2011)).

The Bayesian interpretation comes from updating about a published portfolio $j$ within the context of the model. One can think of the distribution of $\mu_{r,i}$ as forming a “prior” about $j$’s true mean return. Within the model, this prior is effectively an infinitely large sample: it gives all the information one needs to know about $j$, aside from the particular statistics that describe $j$: $\bar{r}_j$ and $\sigma_{r,j}$. As a result, conditioning on the prior already accounts for the fact that $j$ comes from a selected sample. Mathematically, this means

$$E(\mu_{r,j}|\bar{r}_j, \sigma_{r,j}, \bar{r}_j > \bar{r}_{min}, \sigma_{r,j} > \sigma_{min}, \mu_{\mu}, \hat{\sigma}_{\mu}) = E(\mu_{r,j}|\bar{r}_j, \sigma_{r,j}, \mu_{\mu}, \hat{\sigma}_{\mu})$$

(16)

$$= \hat{\mu}_{r,j},$$

(17)

where the last equality comes from the normal-normal updating formula. For readers accustomed to frequentist thinking, this reasoning may be hard to digest—indeed, Dawid (1994) calls it a paradox. However, its effectiveness can be clearly demonstrated in simulations (Section 2.3).
Table 2: Estimations Using Simulated Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Estimations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Skeptical</td>
</tr>
<tr>
<td>$\mu_{\mu}$</td>
<td>0.05</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma_{\mu}$</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td>$\mathbb{E}(\sigma_{r,i})$</td>
<td>0.3</td>
<td>0.29</td>
</tr>
<tr>
<td>$\sigma_{\sigma}$</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>$r_{min}$</td>
<td>0.25</td>
<td>0.27</td>
</tr>
<tr>
<td>$t_{min}$</td>
<td>2</td>
<td>2.01</td>
</tr>
<tr>
<td>Mean Shrinkage</td>
<td>0.54</td>
<td>0.57</td>
</tr>
<tr>
<td>Effective Shrinkage</td>
<td>0.56</td>
<td>0.64</td>
</tr>
</tbody>
</table>

All values are monthly. We simulate the model (Table 1) using ‘True Values,’ which are chosen to display strong selection bias. The ‘Skeptical’ estimation assumes $\mu_{\mu} = 0$, and the ‘Moderate’ estimation assumes $\mu_{\mu} = 1/2$ of the mean published return. Both estimations use come from SMM (equation 13). ‘Mean Shrinkage’ is the proportion of shrinking toward the target $\mu_{\mu}$ (equation 14). ‘Effective shrinkage’ is the mean bias-adjusted return / mean in-sample return. Shrinkages in ‘True Value’ column use the true parameter values. Publication cutoffs are the 1st percentiles of published data. Simulations from more parameter values can be found in Section 4.

2.3. Simulation Test

This section demonstrates that the bias adjustment is effective in simulated data. We assume some true parameter values, simulate portfolios using the model (Table 1), and then apply two versions of our bias adjustment: (1) a very skeptical adjustment with $\mu_{\mu} = 0$ and (2) a more moderate one which assumes $\mu_{\mu} = (1/2)$ of the mean published in sample return. The true parameters values are chosen to display large simulated selection bias. Section 4 examines simulations under additional parameter choices.

Table 2 shows the true parameters and estimated values. The parameters chosen outside of SMM are estimated quite accurately. Indeed, the dispersion of log standard errors is quite close to the truth, showing that selection bias has little effect on this parameter.

SMM does a good job estimating the mean standard error, but the dispersion of true means $\sigma_{\mu}$ varies quite a lot across the two assumptions about $\mu_{\mu}$. The
$\mu_\mu = 0$ estimator results in an accurate estimate of this parameter, while the more moderate estimator leads to an excessively low estimate.

Despite this disparity, both estimators result in similar overall bias adjustments. This is seen in the Effective Shrinkage row: both estimators result in shrinkage of roughly 50%, close to the optimal shrinkage of 56%. We will return to this result shortly.

The accurate effective shrinkage estimates mean that both estimators result in a dramatic improvement on the in-sample return. This improvement can be seen in Figure 1, which plots the forecast error as a function of the in-sample return. The standard forecast: in-sample return = predicted return results in a large upward bias, with the bias increasing in the in-sample return. In contrast, both bias adjustments generate largely unbiased forecasts with small errors, regardless of the in-sample return.

