JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS Vol. 58, No. 3, May 2023, pp. 968–1004

doi:10.1017/S0022109022000874

# Zeroing In on the Expected Returns of Anomalies

Andrew Y. Chen
*Federal Reserve Board Research and Statistics*
andrew.y.chen@frb.gov

Mihail Velikov
*Pennsylvania State University Smeal College of Business*
velikov@psu.edu (corresponding author)

## Abstract

We zero in on the expected returns of long-short portfolios based on 204 stock market anomalies by accounting for i) effective bid–ask spreads, ii) post-publication effects, and iii) the modern era of trading technology that began in the early 2000s. Net of these effects, the average anomaly's expected return is a measly 4 bps per month. The strongest anomalies net, at best, 10 bps after controlling for data mining. Several methods for combining anomalies net around 20 bps. Expected returns are negligible despite cost mitigations that produce impressive net returns in-sample and the omission of additional trading costs, like price impact.

## I. Introduction

The literature on stock market anomalies has documented more than 100 predictors of the cross-section of stock returns.[1] Using historical data, these papers demonstrate market-neutral returns that average around 8% per year. These anomalies range from those based on past return patterns, to those based purely

This article originated from a conversation with Svetlana Bryzgalova. We thank Marie Briere (HFPE discussant), Jennifer Conrad (the editor), Victor DeMiguel, Yesol Huh, Markus Ibert, Nina Karnaukh, Alberto Martin-Utrera (FDU discussant), R. David McLean (the referee), Andy Neuhierl, Steve Sharpe, Nitish Sinha, Ingrid Tierens (Jacobs Levy discussant), Tugkan Tuzun, Michael Weber, Haoxiang Zhu, and seminar participants at the Federal Reserve Board, Penn State University, University of Georgia, the 11th Annual Hedge Fund and Private Equity Research Conference, 2019 Finance Down Under Meetings, 2019 Eastern Finance Association Meetings, 2019 Jacobs Levy Frontiers in Quantitative Finance conference, and 2020 INFORMS Annual Meeting for helpful comments. We are grateful to Victor DeMiguel, Alberto Martin-Utrera, Francisco Nogales, and Raman Uppal for making their data available to us and we thank Rebecca John for excellent research assistance. 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]Examining academic and industry publications, as well as working papers, Green, Hand, and Zhang (2013), find 333 "return predictive signals" and Harvey, Liu, and Zhu (2016) list 316 "factors." However, many of these variables were not shown to predict the cross-section of stock-level returns. Using more strict definitions, McLean and Pontiff (2016), Green, Hand, and Zhang (2017), and Chen and Zimmermann (2022) identify and replicate 97, 94, and 205 of cross-sectional predictors, respectively.

on accounting variables, and still others based on institutional stock holdings. Few economic risk factors or behavioral theories are so broad that they can make a dent in this wide variety of return predictors.[2]

Anomalies' expected returns, however, may be much lower than the mean returns found in the literature. The literature largely ignores trading costs, which can significantly reduce expected payoffs and thus expected returns. Moreover, the historical data used in these papers are stale. The literature uses data going back to the 1920s, raising questions about whether returns from so long ago are still relevant. Indeed, data-mining bias and investor learning imply that returns in recent years are much smaller (McLean and Pontiff (2016)). In addition, the early 2000s saw a revolution in information and trading technologies that had similar effects (Chordia, Subrahmanyam, and Tong (2014)). Taken together, these findings imply that the data from earlier decades are not representative of the future.

In this article, we zero in on the expected returns of anomalies by accounting for both trading costs and the staleness of historical data. Our main result is that, net of these effects, expected returns are close to zero.

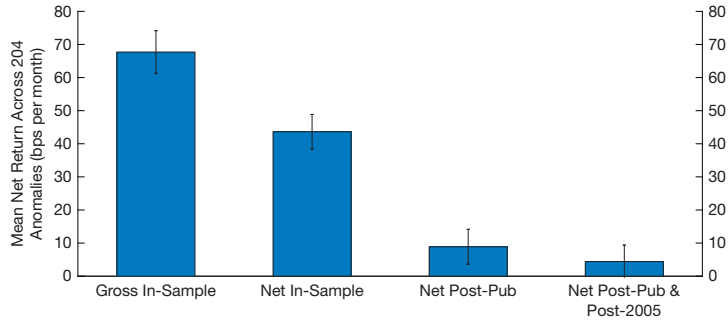
Figure 1 illustrates how we "zero in." To generate this figure, we construct long-short portfolios based on 204 return predictors from Chen and Zimmermann (2022), optimize across cost-mitigation techniques, and reduce portfolio payoffs by half of the effective bid–ask spread whenever a portfolio weight is adjusted. Each bar, moving from left to right, provides a more refined estimate of the average anomaly's expected return.

The first bar is the mean return before trading costs (gross return) within the original papers' sample periods (in-sample), following the original implementations. In our data set, we find an impressive 68 bps per month. Accounting for total trading costs (including cost mitigation) reduces the expected return to 44 bps, which is still a notable 5.3% per year. Adding post-publication effects, however, results in a measly 9 bps per month. Finally, adding the restriction that the sample

FIGURE 1

Anomaly Mean Long–Short Returns

The error bars in Figure 1 show 2 standard errors.




[2]Cochrane (2017) provides an overview of asset pricing theory from a risk-based perspective. Barberis (2018) provides a behavioral perspective.

should only use the modern era of trading technology (post-2005) implies we should expect just 4 bps per month.[3] These results omit additional trading costs such as price impact and short-sale fees. Indeed, short-sale costs average roughly 10–20 basis points per month (Cohen, Diether, and Malloy (2007), Drechsler and Drechsler (2016)) and would likely wipe out the remaining profits.

Though the average anomaly is unprofitable, perhaps the strongest anomalies still offer notable expected returns? Unfortunately, we find that this is not the case.

To zero in on the strongest anomalies, we begin with out-of-sample tests. We replicate McLean and Pontiff's (2016) well-known finding that the gross returns of anomalies decay by roughly 50% post-publication. This decay implies that the 90th percentile anomaly still has a respectable return of about 56 bps per month post-publication. However, if we limit post-publication data to post-2005 observations, the decay increases to 72%. Indeed, after total trading costs (including cost mitigation), the decay is 93%, implying that the 90th percentile anomaly produces an expected return of just 6 bps per month.[4]

Generalizations of the McLean and Pontiff (2016) test lead to similar results. Sorting anomalies on in-sample net returns rather than running regressions leads to similar results, as does sorting on in-sample net Sharpe ratios or turnover. Indeed, none of these sorts produces a reliable pattern in post-publication and post-2005 net returns.

Empirical Bayes estimators provide an intuition for why the strongest anomalies have negligible expected returns (Efron (2012), Azevedo, Deng, Montiel Olea, and Weyl (2019), and Chen and Zimmermann (2020)). These estimators compare the empirical distribution of anomalies to the distribution implied by pure chance, to estimate how much of the heterogeneity is due to luck. We find that the distribution of *t*-stats for post-publication and post-2005 net returns closely resembles a standard normal, with only 7% of *t*-stats exceeding 2.0 in absolute value. Thus, the heterogeneity in mean net returns in recent data can largely be accounted for by luck, and our empirical Bayes estimates imply that the 90th percentile anomaly has an expected return of around 10 bps per month, consistent with our out-of-sample tests.

Even combining anomalies produces small expected returns. We combine anomalies by sorting stocks on the fitted expected gross return from Fama–Macbeth regressions, weighted-average predictor rank, Instrumented Principal Component Analysis (IPCA, Kelly, Pruitt, and Su (2019)), and Least Absolute Shrinkage and Selection Operator (LASSO). Despite the relative simplicity of these strategies, they produce very impressive returns in the 1985–2005 sample. Three out of 4 methods produce a gross return of 380 bps per month, even with microcaps excluded. For comparison, Freyberger, Neuhierl, and Weber's (2020) nonlinear group LASSO produces a gross return of 380 bps over the 1991–2014 sample.[5]

Net of trading costs and stale data, however, the best of our combination strategies earn around 20 bps per month. This weak performance holds even if

[3]We thank Marie Briere for suggesting this analysis. Post-2003 and post-2004 samples lead to similar results.