Figure 2 takes a closer look at the results in Figure 1. The figure shows forecasts and true returns for portfolios sorted by standard error. The deviation between the simple forecasts (black line) and the true returns (blue x’s) highlights the importance of the standard error: the simple forecast’s bias is the worst in the high standard error portfolios (bottom-left panel).

In contrast, the both bias adjustments (red and green lines) run through the middle of the clouds of true returns. The plots also show that the forecasts basically adjust the slope and intercept of the forecasts. While the slope of the simple forecast is 1 (In-Sample = Forecast), the bias adjustments result in flatter slopes.

These slopes illustrate the compensation mechanism built in the estimator. While the moderate adjustment misspecifies the mean of returns as too high (high intercept), it compensates by underestimating the dispersion (shallow slope). The net effect is a forecast cured of selection bias: the moderate forecasts runs neatly through the middle of the cloud of out-of-sample returns. Indeed, Efron and Morris (1976) show that any choice of $\mu_\mu$ improves on the in-sample return in a similar class of estimators.

A key assumption in this analysis is that the model is correctly specified. Misspecification is always difficult to deal with, as there are an innumerable number of ways a model can be misspecified. Nevertheless, Section 4 shows that the results are robust to a few important types of misspecification.
Figure 1: Errors of Various Forecasts in Simulated Data. This figure demonstrates the effectiveness of our selection bias adjustments. Portfolios are sorted into buckets by in-sample return, and markers show the mean forecast error within each bucket, with bars indicating one standard deviation. The standard forecast assumes the in-sample return is equal to the true return. While the simple forecast is systematically biased upwards, the adjusted forecasts have very little bias. The adjusted forecasts apply equations (14), and (15) using estimated values from Table 2.
Figure 2: True Returns and Forecasts by Standard Error in Simulated Data. This figure provides a detailed view of Figure 1. Portfolios are separated into terciles of standard error. The standard forecast (black line) overestimates true returns (blue x’s), and the miss is the worst for high standard errors. Adjusted forecasts (other lines) are approximations of equations (14), and (15) that use the mean value of $\sigma_{r,i}$ within each tercile. All adjustments perform well, including the ‘moderate’ adjustment (green line), which mistakenly assumes the mean of true returns is high (Table 2). The moderate adjustment compensates for this misspecification by estimating a larger shrinkage (shallower slope).
3. Empirical Results

Having described our econometric method and shown that it effectively removes selection bias, we now move on to the empirical results.

3.1. Data Description

Our data begins with Novy-Marx and Velikov (2016)’s (NV’s) sample of 32 long-short portfolios. NV’s published paper uses only 23 of these portfolios, but Novy-Marx kindly provides a larger dataset on his website.

Only 30 of the portfolios have reasonable performance, however. One of these two poor performers is gross margin, which was originally intended to be more of a control variable, and generates a tiny return spread unless one adjusts for HML exposure (Novy-Marx (2013)). The other poor performer is beta arbitrage, which has a surprisingly small t-statistic of 0.16 under NV’s value-weighted construction. We exclude these two poor performers from our analysis.

Of the remaining 30 portfolios, we eliminate another 5 because they are not single sorts. Our framework assumes that the data are drawn from the same normal distribution, a notion which is hard to justify if we mix in a couple double- or triple-sorts. Regardless, the main results are not sensitive to this data cleaning. In the Appendix (Section A.3) we show that we obtain similar shrinkages of between 10 and 15% whether we use our 25 portfolios, the 30 with good performance, or the 23 portfolios featured in the published version of Novy-Marx and Velikov (2016).

While our list of 25 portfolios may appear short, it is representative of a larger sample in several ways. NV’s data is chosen to limit the number of related anomalies. While McLean and Pontiff (2016)’s 97 portfolios include 10 related to valuation, 8 related to momentum, and 8 related to investment, our subset of NV’s data includes just 3 valuation-based, 2 momentum-based, and 1 investment-based portfolio. Our sample size is also justified by Green, Hand, and

\[^3\]Beta arbitrage also introduces a large amount of noise due to its enormous standard error of 2.92 percent, nearly 20 standard deviations larger that the mean standard error. Our estimator seems to be able to handle this extreme observation in that it still produces reasonable shrinkage estimates, and, indeed, recommends near 100% shrinkage for the beta arbitrage portfolio. But including this outlier makes the estimated shrinkage somewhat sensitive to the choice of histogram bins: with 7 bins, we get a mean shrinkage of 24%, but 6 bins leads to 20%. In contrast, our baseline results are essentially insensitive to the choice of bins.
Table 3: Dataset Comparison.

<table>
<thead>
<tr>
<th></th>
<th>Our Data</th>
<th>McLean and Pontiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of predictor portfolios</td>
<td>25</td>
<td>97</td>
</tr>
<tr>
<td>Predictors portfolios with t-statistic &gt; 1.5</td>
<td>22</td>
<td>85</td>
</tr>
<tr>
<td>Mean publication year</td>
<td>2005</td>
<td>2000</td>
</tr>
<tr>
<td>Median publication year</td>
<td>2008</td>
<td>2001</td>
</tr>
<tr>
<td>Mean portfolio return in-sample</td>
<td>0.731</td>
<td>0.582</td>
</tr>
<tr>
<td>Std. dev. of in-sample means</td>
<td>0.451</td>
<td>0.395</td>
</tr>
<tr>
<td>Mean observations in-sample</td>
<td>363</td>
<td>323</td>
</tr>
<tr>
<td>Mean portfolio return out-of-sample</td>
<td>0.506</td>
<td>0.402</td>
</tr>
<tr>
<td>Std. dev. of out-of-sample means</td>
<td>0.785</td>
<td>0.651</td>
</tr>
<tr>
<td>Mean observations out-of-sample</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>Mean portfolio return post-publication</td>
<td>0.122</td>
<td>0.264</td>
</tr>
<tr>
<td>Std. dev. of post-publication means</td>
<td>0.386</td>
<td>0.516</td>
</tr>
<tr>
<td>Mean observations post-publication</td>
<td>157</td>
<td>156</td>
</tr>
</tbody>
</table>

This table compares our variation of the Novy-Marx and Velikov data to McLean and Pontiff (2016).

Zhang (2014)'s finding that 100 characteristics can be summarized by 24 factors (see additional discussion in the Introduction).

Most important, our data features summary statistics which are very close to those of McLean and Pontiff (2016), as seen in Table 3. Our data and McLean and Pontiff (2016)'s are similar in terms of mean in-sample returns, mean out-of-sample returns, mean observations in-sample, mean observations out-of-sample, and more. The two samples differ somewhat in their post-publication mean returns, but the difference is well within one standard error.

These similar summary statistics come in spite of the fact that McLean and Pontiff (2016)'s portfolios are equal-weighted quintile sorts and NV's data consists of value-weighted decile sorts. This may be the result of the fact that the two differences in portfolio construction balance: while value-weighting tends to produce lower and less volatile returns, deciles tend to produce higher and more volatile returns. We consider these similarities a reassuring sign for anomalies in general. As there are many ways of forming portfolios based off of a given anomaly variable, it's nice to know that there is some amount of robustness to the choice of formation method.
3.2. Main Results: Empirical Estimation and Bias-Adjustment

Finally, we are in a position to show the main results. Table 4 begins with parameter estimates of our model of anomaly publication (Section 2.1) using empirical data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimations</th>
<th>Skeptical</th>
<th>Moderate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_\mu$</td>
<td>mean of true means</td>
<td>0</td>
<td>0.41</td>
</tr>
<tr>
<td>$\sigma_\mu$</td>
<td>dispersion of true means</td>
<td>0.72</td>
<td>0.54</td>
</tr>
<tr>
<td>$\mathbb{E}(\sigma_r,i)$</td>
<td>mean standard error</td>
<td>0.25</td>
<td>0.26</td>
</tr>
<tr>
<td>$\sigma_\sigma$</td>
<td>dispersion of log std. err.</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>$r_{min}$</td>
<td>min $\tilde{r}_i$ for publication</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>$t_{min}$</td>
<td>min t-stat for publication</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>Mean Shrinkage</td>
<td></td>
<td>0.148</td>
<td>0.223</td>
</tr>
<tr>
<td>Effective Shrinkage</td>
<td></td>
<td>0.147</td>
<td>0.111</td>
</tr>
</tbody>
</table>

All values are monthly. We apply our SMM estimator (Section 2.2) to our empirical sample of anomaly returns. The Skeptical estimator assumes $\mu_\mu = 0$, while the Moderate estimator assumes $\mu_\mu = 0.5 \times$ the mean in-sample return. ‘Mean Shrinkage’ is the proportion of shrinking toward the target $\mu_\mu$ (equation (14)). ‘Effective shrinkage’ is the mean bias-adjusted return / mean in-sample return. Publication cutoffs are the 5th percentile of published statistics.

As in Section 2.3, consider both a skeptical bias adjustment and a more moderate one. The skeptical adjustment assumes the mean of true returns $\mu_\mu = 0$, while the moderate one assumes $\mu_\mu = (1/2)$ the mean published in sample return. This calculation results in the mean true return being 41 basis points per month, which is economically far from 0.