[4]In our data, the 90th percentile in-sample gross return is 127 bps per month, and the 90th percentile in-sample net return (cost-mitigated) is 91 bps per month.

[5]This gross return does not exclude microcaps but does include the lower returns in post-2005 data.

we optimize over cost mitigations using pre-2006 data, and even though this optimization nets around 180 bps per month pre-2006.

In contrast to our results, Frazzini, Israel, and Moskowitz (2015), Novy-Marx and Velikov (2016), Briere, Lehalle, Nefedova, and Raboun (2019), and DeMiguel, Martin-Utrera, Nogales, and Uppal (2020) find that anomalies remain profitable after trading costs. We reconcile with these studies by accounting for differences in anomaly selection, sample periods, and trading cost measurement. For all of these studies, we find that differences in sample periods account for most of the differences in results. The aforementioned papers all focus on net returns using sample periods that include pre-2006 data, during which anomaly predictability was much stronger.[6]

Our effective spread measurement leads to two additional results that are of independent interest for microstructure researchers. The first is that low-frequency (LF) effective spreads (i.e., spreads calculated from daily CRSP data instead of high-frequency (HF) intraday data) are upward biased compared to the traditional HF spreads by about 25–50 bps post-2005. This bias is seen in all four LF spreads we examine: Hasbrouck (2009), Corwin and Schultz (2012), Kyle and Obizhaeva (2016), and Abdi and Ranaldo (2017). In contemporaneous work, Jahan-Parvar and Zikes (2019) find a similar bias and show that it is closely related to volatility. Taken together, these results suggest that LF spread estimates may no longer be valid in the modern era of electronic trading. Due in part to their accessibility, recent papers in the anomalies literature have used exclusively LF spreads (Novy-Marx and Velikov (2016), DeMiguel et al. (2020), and Freyberger et al. (2020)). To help other researchers use the HF data, we provide easy-to-use code for generating HF spreads that go back to 1983 at https://github.com/chenandrewy/hf-spreads-all.[7]

The second result of independent interest is that averaging LF effective spreads provide a more accurate estimate of HF spreads than any individual LF spread measure. This result is important because trade-level data is largely unavailable before 1983, making LF data the only option for measuring this fundamental trading cost. This improvement is consistent with the literature on economic forecasting, which finds that simple averages of forecasts (or backcasts) often outperform individual forecasts (Bates and Granger (1969), Timmermann (2006)). We make our combined low-frequency measure available at https://sites.google.com/site/chenandrewy/.

The literature on the liquidity effects on anomalies is large,[8] as is the literature that finds that stale data may bias upward expected returns.[9] Among these papers,

[6]We are grateful to Victor DeMiguel, Alberto Martin-Utrera, Francisco Nogales, and Raman Uppal for making their data available to us, and helping us understand this reconciliation.

[7]This code generates HF spreads going back to 1983 in roughly 1 hour by combining WRDS' calculations for TAQ (from the WRDS Intraday Indicators data set) with our own calculations for ISSM spreads.

[8]See Schultz (1983), Stoll and Whaley (1983), Ball, Kothari, and Shanken (1995), Knez and Ready (1996), Pontiff and Schill (2001), Korajczyk and Sadka (2004), Lesmond, Schill, and Zhou (2004), Hanna and Ready (2005), McLean(2010), Frazzini et al. (2015), Hou, Kim, and Werner (2016), Briere et al. (2019), Patton and Weller (2020), and Detzel, Novy-Marx, and Velikov (2021), among others).

[9]Schwert (2003), Marquering, Nisser, and Valla (2006), Huang and Huang (2013), Chordia et al. (2014), Chen and Zimmermann (2020), and Jacobs and Müller (2020).

ours is unique in that it applies high-frequency data, a very large set of anomalies, and data-mining adjustments. All three of these elements are required to make confident estimates of the expected returns on the best anomalies. Low-frequency data overstates trading costs in recent data, a small set of anomalies leads to noisy estimates after excluding stale data, and data-mining adjustments are required to control for the bias that comes from examining the best performers. Huang and Huang (2013) come the closest in spirit to our approach, but they impute trading costs based on statistics reported in the literature and study only 14 anomalies.

We cannot rule out the existence of profitable long-short strategies. A perfectly efficient market is impossible (Grossman and Stiglitz (1980)). Indeed, Freyberger et al. (2020) show that a well-designed non-linear model can net 210 bps per month in the 1991–2014 sample, if trading on microcaps is allowed. This result suggests that profits can still be found in recent data, though they are not easy to find.

Code to reproduce our results is found at https://github.com/velikov-mihail/Chen-Velikov.

Section II describes our methods. Section III presents results for the average anomaly. Section IV examines the strongest anomalies, and Section V combines anomalies. We reconcile our results with selected papers in Section VI. Section VII concludes.

## II. Data and Methods

In Section II we describe our anomalies data, trading cost measurement, and portfolio implementation.

### A. Anomalies Data

Our anomalies data come from Chen and Zimmermann's (2022) "open source" asset pricing project. This project shares code and data to reproduce 205 cross-sectional predictors. We refer to all of these predictors as "anomalies" for simplicity. We exclude one anomaly because it relies on trading only a handful of stocks (institutional ownership among stocks with very high short interest from Asquith, Pathak, and Ritter (2005)), leading to our main data set of 204 anomalies. We use the Apr 2021 data release, downloaded from www.openassetpricing.com.

The CZ data set aims to provide comprehensive coverage of published anomalies. It covers all predictors in Hou, Xue, and Zhang (2020) and all but two predictors from McLean and Pontiff (2016).[10] The CZ data also covers 90% of firm-level cross-sectional predictors with clear evidence of long-short significance from Harvey et al. (2016) and Green et al. (2017).

In contrast, the Novy-Marx and Velikov (NV) (2016) data have a more limited scope, covering "23 of the best known, and strongest performing, anomaly strategies." We will see that the anomalies literature as a whole (as drawn from the CZ data) perform significantly worse than the anomalies selected by NV (Section VI.A).

[10]Hou et al.'s (2020) 452 "anomalies" derive from only 240 characteristics, and only 118 of these showed clear evidence of long-short significance in the original papers.

## B. Direct Trading Cost Measurement

We measure returns before trading costs using the ubiquitous monthly CRSP data. To adjust for trading costs, we track portfolio weights, and each time a position is entered or exited, we assume the effective half spread is paid. This notion of trading costs is also studied in Korajczyk and Sadka (2004), Hanna and Ready (2005), and Novy-Marx and Velikov (2016).

To understand this trading cost measure, it helps to know that CRSP returns are predominantly determined by closing auctions.[11] The hypothetical anomaly strategies studied by academics would have added additional demand or supply to these auctions, increasing the prices for buys and decreasing the prices for sells. These price deviations, then, would reduce returns compared to the CRSP benchmark. Our trading cost aims to measure the minimum amount by which these prices would have been moved.[12] An alternative method for measuring trading costs is to exclusively use intraday data, as in Knez and Ready (1996), but this would deviate significantly from the anomalies literature which is based on closing prices.

Our measure of the minimal price deviation is the effective half bid–ask spread (i.e., the absolute difference between the trade price and the prevailing quoted midpoint). Supposing that the prevailing midpoint is an unbiased estimate of the frictionless price, a buy trade "overpays" by the effective half spread, and a sell trade receives too little by the same amount. Effective spreads use trades that are actually executed and typically imply smaller spreads than quoted prices due to price improvement (Stoll (2003)).