The key parameter, the dispersion of true returns, is estimated to be 72 basis points per month in the skeptical estimation. This value is intuitive when compared to the dispersion of 45 basis points seen in the published returns (Table 3). If we assume that the mean of true returns is zero, then the published data is missing roughly half (the left half) of the distribution of all in-sample returns.

The moderate estimator’s dispersion of true returns is smaller as it assumes
a higher mean. Intuitively, the moderate estimator assumes that the published data is more representative of the full sample, and thus less of the full dispersion is missing.

Despite their different parameter estimates, both estimators lead to similar shrinkage, as seen in the bottom rows of Table 4. The skeptical estimator shrinks in-sample returns by 15% toward a target of 0, while the moderate estimate shrinks 22% toward a target of 41 basis points. The net effect is similar and captured by the ‘effective shrinkage,’ that is, the mean bias adjusted return divided by the mean in-sample return. The effective shrinkages are 14.7% and 11.1% for the skeptical and moderate estimators, respectively. This consistent shrinkage in spite of different parameter estimates is the result of the compensation mechanism discussed in Section 2.2.

The small overall shrinkage can be understood through the Bayes-Stein bias adjustment formula (14) and (15). The shrinkage \( s_j \) basically compares the amount of dispersion in true returns \( \sigma_{\mu} \) (signal) to the standard error \( \sigma_{r,j} \) (noise). As the estimated dispersion of true returns of 70 basis points is much larger than the standard error of 30 basis points, the estimator considers the bulk of the in-sample return as signal. The exact formula (15) squares the dispersions, which amplifies the difference and leads to the small shrinkage of 15%.

The mean shrinkage figures hide some nuances, as the shrinkage for a given anomaly depends on its standard error. This heterogeneity is shown in the top panels of Figure 3 which plots the distribution of shrinkage. The distributions are right skewed for both estimators: while the vast majority of shrinkage values is about 10-15%, a handful of anomalies are shrunk by more than 40%. These outliers are the anomalies with very volatile returns (idiosyncratic volatility) or a short sample (SUE). The idea of applying more shrinkage to these kinds of anomalies is intuitive: anomalies with large measurement error are more likely to have had lucky in-sample returns, and thus exhibit more selection bias (on average).

Despite this heterogeneity, the mean effective shrinkage numbers from Table 4 are representative of the full distribution. The bottom panel of Figure 3 shows that the distributions of in-sample and bias-adjusted returns and are quite similar. In a couple of the bins, one can see the reallocation of mass from higher returns to lower returns, but overall this adjustment is small.
Figure 3: Distributions of Shrinkage and Bias-Adjusted Returns from Empirical Data. Shrinkage depends on the standard error (equation (15)), which leads to significant heterogeneity in shrinkages, as seen in the top panels. Despite this heterogeneity in shrinkage, the mean shrinkage of 15% is an accurate description of the mild overall bias adjustment (bottom panel).
**Implied Trading Effects**  The discovery of portfolios with high returns may lead to market incentives which reduce those returns. Such “trading effects” are an important issue because they shed light on the source of the returns: returns due to mispricing will be eliminated, however, returns due to risk will remain (Q. Liu et al. (2015)).

Measuring trading effects is complicated, however, by selection bias. As demonstrated in Figure 1, selection bias also leads to the elimination of returns out-of-sample.

The existence of selection bias forces McLean and Pontiff (2016) (MP, hereafter) to make somewhat conservative interpretations of their out-of-sample results. Defining out-of-sample as ‘after the sample used in the publication and before the publication date,’ MP find that out-of-sample returns are roughly 0.15 percentage points lower than in-sample returns. Without a method to decompose this decay into trading and selection bias effects, they simply conclude that this 0.15 percentage points is an upper bound on selection bias.

Our bias adjustment allows for more refined estimates and stronger conclusions. In fact, the adjustment gives us trading effects as a function of in-sample returns. These effects are shown in Figure 4 which plots out-of-sample decay (in-sample returns minus out-of-sample returns, following MP’s convention) for portfolios with below median in-sample returns and above median in-sample returns.

The plot shows that the out-of-sample decay in our data is slightly higher than MP’s, though it is well within one standard error. Selection bias, that is, the amount of decay implied by our estimated bias adjustment, accounts for a minority of this decay (green bars). The large contribution of trading effects (blue bars) shows that market participants are actively learning about mispricing from academia, taking advantage of information in working papers even before they are published.