We use high-frequency (HF) data to compute spreads whenever it is available. We coalesce daily spread average spreads from the Daily TAQ, Monthly TAQ, and ISSM data sets. TAQ spreads are calculated by WRDS (from the WRDS Intraday Indicators data set) and ISSM spreads use our own calculations. To match the monthly data frequencies used in the anomalies literature, we first aggregate to a daily level by taking a dollar-weighted average of intra-day spreads, and then aggregate across days within each month by taking a simple average following Abdi and Ranaldo (2017).

Anomaly returns are measured using end-of-month closing prices and thus one may argue that end-of-month spreads are a better match. However, averaging across the month ensures that our spreads are not sensitive to outliers. Moreover, this method is used by previous papers that apply HF trading cost data to anomalies (Hanna and Ready (2005)), and papers that study LF spreads typically compare to monthly averages as well (Abdi and Ranaldo (2017)). For additional details, see Appendix A or https://github.com/chenandrewy/hf-spreads-all.

Our HF data provide a mostly continuous history of transactions on the NYSE and AMEX from 1983–2020.[13] These data sets are sufficient for estimating trading

[11]The NYSE and NASDAQ closing auctions are described at https://www.nyse.com/article/nyse-closing-auction-insiders-guide and https://www.nasdaqtrader.com/content/productsservices/Trading//ClosingCrossfaq.pdf.

[12]We are grateful to Haoxiang Zhu for suggesting this interpretation.

[13]Data for NASDAQ stocks is somewhat shorter (1987–2020), as ISSM is missing NASDAQ data before 1987. The older ISSM data also features several gaps in data. NASDAQ data is missing in Apr. and May 1987, Apr. and July 1988, and Nov. and Dec. 1989. In addition, there are 46 trading days with

costs of anomalies post-publication, as 97% of anomalies are published after 1983 (see Figure C3). However, we also wish to study the effects of cost optimization. To avoid data-mining bias, we run our optimizations on pre-publication data.

Thus, we compute effective spreads pre-1983 (and whenever HF data is missing) using low frequency (LF) proxies based on daily CRSP data. Rather than choose any particular LF proxy, we compute four different LF proxies and use the simple average as our spread. The four LF proxies we use are Hasbrouck's (2009) Gibbs estimate (Gibbs), Corwin and Schultz's (2012) high-low spread (HL), Abdi and Ranaldo's (2017) close-high-low spread (CHL), and Fong, Holden, and Tobek's (2017) implementation of Kyle and Obizhaeva's (2016) invariance-based volume-over-volatility measure (VoV).

This approach is motivated by the idea that the LF proxies are a forecast (or backcast) of the unobserved high-frequency effective spread. The literature on economic forecasting has shown that a simple average of forecasts (a.k.a. combination forecasts) significantly outperforms individual forecasts in a wide variety of settings (Bates and Granger (1969), Timmermann (2006)). This improvement can be understood from a simple diversification argument: the predictive power of a particular forecast varies across observations and combining multiple forecasts averages out these errors. The averaging of multiple LF illiquidity proxies is also used in Karnaukh, Ranaldo, and Soderlind (2015), which finds that averaging improves on using the constituent proxies alone. Indeed, we find that our LF average outperforms any individual LF proxy in terms of its ability to match HF data. For further details, see Appendix B.

Table 1 illustrates the performance of our LF average proxy. Panel A begins by showing that our four LF proxies, while highly correlated, still contain distinct information. The typical correlation is around 75% but can be as low as 0.5 (between HL and VoV). These results suggest that the logic of combination forecasts applies here: by combining proxies we can average out their errors.

Panels B and C of Table 1 show that this logic works. These panels compare our LF average with HF spreads when they are available. The LF average has the highest correlation with TAQ spreads, at 90%. For comparison, the best individual LF proxy is VoV, which has an 85% correlation with TAQ. Panel C shows similar results for ISSM. The LF average has a 92% correlation with ISSM spreads, compared to 88% for the best individual LF proxy, Gibbs.

Though LF spreads are highly correlated with HF spreads, they exhibit a strong bias, especially in recent data. This problem is shown in Figure 2, which plots the median difference between LF and HF spreads over time. Post-2003, spreads are biased upward by 25–50 basis points. This bias indicates that it is important to use HF data to examine trading costs in recent years, and that the LF trading costs used by Novy-Marx and Velikov (2016) overestimate expected costs going forward.

Figure 3 illustrates how our combined effective spread measure has evolved over time. Trading costs rise sharply in the early 1970s as NASDAQ stocks enter the CRSP universe. Costs rise further in the late 1980s, a phenomenon which is seen in

no data for NASDAQ stocks between 1987 and 1991, and 146 trading days with no data for NYSE/AMEX. These data gaps are also found by Barber, Odean, and Zhu (2008).

TABLE 1

## Correlations Between Low-Frequency Proxies and High-Frequency Effective Bid–Ask Spreads

Table 1 examines four low-frequency proxies: Gibbs is Hasbrouck's (2009) Gibbs estimate of the Roll model, HL is Corwin and Schultz's (2012) high-low spread, CHL is Abdi and Ranaldo's (2017) close-high-low, and VoV (volume-over-volatility) is Fong et al.'s (2017) implementation of Kyle and Obizhaeva (2016) microstructure invariance hypothesis. Correlations are pooled. LF_AVE is the equal-weighted average of the four low-frequency proxies. TAQ and ISSM are computed from high-frequency data. The low-frequency measures are imperfectly correlated, suggesting that they contain distinct information. LF_AVE has the highest correlation with high-frequency spreads. Code is found at https://github.com/chenandrewy/hf-spreads-all and https://github.com/velikov-mihail/Chen-Velikov. LF spread data is at https://sites.google.com/site/chenandrewy/.

*Panel A. LF Spread Correlations (1926–2020)*

| | Gibbs | HL | CHL | VoV |
|---|---|---|---|---|
| Gibbs | 1.00 | | | |
| HL | 0.63 | 1.00 | | |
| CHL | 0.74 | 0.86 | 1.00 | |
| VoV | 0.74 | 0.53 | 0.73 | 1.00 |

*Panel B. Correlations with TAQ (1993–2020)*

| | TAQ | Gibbs | HL | CHL | VoV | LF_AVE |
|---|---|---|---|---|---|---|
| TAQ | 1.00 | | | | | |
| Gibbs | 0.84 | 1.00 | | | | |
| HL | 0.64 | 0.60 | 1.00 | | | |
| CHL | 0.79 | 0.72 | 0.85 | 1.00 | | |
| VoV | 0.84 | 0.72 | 0.53 | 0.74 | 1.00 | |
| LF_AVE | 0.90 | 0.89 | 0.82 | 0.93 | 0.86 | 1.00 |

*Panel C. Correlations with ISSM (1983–1992)*

| | ISSM | Gibbs | HL | CHL | VoV | LF_AVE |
|---|---|---|---|---|---|---|
| ISSM | 1.00 | | | | | |
| Gibbs | 0.88 | 1.00 | | | | |
| HL | 0.77 | 0.74 | 1.00 | | | |
| CHL | 0.83 | 0.78 | 0.88 | 1.00 | | |
| VoV | 0.86 | 0.81 | 0.62 | 0.74 | 1.00 | |
| LF_AVE | 0.92 | 0.94 | 0.88 | 0.93 | 0.87 | 1.00 |

FIGURE 2

## The Bias in Low-Frequency Effective Spread Proxies

In Figure 2, we take the difference between low-frequency effective spreads and TAQ effective spreads at the firm-month level and then take the median across firms to calculate the median error in each month. Low-frequency spreads are from Hasbrouck (2009) (Gibbs), Corwin and Schultz (2012) (HL), Abdi and Ranaldo (2017) (CHL), and Kyle and Obizhaeva (2016) (VoV). Post-decimalization, low-frequency proxies are biased upward by roughly 25–50 bps. LF spread data are found at https://sites.google.com/site/chenandrewy/, HF spread code is at https://github.com/chenandrewy/hf-spreads-all, and replication code is at https://github.com/velikov-mihail/Chen-Velikov.

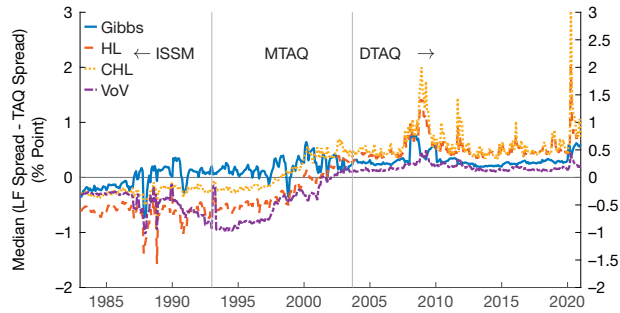
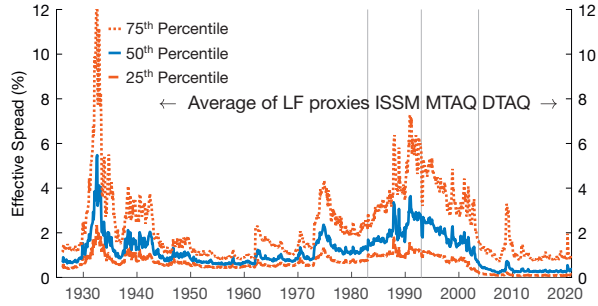


FIGURE 3

Combined Effective Spreads over Time

Spreads in Figure 3 combine high-frequency and low-frequency data. We use high-frequency Daily TAQ (DTAQ), Monthly TAQ (MTAQ), and ISSM when available. Otherwise, we use the average of four low frequency proxies: Gibbs (Hasbrouck (2009)), HL (Corwin and Schultz (2012)), CHL (Abdi and Ranaldo (2017)), and VoV (Kyle and Obizhaeva (2016)). The combined spread tracks well-known structural changes like the entry of NASDAQ (early 1970s) and decimalization (early 2000s).




other papers (Corwin and Schultz (2012), Abdi and Ranaldo (2017)). Trading costs plummet in the 2000s as electronic trading and decimalization have improved liquidity. Overall, our combined effective spread is consistent with key features of stock market history.

## C. Portfolio Implementations

We examine three implementations for each anomaly: i) the implementation in the original paper, ii) a constrained cost optimization that allows for equal-weighting, and iii) a constrained cost optimization that enforces value-weighting.

Implementation is important because trading costs include not only the direct costs of trades (e.g., effective spreads), but also the lost returns that come from avoiding the direct costs (Perold (1988)). Thus, a full accounting of trading costs requires the study of cost optimization. Moreover, the relevant implementation depends on the investor in question, so we study two versions of our constrained optimized implementation.

The implementations used in the original papers are hand-collected by Chen and Zimmermann (2022). The vast majority of the original papers use either equal-weighted portfolios or simple regressions, which also imply equal-weighting (consistent with Green et al. (2013)). Roughly half rebalance monthly, and roughly half rebalance annually. About 20% are discrete signals, about 40% use decile sorts, and the remainder mostly uses quintile sorts, though a few use more coarse groupings. Only a handful use NYSE breakpoints. For further details, see the SignalDoc.csv file in the Chen and Zimmermann (2022) Github repo (accessible via www.openassetpricing.com).

To examine indirect trading costs, we study constrained cost-optimized implementations. For each anomaly, we examine up to 21 cost-mitigated implementations that build on the techniques studied in Novy-Marx and Velikov (2019). The number of cost-mitigated implementations depends on the anomaly, as some cost mitigations are not applicable to the original strategy. For each anomaly, we select the implementation that produces the highest in-sample net return, while including

the original paper implementation in this choice set. We call this "constrained cost optimization," because we optimize over many cost-mitigation techniques, but do not do a full-blown optimization to maintain transparency and limit overfitting.

The cost-mitigated implementations we optimize over include equal- and value-weighted versions of the following three techniques:

- Low-cost universe: This technique limits trading to stocks in the bottom n-tile of trading costs within each NYSE size decile. We consider both the bottom tercile and bottom half of trading costs.
- Reduced rebalancing: This technique simply reduces the rebalancing frequency relative to the one used in the original paper. We consider 3-, 6-, and 12-month rebalancing frequencies.
- Buy/hold spreads (a.k.a. banding): This technique is best described with an example: a 20/40 buy/hold spread goes long stocks with signals that are in the top 20th percentile, but only exits stocks that have signals below the top 40th percentile (and similarly for the short end). The buy/hold spreads we consider depend on whether we are examining equal-weighting or value-weighting. For value-weighted implementations, we examine four buy/hold spreads: 10/20, 10/30, 10/40, and 10/50. For equal-weighted implementations, we examine 20/25, 20/30, …, and 20/50 buy/hold spreads, since equal-weighted 10/20 implementations may overemphasize stocks with very high effective spreads. This technique can only be applied to continuous predictors.

We examine two optimizations: the first choose the implementation with the highest in-sample net return across all implementations described above. The second limits the choices to value-weighted implementations.

Our cost optimizations are clearly constrained, as there are many more ways to implement each individual predictor. Optimizing over additional choices would, by construction, improve performance in-sample, but would lead to more overfitting. We will see, however, that our constrained optimization dramatically improves net returns in-sample, suggesting that the cost of more overfitting outweighs the benefits. These costs tend to be large in portfolio choice (e.g., DeMiguel, Garlappi, and Uppal (2009)).

We emphasize that our cost optimizations use only in-sample information, for similar reasons. Our main object of interest is the mean net return post-publication and post-2005. Optimizing using only in-sample information ensures that our main object of interest is not affected by data-mining bias coming from optimization. Further details are in Appendix C.

## III. Zeroing In on the Average Anomaly

Having described our methods, we can now zero in on expected returns. We begin with the original paper implementations because they are widely understood. We then present our first main result, which examines cost-mitigated implementations.

### A. The Average Academic Implementation

Table 2 shows that the original papers' implementations offer no expected returns at all. Though the historical gross return (in-sample) was 68 bps per month,

TABLE 2

Zeroing In on the Average Anomaly's Expected Return

Table 2 estimates the average net return (e) of 204 anomaly long-short portfolios after accounting for effective bid–ask spreads and stale data. All figures are in bps per month except for turnover, which is in percent per month. Figures average across months and then across anomalies, with standard errors in parentheses. Panel A examines the original papers' implementations. Panels B and C examine constrained cost-optimized implementations (Section II.C). Columns a–d report an approximate net return decomposition. Anomalies are drawn from Chen and Zimmermann's (2022) predictors. After accounting for trading costs and stale data, the expected return is approximately 0.

| | a | b | c | d ≈ b × c | e = a − d |
|---|---|---|---|---|---|
| | Gross Return | Turnover (2-Sided) | Ave. Spread Paid | Return Reduction | Net Return |
| *Panel A. Original Paper Implementations* | | | | | |
| In-sample | 68 | 39 | 206 | 74 | −7 |
| | (3) | (3) | (6) | (7) | (6) |
| Post-publication | 28 | 40 | 85 | 30 | −1 |
| | (4) | (3) | (4) | (3) | (5) |
| Post-pub and post-2005 | 19 | 41 | 68 | 24 | −5 |
| | (2) | (4) | (2) | (2) | (3) |
| *Panel B. Cost-Mitigated Implementation Selected to Maximize In-Sample Net Return* | | | | | |
| In-sample | 61 | 16 | 137 | 17 | 44 |
| | (3) | (2) | (6) | (1) | (3) |
| Post-publication | 16 | 17 | 51 | 7 | 9 |
| | (3) | (2) | (4) | (1) | (3) |
| Post-Pub and Post−2005 | 9 | 17 | 40 | 5 | 4 |
| | (2) | (2) | (3) | (1) | (2) |
| *Panel C. Cost-Mitigated, Value-Weighted Implementations Only, Selected In-Sample* | | | | | |
| In-sample | 47 | 18 | 78 | 12 | 35 |
| | (3) | (2) | (3) | (1) | (3) |
| Post-publication | 5 | 20 | 21 | 4 | 1 |
| | (3) | (2) | (2) | (1) | (3) |
| Post-Pub and Post-2005 | 1 | 20 | 15 | 3 | −2 |
| | (3) | (2) | (1) | (1) | (3) |

one should expect −1 bps per month going forward (net of costs and post-publication). Notably, our large set of anomalies produces a standard error on the post-publication net return of just 5 bps.