The plot also shows that portfolios with higher in-sample returns experience larger decay, an effect highlighted in MP. The much larger green section of the high returns bar shows that, though trading effects play a role, this larger decay is primarily due to selection bias. Intuitively, high in-sample returns are more likely to be due to luck, and thus need a larger bias adjustment (on average). Moreover, high in-sample returns means that the portfolio can have a large standard error and still make the typical t-stat publication cutoff. Larger standard errors also
Figure 4: Trading Effects vs Selection Bias in Out-of-Sample Returns. Out-of-sample returns are the mean return between the end of the sample in the original paper and the publication date of the paper (same as in McLean and Pontiff (2016)). Out-of-sample decay is the difference between the in-sample return and the out-of-sample return (in percentage points). The selection bias effect is the difference between the in-sample return and the bias adjusted return. The trading effect is the difference between the out-of-sample decay and the selection bias effect. We divide the data in two groups: low in-sample returns are below the median in-sample return, and high in-sample returns are above the median.

3.3. Estimation Details

Having seen the main results, we now take a closer look at the estimation. Here we show that parameter estimates are a natural result of the data. Perhaps the first question is: did the nonlinear SMM minimization work? Because we can estimate all but 2 parameters outside of the SMM minimization, a simple 2D plot clearly answers this question. Figure 5 shows the objective for the skeptical estimator, and we see that the objective is very well-behaved.

Over most of the domain, the objective appears continuous and differentiable. The objective forms a nice bowl with a clear region of optimality. Using the grid
Figure 5: SMM objective function for the Skeptical Estimation on Empirical Data. This figure shows that the SMM objective (problem (13)) is well-behaved. Each line shows a different mean standard error ($E(\sigma_{r,i})$), with lighter lines indicating a higher mean standard error. The objective shows a clear region of optimality around the grid search minimum (red x). The algorithm uses the grid search minimum (red x) as an initial guess for a Nelder-Mead non-linear optimization.

This well-behaved objective is reflected in a good fit for the targeted data. The top panel of Figure 6 shows the SMM targets: the histogram counts of in-sample returns and the histogram counts of standard errors. Both histograms follow patterns predicted by the model: the distributions are right skewed, but still bell-shaped overall. Moreover, the decays in the histograms are fairly smooth, as predicted. There is a slight indication of a second mode near the in-sample return of 100 basis points, but we shall see in the robustness section (Section 4) that our estimator is effective under this kind of misspecification.

The large estimated dispersion in true returns is due to the large dispersion
Figure 6: SMM Fit and Estimated Distributions of All Narrative In-Sample Returns for Empirical Data. The top two panels show the quality of the SMM target fits. The SMM targets published data on (1) in-sample returns and (2) standard errors. The bottom panel shows the implied distribution of in-sample return returns, which is critical for the bias adjustment (equations [14], and [15]).
of in-sample returns. This large dispersion can be seen in the top left panel of Figure 6, which plots the histogram of in-sample returns. Though most of the published returns are near 50-60 basis points (i.e. value, asset growth), a significant portion of the data lies above 100 basis points (i.e. SUE, momentum). To put this dispersion in perspective, the typical standard error in our sample is about $0.30 \approx \frac{500}{\sqrt{360}}$. Thus, the large dispersion in the histogram shows that it’s very easy to find in-sample returns more than one standard error apart. As a result, the large dispersion in in-sample returns cannot be entirely accounted for by noise. This logic leads to the estimated dispersion of true returns being significantly larger than the typical standard error, at 70 and 50 basis points for the skeptical and moderate estimators, respectively.

The bottom panel of Figure 6 provides an additional reassurance that our estimator is doing something that makes sense. The panel shows the estimated distributions of all in-sample returns (published and not). As the skeptical estimator assumes that the mean true return is zero, it views the published sample (top left panel) as just the right half of the distribution of all narratives. Thus, the skeptical distribution of all narratives (red line) basically reflects the published distribution across zero. In contrast, the moderate estimator (green line) assumes a higher mean of 40 basis points, leading to a less dispersed distribution of all narrative in-sample returns.

4. Robustness: Bias Adjustments in Other Settings

We’ve seen that our bias adjustment is effective at curing selection bias in the simulated data and have shown that our empirical results are consistent with out-of-sample evidence.

This section aims to alleviate remaining concerns about the effectiveness of our bias adjustment. Here we show that the small forecast errors in the demo simulation remain for a variety of other settings, including those under which the estimator’s model is significantly misspecified. Moreover, we show that this robustness results from compensation mechanisms built into the estimator.

Specifically, we simulate 6 different ‘settings’ and apply our bias adjustment in each. Each setting is simply simulated data from a different model, but we use the term ‘setting’ to avoid confusion with the model that is embedded in the bias adjustments.
Table 5 outlines the 6 settings that we examine for robustness. The demo simulation’s parameters are included for ease of comparison. All alternative settings begin with the parameter values in our baseline simulation (Table 2) and then modify the model. The first four of the settings simply alter selected parameters. These deviations are chosen to produce settings in which the main results could conceivably fall apart.

The last two settings add an additional, arguably intuitive, kind of selection bias. In these settings, we assume that the journals have an additional preference for portfolios with high in-sample returns. As the model used in the bias adjustments does not account for this additional bias, these two settings introduce misspecification problems. We discuss the details of this additional bias alongside their respective estimation results in Section 4.3.

Table 5: Alternative Settings for Robustness Checks

<table>
<thead>
<tr>
<th>Name of Model</th>
<th>mean of true returns</th>
<th>dispersion of true returns</th>
<th>dispersion of log std err</th>
<th>increased prob of pub if ( \hat{r}_t &gt; 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Selection Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Sim (for Reference)</td>
<td>0.05</td>
<td>0.20</td>
<td>0.45</td>
<td>1</td>
</tr>
<tr>
<td>Dispersed Standard Errors</td>
<td>0.05</td>
<td>0.20</td>
<td>0.90</td>
<td>1</td>
</tr>
<tr>
<td>Small Selection Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dispersed True Returns</td>
<td>0.05</td>
<td>0.50</td>
<td>0.45</td>
<td>1</td>
</tr>
<tr>
<td>Dispersed and High True Returns</td>
<td>0.50</td>
<td>0.50</td>
<td>0.45</td>
<td>1</td>
</tr>
<tr>
<td>Misspecification Problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High True Returns</td>
<td>0.50</td>
<td>0.20</td>
<td>0.45</td>
<td>1</td>
</tr>
<tr>
<td>Moderate Additional Bias</td>
<td>0.05</td>
<td>0.20</td>
<td>0.45</td>
<td>2</td>
</tr>
<tr>
<td>Large Additional Bias</td>
<td>0.05</td>
<td>0.20</td>
<td>0.45</td>
<td>3</td>
</tr>
</tbody>
</table>

We examine the accuracy of selection-bias adjustments under a variety of settings. Parameter values not shown are the same as those in the baseline simulation (Table 2). ‘Additional Bias’ models have a multiplicative over-representation of \( \hat{r}_t > 10 \) among published returns (see Figure 11).

Figures 7 and 8 show the main results of this section: forecast errors and forecasts under these alternative settings. The bias adjustments work well in every case: regardless of the setting, the adjusted forecasts have small errors centered
Figure 7: Return Forecasting Errors under Alternative Settings. We simulate a variety of models (Table 5), perform our bias adjustments on each simulation, and plot the return forecasting errors (see Figure 1 for more info). For all of the models, adjusted forecasts have small, unbiased errors and are a vast improvement on the standard forecast (in-sample return = true return).
Figure 8: True Returns and Forecasts under Alternative Settings. This figure provides a more detailed look at Figure 7. Portfolios are separated into terciles of standard error, and we plot only the middle tercile due to space. Regardless of the model, the adjusted forecasts (red and green lines) run through the middle of the true returns (blue x’s) and are a vast improvement upon the standard forecast (black lines).
around zero. We discuss each panel in the coming subsections.

4.1. Large Selection Bias Settings

We’ve already seen that our forecast adjustments lead to unbiased forecasts in a setting with large selection bias (the baseline simulation). Here, we simply check that it still works if heteroskedasticity is more pronounced.

The ‘Dispersed Standard Errors’ setting assumes that the dispersion of log standard errors doubles to 0.90 (compared to the baseline value of 0.45). Quantitatively, this means that it’s quite common to find standard errors that are twice as large as the typical standard error of 0.30% per month. This setting could lead to forecast errors, as the estimation simply uses published standard errors to estimate the dispersion of all standard errors (Table 2).

The top left panels of Figures 7 and 8 show that the estimators have no problem dealing with this setting. Both the skeptical and moderate estimators produce unbiased forecasts despite the presence of severe heteroskedasticity.