Table 2 offers a few decompositions for understanding this lack of expected returns. The post-publication row shows that roughly half of the in-sample gross returns are eliminated by data-mining bias and changes in the investing environment, consistent with McLean and Pontiff (2016). Though this decay is large, post-publication data still imply a notable 28 bps per month of expected returns (4% per year) before trading costs.

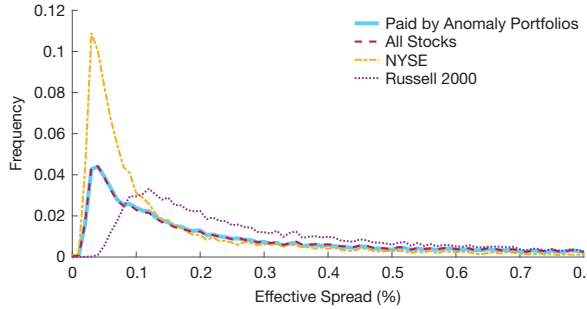
Trading costs wipe out the remaining expected returns, however. A second decomposition shows that this return reduction (column d) is roughly equal to the product of 2-sided turnover (column b) and the average spread paid (column c). As the typical anomaly turns over 20% of its long portfolio and 20% of its short portfolio each month, the total 2-sided turnover is about 40%. Multiplying this turnover by the average paid post-publication spread of 85 bps (column c) leads to a return reduction of about 30 bps, completely eliminating the post-publication gross return.

The large impact of trading costs may be surprising, since decimalization implies that the quoted spread on many stocks is just one penny. Dividing $0.01

FIGURE 4

Distribution of Spreads Paid by Academic Implementations in 2014

In Figure 4, we compare the effective spreads paid by academic implementations with those of all stocks, NYSE stocks, and Russell 2000 stocks. "Paid by anomaly portfolios" pools across all trades implied by 204 academic implementations in 2014. Other distributions are pooled across all stock months in 2014. Academic implementations trade stocks across the entire liquidity spectrum, resulting in large trading costs despite the near-zero modal spreads of recent years.




by the typical share price of $20 leads to a tiny spread of 5 bps, far from the 83 bps post-publication spread paid in Table 2.

Trading costs are extremely right-skewed, however, and anomaly strategies require trading stocks from all over the liquidity spectrum. Thus, the typical spread paid by an anomaly strategy is more similar to the mean spread, and much larger than the modal spread one typically sees at a brokerage.

This skewness is seen in Figure 4, which compares distributions of spreads in 2014. NYSE spreads (dash-dotted line) display a mode at around 5 basis points, consistent with the tiny spread implied by decimalization. The NYSE contains many stocks with much larger spreads, however, as seen in the long right tail of the distribution. Indeed, about 20% of NYSE stocks have effective spreads in excess of 20 bps.

Anomaly portfolios load up on this right tail. The distribution of spreads paid by academic implementations in 2014 (solid line) shares the same mode as the NYSE distribution, but the peak is only half as tall, and the missing mass is shifted into the right tail. As a result, the mean spread paid by anomaly strategies in 2014 is 67 bps, more than 4 times the average NYSE spread of 16 bps.

While academic portfolios tend to trade stocks that are more illiquid than the NYSE, their trading costs are similar to that of the broad universe of stocks. Indeed, the anomaly paid spread distribution (solid line) lines up closely with the distribution for all stocks (dashed line), and is significantly shifted to the left compared with the distribution for the Russell 2000 (dotted line).

Returning to Table 2, the "in-sample" row shows that academic implementations are not even profitable in-sample. Compared to post-publication results, turnover is about the same in-sample, but the average spread paid is more than twice as large, and thus the return reduction more than doubles to 74 bps per month. This return reduction completely wipes out the in-sample gross return of 68 bps per month. Thus, one may consider academic strategies to be too naive, so we investigate cost-mitigated implementations in what follows.

## B. The Average Cost-Mitigated Anomaly

Our cost mitigations are very effective in-sample. Panel B of Table 2 shows that, relative to the academic implementation, cost optimization improves in-sample net returns by 51 bps per month, leading to a noteworthy 44 bps net return. This improvement comes from a 59% decrease in turnover and a 33% decrease in the spreads paid, while the lost gross returns are just 7 bps per month (68–61 bps). Figures C1 and C2 (Appendix C) provide additional details on our cost optimization.

Post-publication, however, the mean net return is just 9 bps per month. This negligible return comes from the fact that the gross return drops to just 16 bps post-publication. Thus, even with a miniscule return reduction of 7 bps, the net return is tiny.

Figure 5 provides a more graphic view of this deterioration. This figure shows our estimates as an event study: we average net returns across 204 anomalies within each month relative to publication (light line). The extreme volatility of the light line is a reminder that anomalies are not at all sure bets.

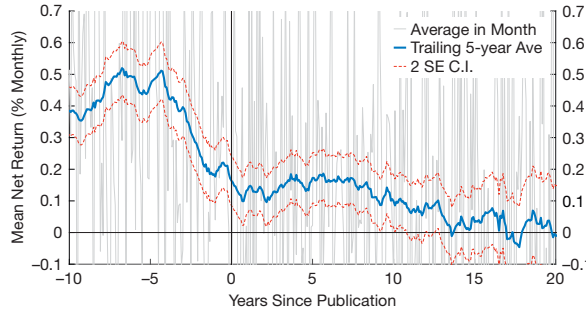
The dark line shows the trailing 5-year moving average net return, once again averaging across 204 anomalies. This moving average declines sharply around publication, dropping from about 45 bps 5 years before publication to around 10 bps afterward.

Returning to Table 2, the "Post-Pub & Post-2005" row further isolates expected returns by accounting for the change in trading technologies that happened during the early 2000s. This change saw an explosion in trading volume and institutional activity, which implies that the data pre-2005 is unlikely to be representative of the future (Chordia et al. (2014)). We account for this change by limiting the data to anomaly months that are both post-publication and post-2005.[14] In this more refined isolation, the typical anomaly is expected to return only 4 bps per month, with a standard error of just 2 bps.

FIGURE 5

Event-Time Net Returns for Cost-Mitigated Implementations

In Figure 5, for a given month relative to publication, light lines plot the mean net return across all anomalies. Dark lines show the trailing 5-year moving average of mean returns, and dashed lines show 2 standard error confidence bounds. Cost mitigation is effective before publication, but net returns become tiny afterward.




[14]Using only post-2003 or post-2004 data leads to very similar results.

Even this tiny 4 bps per month may be unachievable on larger scales, as Panel B of Table 2 examines portfolio implementations that allow for equal weighting. Limiting implementations to those that use value-weighting leads to a negative −2 bps per month of expected returns (Panel C).

## IV. Zeroing In on the Strongest Anomalies

We've seen that the average anomaly's expected return is effectively zero. But what should we expect from the strongest anomalies? This section presents our second main result: The strongest anomalies' expected returns are only around 10 bps per month.