As in the baseline simulation, this setting implies a large selection bias. The mean of true returns is quite small at 0.05%, while the mean published in sample return is 0.54% (not shown). This large bias means that the moderate estimator is severely misspecified. In spite of this misspecification, the moderate bias adjustment results in good forecasts. This robustness is due to a compensation mechanism that is built into the estimator.

Figure 9 illustrates this compensation mechanism. The bottom panel shows estimated distributions of all in-sample returns (published and not). The mean of the moderate distribution (green dotted line) is far too high, due to the high assumed mean of true returns. This high mean, however, implies that a small dispersion best fits the data on in-sample returns (top-left panel). In contrast, skeptical estimate does not compensate its dispersion (bottom panel, red line), as it more-or-less correctly specifies the mean of true returns.

4.2. Small Selection Bias Settings

It could be that our estimator is geared to fit strong selection bias, and that, given more innocent data, the estimator will be excessively pessimistic. Here we show that this is not the case.
Figure 9: SMM Fit under the Dispersed Standard Errors Setting. This figure provides a detailed look at the estimation for the ‘Dispersed Standard Errors’ setting (Table 5). The top two panels show the quality of the SMM target fits. The SMM targets published data on (1) in-sample returns and (2) standard errors. The bottom panel shows the implied distribution of in-sample return returns, which is critical for the bias adjustment (equations (14), and (15)). The skeptical estimation effective recovers the truth distribution well. The moderate estimation overshoots the mean true return but compensates by underestimating the dispersion, resulting in an effective bias adjustment.
Two of the alternative settings, Dispersed True Returns and Dispersed and High True Returns, exhibit very little selection bias. This small bias can be seen in Figure 7. The black squares in the middle left and bottom left panels are not far from zero, unlike in the baseline simulation. The red x and green triangles in these panels show that the estimators are still quite accurate. Note that these forecast errors are not absolute values, so excessive shrinkage could result in plots below zero.

The modest shrinkage can be seen in Figure 8. The middle left and bottom left panels show that the skeptical and moderate forecasts (red and green lines) have steep slopes, indicating minimal (appropriate) shrinkage. These steep slopes are accurate estimates in that they pass through the center of the cloud of true returns (blue x's).

4.3. Misspecification Problem Settings

As with any econometric method, misspecification has the potential to throw off the results. We've already seen that the estimators work quite well even in the presence of one type of misspecification. In the baseline simulation, the moderate estimator assumed that the mean of true returns was 0.32 percent, significantly above the true value of 0.05 (Table 2). In spite of this misspecification, the moderate estimator produced unbiased forecasts (Figure 1).

Here we examine three more settings in which the model used for estimation is misspecified. The first setting, ‘High True Returns,’ leads to misspecification similar to that in the baseline model. Now, it is the skeptical estimator which is misspecified, as it assumes that the true mean return is zero, far below the true value of 0.50 percent.

Despite the large magnitude of the misspecification, the skeptical adjustment still effectively removes selection bias. This is seen in the top right panels of Figures 7 and 8, which shows that the misspecified estimator still produces forecast errors that are unbiased and small.

Figure 10 shows how the estimator overcomes misspecification. The figure shows SMM estimation details, and in particular, the estimated distributions of all in-sample returns. Misspecification results in the mean of the skeptical distribution (red line) is being far below that of the truth (blue dashed line). The skeptical estimator compensates for this misspecification, however, by overesti-
Figure 10: SMM Estimation Details under the High True Returns Setting. In this setting, mean true return is 0.50 percent per year, far above the skeptical estimator's assumed mean of zero. The skeptical estimate compensates for its misspecified mean by overestimating the dispersion of in-sample returns (red line, bottom panel), leading to less shrinkage (equation 15), and unbiased forecasts of returns (Figure 7).

mating the dispersion of in-sample returns. This large dispersion leads to less shrinkage (equation 15), higher forecasted returns, and, ultimately, the unbiased forecasts seen in Figures 7 and 8.

The last two settings, ‘Moderate Additional Bias’ and ‘Large Additional Bias,’ assume that the data exhibits an additional selection bias that is overlooked by the estimator. Specifically, we assume that if the in-sample return exceeds 0.83 percent per month (10 percent per year), then the probability of publication increases by some factor. The Moderate Additional Bias case assumes the probability of publication doubles while the Large Additional Bias assumes a tripling of publication chances (Table 5).

This additional bias can be thought of as the journals being attracted to dramatic results. The presence of double-digit annualized returns could attract attention and benefit the journal, or, it could attract the subconscious attention of
the editorial staff. While the story might seem plausible, the data does not provide strong indications that this additional bias exists (see Figure 1 of McLean and Pontiff (2016)). Still, we include it to demonstrate the robustness of our estimator.

Figure 11 provides a graphical illustration of this additional bias. The plot shows histograms of in-sample and true returns under the Large Additional Bias setting. The published narratives in the left panel display a large bump in the distribution of in-sample returns (black line) at 0.80 percent, coincident with the tripling of publication probabilities. The distribution of true returns (blue dashed line) is unchanged from the baseline simulation, however, and thus this setting demands even larger shrinkage than the baseline setting.

**Figure 11: Selection Bias in the Large Additional Bias Setting.** In this model, portfolios with in-sample returns > 10 percent per year are three times more likely to be published. This results in the sharp bump in the distribution of published in-sample returns (black line, left panel), and leaves the true returns (blue dashed, left panel) unchanged from the baseline. This leads to more selection bias and misspecification of the model used for estimation, but the estimated adjusted returns are still bias-free (Figure 7).

Figures 7 and 8 show that the selection bias adjustments are effective, in spite of the misspecification problems. The middle right and bottom right panels show that, as before, both the skeptic and moderate adjustments lead to unbiased forecasts.

Figure 12 provides some detail on how the estimator overcomes the misspec-
**Figure 12: SMM Estimation Details in the Large Additional Bias Setting.** In this model, portfolios with in-sample returns > 10 percent per year are three times more likely to be published. The SMM target fit (top panels) is worsened compared to the baseline model (Figure 9), but the skeptical estimator still uncovers the editor’s data quite well (bottom panel). As in the baseline case, the moderate estimate compensates for misspecification of the mean with an underestimated dispersion, leading to unbiased return forecasts (Figure 7).

The figure shows details of the SMM. The top panels show that the SMM target fit is worsened compared to the baseline model as a result of misspecification. Still, the fit looks fairly good. Even with a 3-fold increase in density above returns of 0.83, the decay of in-sample returns in the histogram is primarily driven by the fundamental properties of narrative returns (equations (4) and (6)).

### 5. Conclusion

We find that enhanced interrogation of the CRSP tapes is surprisingly effective at discovering true variation in expected returns. Bias-adjusted returns are
a modest 10-15% smaller than in-sample returns. These results come from the estimation of a simple model of anomaly publication which effectively accounts for selection bias in a wide variety of settings. Since this selection effect is much smaller than the out-of-sample decline in anomaly returns, these results imply that investors learn about mispricing from academic research.
A. Appendix

A.1. Simulation and Estimation Details

Simulations draw 200,000 portfolios. This large number is chosen for reproducibility. Simulations with a smaller number of draws produce similar, but noisier results.

A.2. Estimation Details

The results are insensitive to the histogram construction. A more formal approach could choose some kind of optimal histogram, but we simply choose $N_{\text{hist}} = 8$ with linearly spaced histogram bins that span 98% of the sample.

The minimization is done in two steps. The first step is a grid search. We simply look for the best fit over a grid of 100 values of $\sigma_\mu$ and $E(\sigma_r, i)$. We then use the result of this grid search as an initial guess in Matlab’s fminsearch function, which uses the Nelder-Mead simplex method.

This well-behaved objective means that the two step minimization is stable and fast. Without any tuning, the algorithm converges within a handful of seconds on a run-of-the-mill (government-issued) desktop. The reliability of the estimation also reflects clean identification. The dispersion of true returns controls is tied to the dispersion of published in-sample returns and the mean standard error is linked to, well, the mean published standard error.

A.3. Estimations on Other Variations of the Empirical Sample

Here we show figures from alternative variations of the empirical sample. We consider both all 32 portfolios in the Novy-Marx and Velikov dataset, as well as the 23 portfolios from the published Novy-Marx and Velikov (2016). The 23 portfolios lead to a bimodel distribution as they overemphasize some multiple-sort portfolios. Nevertheless, both variations of the data result shrinkage that is similar to our baseline results with 25 portfolios.
Figure A.13: SMM Fit and Estimated Distributions of All Narrative In-Sample Returns: All 32 Novy-Marx and Velikov Portfolios.
Figure A.14: Distributions of Shrinkage and Bias-Adjusted Returns: All 32 Novy-Marx and Velikov Portfolios.
Figure A.15: SMM Fit and Estimated Distributions of All Narrative In-Sample Returns: 23 Portfolios from Novy-Marx and Velikov (2016).
Figure A.16: Distributions of Shrinkage and Bias-Adjusted Returns: 23 Portfolios from Novy-Marx and Velikov (2016).
References