To examine the strongest anomalies, we need to deal with data-mining bias. Data-mining bias comes from the fact that sample mean return of predictor $i$ in recent data can be broken down into two components

$$\bar{r}_i = \mu_i + \varepsilon_i, \tag{1}$$

where $\bar{r}_i$ is the sample mean, $\mu_i$ is the true expected return, and $\varepsilon_i$ is a zero mean noise term due to sampling variability. If we focus on anomalies where $\bar{r}_i$ is larger than the 80th percentile $\bar{r}_{80}$, we get a biased estimate of $\mu_i$:

$$\mathbb{E}(\bar{r}_i|\bar{r}_i > \bar{r}_{80}) = \mathbb{E}(\mu_i|\bar{r}_i > \bar{r}_{80}) + \underbrace{\mathbb{E}(\varepsilon_i|\bar{r}_i > \bar{r}_{80})}_{>0}. \tag{2}$$

The noise term $\mathbb{E}(\varepsilon_i|\bar{r}_i > \bar{r}_{80})$ is positive because mining for large mean returns also selects for large realizations of noise, leading to an upward bias in $\mathbb{E}(\bar{r}_i|\bar{r}_i > \bar{r}_{80})$, compared to the true return $\mathbb{E}(\mu_i|\bar{r}_i > \bar{r}_{80})$.

We examine two approaches to removing the bias $\mathbb{E}(\varepsilon_i|\bar{r}_i > \bar{r}_{80})$. Section IV.A uses out-of-sample tests, and Section IV.B uses an empirical Bayesian adjustment. Though the methods differ, they lead to similar results.

Throughout this section, we refer to "post-publication and post-2005," so to simplify exposition we abbreviate this expression as "post-pub05."

### A. Data-Mining Adjustments Using Out-of-Sample Tests

A simple way to remove the bias in equation (2) is with out-of-sample tests. One can regress post-pub05 mean returns $\bar{r}_i$ on the in-sample mean returns

$$\bar{r}_i = \alpha + \beta \bar{r}_{i,\text{IS}} + \delta_i, \tag{3}$$

and since monthly returns have near-zero autocorrelation ($\text{cov}(\bar{r}_{i,\text{IS}}, \delta_i) = 0$) we have an unbiased estimator

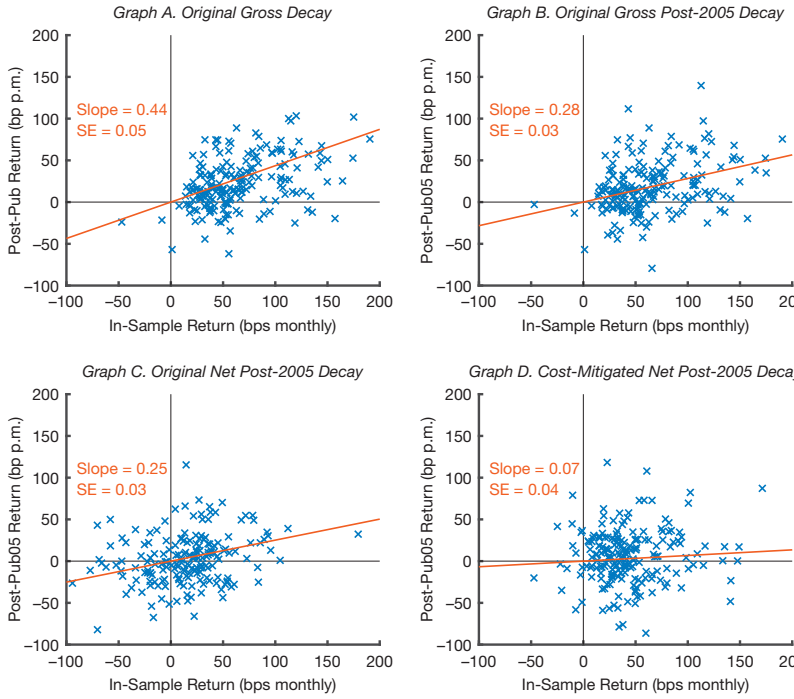
$$\mathbb{E}(\bar{r}_i|\bar{r}_{i,\text{IS}}) = \widehat{\alpha} + \widehat{\beta}\bar{r}_{i,\text{IS}}, \tag{4}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are OLS estimates of equation (3). McLean and Pontiff's (2016) main table can be understood as a refinement of these equations. We then have an unbiased estimate of the strongest mean returns by conditioning equation (4) on a

FIGURE 6

Post-Publication Performance Decay Net of Costs and Stale Data

Each marker in Figure 6 is one anomaly. Fit line is OLS with zero intercept, but allowing for an intercept leads to similar results. Graph A replicates the fact that the original implementations' returns decay by roughly 50% post-publication (McLean and Pontiff (2016), Chen and Zimmermann (2020)). Decay increases to 72% if stale (pre-2005) data is excluded from post-publication performance (Graph B). Decay increases to 93% after trading costs (Graph D), showing even the strongest anomalies have tiny expected returns. Cost mitigation is necessary when accounting for trading costs because many original strategies have negative net returns even in-sample (Graph C).




large $\bar{r}_{i,\mathrm{IS}}$. By applying various refinements of the mean return (a la Figure 1), we can zero in on the strongest anomalies.

Figure 6 shows the results. Graph A begins by regressing post-publication gross returns on in-sample gross returns, using the original papers' strategies. We find a slope of 44%, replicating MP's result that gross returns decay by roughly 50% post-publication (see also Chen and Zimmermann (2020)). As the 90th percentile in-sample gross return is 127 bps per month, this slope implies that the 90th percentile anomaly still has a respectable post-publication gross return of about 56 bps per month after adjusting for data-mining bias.

Graph B shows that expected returns are much smaller after controlling for stale data, however. This panel limits the post-publication data to observations that are also post-2005, and produces a slope of 28%. Thus, a large chunk of the post-publication profitability found by MP and others likely comes from observations that pre-date the modern era of information technology.

Graph C adds trading cost adjustments by replacing gross returns with net returns. The slope drops to 25%, but this decay is difficult to interpret, as many of the in-sample net returns are negative. Like Graphs A and B, Graph C uses the

TABLE 3

The Best Expected Returns Using Out-of-Sample Tests

In Table 3, to account for data-mining bias, we sort anomalies based on in-sample statistics and examine average net returns post-publication and post-2005. All portfolio implementations use cost mitigation following Section II.C. Parentheses denote standard errors. Panel B restricts implementations to value-weighting. There are no robust predictors of post-publication and post-2005 performance, indicating that even the best anomalies have approximately zero expected return.

| | Post-Pub Post-05 Net Return (bps Monthly) | | | |
|---|---|---|---|---|
| | Predictor Quartile | | | |
| In-Sample Predictor | 1 (Worst) | 2 | 3 | 4 (Best) |
| *Panel A. Including Equal-Weighted Implementations* | | | | |
| Net return | 7.5 | −0.9 | 3.1 | 9.5 |
| | (4.2) | (5.5) | (4.5) | (5.3) |
| Net sharpe | 8.2 | −0.0 | 0.5 | 10.5 |
| | (5.6) | (5.2) | (4.1) | (4.6) |
| 1/Turnover | 8.7 | 6.8 | 3.7 | 0.2 |
| | (4.7) | (6.5) | (4.3) | (3.9) |
| *Panel B. Value-Weighted Implementations Only* | | | | |
| Net return | −0.5 | 0.7 | −9.7 | 2.1 |
| | (4.7) | (6.5) | (4.9) | (5.0) |
| Net sharpe | −1.3 | 1.9 | −9.1 | 1.1 |
| | (4.9) | (6.6) | (5.1) | (4.4) |
| 1/Turnover | 1.1 | −4.9 | 1.4 | −5.2 |
| | (5.0) | (6.2) | (4.1) | (5.8) |

original papers' implementations (following MP), leading to large trading costs and many anomalies with negative net returns.

Thus, to effectively study anomaly decay, we need to use cost-mitigated implementations. Graph D shows regressions of post-publication and post-2005 net returns on in-sample net returns where all returns come from the cost mitigation described in Section II.C. The slope drops to 7%, implying a 93% decay in future performance relative to in-sample performance. As the 90th percentile cost-mitigated in-sample net return is 91 bps per month, this regression implies that the strongest anomalies earn expected returns of around 6 bps per month, once data-mining is accounted for.

For robustness, Table 3 examines anomaly sorts. Just as portfolio sorts are a nonparametric version of Fama–Macbeth regressions, the anomaly sorts in Table 3 are a nonparametric form of MP's regressions. The table also extends the sorting variable beyond in-sample net returns by checking the in-sample net Sharpe ratio and in-sample turnover for predictive power.[15]

The table shows that predictability of post-pub05 net returns is weak regardless of the in-sample performance measure. In implementations that allow for equal-weighting (Panel A of Table 3), the best net returns come from using the net Sharpe ratio, with the top quartile producing expected returns of 10.5 bps per month. However, the net returns from this sort are nonmonotonic, and none of the predictors produces a reliable pattern in expected returns. Indeed, turnover actually

[15]In earlier versions of this article, we also find that in-sample return reduction leads to similar results.

predicts net returns with the wrong sign, with high turnover implying *lower* net returns.

Predictability is even worse when using only value-weighted implementations (Panel B of Table 3). Under this restriction, the in-sample net return is the best predictor, and it produces only 2.1 bps per month in its top quartile. Anomalies with low in-sample turnover actually produce negative net returns post-publication and post-2005 (when implementations are restricted to value-weighting).

Overall, post-pub05 mean net returns show little predictability in out-of-sample tests. Taken together, these results lead us to conclude that the strongest anomalies offer at most 10 bps per month, once data-mining bias is accounted for.

## B. Data-Mining Adjustments Using "Empirical Bayes"

As an alternative data-mining adjustment, we study an "empirical Bayes" estimator. Empirical Bayes has been shown to effectively remove data-mining bias in a variety of settings (Efron (2011), Azevedo et al. (2019), Chen and Zimmermann (2020), and Liu, Moon, and Schorfheide (2020)).

Empirical Bayes estimation can be motivated by equation (2). Data-mining bias comes from the noise term $\mathbb{E}(\varepsilon_i|\overline{r}_i > \overline{r}_{80})$. Thus, one can remove bias by directly estimating $\mathbb{E}(\mu_i|\overline{r}_i > \overline{r}_{80})$. In other words, what the econometrician really wishes to know is $\mu_i$ for the strongest anomalies, and thus our goal is not the conditional sample mean $\mathbb{E}(\overline{r}_i|\overline{r}_i > \overline{r}_{80})$, but the conditional expectation of true returns $\mathbb{E}(\mu_i|\overline{r}_i > \overline{r}_{80})$. Given an estimated model, Bayes rule provides the logic for computing this expectation. To generate an estimated model, we specify a DGP and fit it to empirical data using frequentist methods. This combination of empirical frequentist methods and Bayesian logic gives the name "empirical Bayes."

We first develop the adjustment and then examine adjusted expected returns. Throughout this section, we refer to mean returns that are post-publication, post-2005, and net of trading costs. For ease of reading, we drop all of the qualifiers in what follows ("Sharpe ratio" refers to the post-publication, post-2005, net Sharpe ratio).

### 1. Empirical Bayes Model and Estimation

Suppose the Sharpe ratio for predictor $i$ is normally distributed around the true Sharpe ratio

$$\frac{\overline{r}_i}{\sigma_i} \sim N\left(\frac{\mu_i}{\sigma_i}, \mathrm{SE}(\mathrm{SR}_i)\right), \tag{5}$$

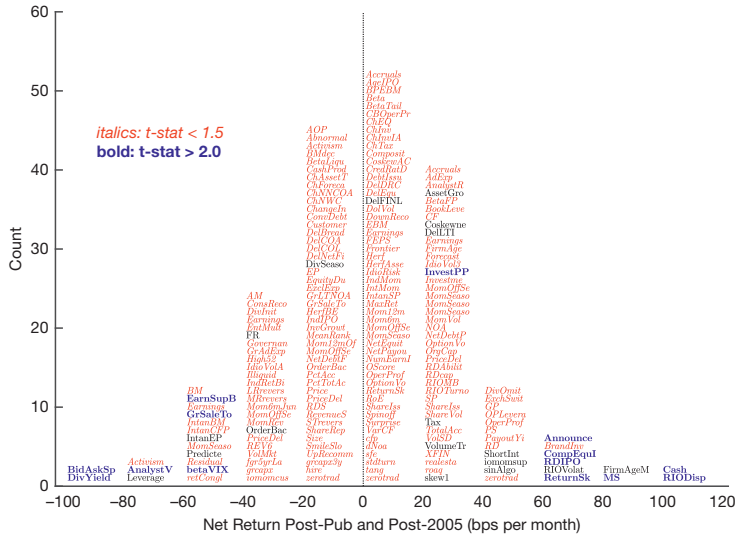
where $\sigma_i$ is the volatility of net returns and $\mathrm{SE}(\mathrm{SR}_i)$ is the standard error for Sharpe ratio $i$. The normal distribution is justified by the central limit theorem and the fact that the sample sizes are in the order of hundreds.

Modeling Sharpe ratios rather than mean returns effectively rescales portfolios to have the same volatility. We find that modeling mean returns leads to even smaller expected returns, consistent with the relatively strong performance of net Sharpe ratios as in Table 3.

FIGURE 7

Distribution of Cost-Mitigated Mean Net Returns Post-Publication and Post-2005

Each acronym in Figure 7 represents one anomaly. For full references, see https://github.com/OpenSourceAP/CrossSection/blob/master/SignalDoc.csv. The distribution closely resembles the null of no predictability, leading to the large data-mining adjustments in Table 4.




We assume $\sigma_i$ is observed. This assumption can be justified by the small standard error in sample volatility for samples of 360 months.[16] Under this assumption and the standard assumption of zero autocorrelation in monthly returns, $\mathrm{SE}(\mathrm{SR}_i) = \mathrm{SE}(\bar{r}_i)/\sigma_i = 1/\sqrt{T_i}$.

True Sharpe ratios are location-scale $t$-distributed

$$\frac{\mu_i}{\sigma_i} \sim t(\mu_{\mathrm{SR}}, \sigma_{\mathrm{SR}}, \nu_{\mathrm{SR}}), \tag{6}$$

where $\mu_{\mathrm{SR}}$ is the location (mean), $\sigma_{\mathrm{SR}}$ is the scale (dispersion), $\nu_{\mathrm{SR}}$ is the degrees of freedom parameter. This bell-shaped distribution is consistent with the data (Figure 7). Using a $t$-distribution allows for fat tails and thus the idea that there may be a few predictors that are truly exceptional.

Equations (5) and (6) summarize the model. The model has just three parameters: $\mu_{\mathrm{SR}}$, $\sigma_{\mathrm{SR}}$, and $\nu_{\mathrm{SR}}$. For simplicity, we fix $\nu_{\mathrm{SR}}$ at different values to examine how our results change.

Given $\nu_{\mathrm{SR}}$, method of moments implies a simple estimate (Xie, Kou, and Brown (2012))[17]

[16] If the monthly return is normally distributed, then sample volatility is $\widehat{\sigma}_i = \frac{\sigma_i}{\sqrt{T-1}}\chi_{T-1}$. Then the standard error of $\widehat{\sigma}_i = 0.037s$ for a sample size of 30 years.

[17] To see this, note

$$\mathbb{E}\left[(\bar{r}_i/\sigma_i - \mu_{\mathrm{SR}})^2\right] = \mathbb{E}\left[(\mu_i/\sigma_i - \mu_{\mathrm{SR}})^2 + (\mu_i/\sigma_i - \mu_{\mathrm{SR}})\delta_i + \delta_i^2\right], \tag{7}$$

where $\delta_i$ is a noise term. The cross term drops out, and then population moments are replaced by sample moments to arrive at (8). Restricting the parameter set to positive $\sigma_{\mathrm{SR}}^2$ results in the max operation.

$$\widehat{\sigma}_{\text{SR}} \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{\overline{r}_i}{\sigma_i}, \tag{8}$$

$$\widehat{\sigma}^2_{\text{SR}} \equiv \max\left\{\left(\frac{\nu_{\text{SR}}-2}{\nu_{\text{SR}}}\right)\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\overline{r}_i}{\sigma_i}-\widehat{\mu}_{\text{SR}}\right)^2-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\right], 0\right\}. \tag{9}$$

Intuitively, the grand mean $\mu_{\text{SR}}$ is estimated using the average of all Sharpe ratios, and the dispersion $\sigma^2_{\text{SR}}$ is estimated as the dispersion in Sharpe ratios $\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\overline{r}_i}{\sigma_i}-\widehat{\mu}_{\text{SR}}\right)^2$ that cannot be accounted for by noise $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}$. The factor $\left(\frac{\nu_{\text{SR}}-2}{\nu_{\text{SR}}}\right)$ adjusts for the assumed fat tail parameter $\nu_{\text{SR}}$. Like other method of moments estimators, weak dependence is sufficient for consistency (Wooldridge (1994)), and the fact that anomaly correlations cluster around zero implies weak dependence (McLean and Pontiff (2016), Chen and Zimmermann (2020), and Chen (2021)).

With estimated parameters in hand, we calculate the bias-adjusted expected return for predictor $i$ with

$$\widehat{\mu}_i \equiv \mathbb{E}\left[\frac{\widehat{\mu}_i}{\sigma_i}\middle|\overline{r}_i, \sigma_i, \widehat{\mu}_{\text{SR}}, \widehat{\sigma}_{\text{SR}}, \nu_{\text{SR}}\right]\sigma_i. \tag{10}$$

That is, the bias adjusted return is the conditional expectation of the true Sharpe ratio given all available information, rescaled by volatility.

Equation (10) is free of data-mining bias, even for predictors with large $\overline{r}_i$. This phenomenon happens because $\widehat{\mu}_{\text{SR}}$ and $\widehat{\sigma}_{\text{SR}}$ are estimated from the entire distribution of predictors, and thus equation (10) conditions on a thorough exploration of the data. In other words, evaluating equation (10) at large $\overline{r}_i$ is equivalent to simulating a large set of predictors, selecting only those with large $\overline{r}_i$, and then computing the mean $\mu_i$.[18]

The mechanics of the adjustment can be seen in the special case $\nu_{\text{SR}} \to \infty$. In this case, normal-normal updating formulas imply

$$\widehat{\mu}_i = \widehat{s}_i\widehat{\mu}_{\text{SR}}\sigma_i + (1-\widehat{s}_i)\overline{r}_i, \tag{11}$$

where the "shrinkage" $\widehat{s}_i$ is given by

$$\widehat{s}_i \equiv \frac{1/T_i}{\widehat{\sigma}^2_{\text{SR}}+1/T_i}. \tag{12}$$

Intuitively, we shrink large $\overline{r}_i$ toward the grand mean $\widehat{\mu}_{\text{SR}}\sigma_i$. Predictors with smaller samples are shrunk more, as they are more vulnerable to data-mining bias. The overall shrinkage is determined by $\widehat{\sigma}_{\text{SR}}$, where in the extreme case that there

[18]This property is related to selection "paradoxes" in Bayesian inference (Dawid (1994), Senn (2008)), though we note that empirical Bayes methods are actually frequentist from a philosophical perspective.

is no dispersion in true Sharpe ratios, shrinkage is 100%. Equation (11) shows our estimator is closely related to the celebrated James and Stein (1961) estimator. Thus, similar estimators can also be derived from quadratic loss arguments, as well as Galtonian reverse regression (Stigler (1990)).

## 2. Empirical Bayes Results

Table 4 describes the estimation results and bias adjusted returns. Panel A shows our baseline cost optimizations, which allow for equal weighting. Assuming true Sharpe ratios are normally distributed, ($\nu_{SR} = 100$), the standard deviation of true Sharpe ratios is 0.15 (annualized). Considering that the mean standard error on the observed net Sharpe ratio is about 0.30, this implies that the bias adjustment is very large (equation (11)). Indeed, 90th percentile adjusted net post-pub05 returns are only about 13 bps per month. Assuming that true Sharpe ratios are fat-tailed ($\nu_{SR} = 4$) has almost no effect on the results. These results are quantitatively close to those from our predictability-based adjustment (Table 3).

Bias adjustments for implementations that use only value-weighting (Panel B) are even stronger. Though the dispersion in the true Sharpe ratios is a bit higher than in Panel A, the mean true Sharpe ratio falls to zero, as the mean net return when restricted to value-weighting is −2 bps per month (Table 2). As a result, the 90th percentile of adjusted net returns is still only about 10 bps per month.

To understand the intuition, it helps to examine the distribution of post-pub05 net returns. Figure 7 plots this distribution. Some anomalies have notable net returns. Cash to assets (Cash), the Mohanram G-score (MS), and momentum for young firms (FirmAgeM) all produce net returns in excess of 80 bps per month in this recent sample.

TABLE 4

The Best Expected Returns Using Empirical Bayes

Table 4 shows the large mean net returns in post-publication and post-2005 (post-pub05) samples for data-mining using empirical Bayes. Bootstrapped standard errors are in parentheses. Adjustments assume Sharpe ratios are the sum of the true Sharpe ratio and an error term, and true Sharpe ratios are *t*-distributed with d.o.f. $\nu_{SR}$, scale $\sigma_{SR}$, and mean $\mu_{SR}$. Given $\nu_{SR}$, we estimate $\sigma_{SR}$ and $\mu_{SR}$ by method of moments (equation (8)). Adjusted expected returns are computed from the conditional expectation of true Sharpe ratios (equation (10)). Even the strongest anomalies have expected returns of only about 10 bps per month, consistent with Table 3.

*Panel A. Including Equal-Weighting Implementations*

| Parameters (Annualized) | | | Bias-Adjusted Net Return (bps Monthly) | | | |
|---|---|---|---|---|---|---|
| Assumed | Estimated | | Percentile | | | |
| $\nu_{SR}$ | $\widehat{\sigma}_{SR}$ | $\widehat{\mu}_{SR}$ | 50 | 70 | 80 | 90 |
| 100 | 0.15 | 0.05 | 4.7 | 8.7 | 10.1 | 13.4 |
| | (0.06) | (0.03) | (2.5) | (3.5) | (3.9) | (5.9) |
| 4 | 0.10 | 0.05 | 4.7 | 7.9 | 9.4 | 12.0 |
| | (0.05) | (0.03) | (2.5) | (3.2) | (3.6) | (5.1) |

*Panel B. Value-Weighted Implementations Only*

| Parameters (Annualized) | | | Post-Pub05 Net Return (bps Monthly) | | | |
|---|---|---|---|---|---|---|
| Assumed | Estimated | | Percentile | | | |
| $\nu_{SR}$ | $\widehat{\sigma}_{SR}$ | $\widehat{\mu}_{SR}$ | 50 | 70 | 80 | 90 |
| 100 | 0.22 | −0.01 | −0.5 | 5.6 | 8.3 | 13.0 |
| | (0.04) | (0.03) | (3.1) | (3.4) | (3.8) | (4.9) |
| 4 | 0.15 | −0.01 | −0.7 | 3.9 | 6.1 | 10.7 |
| | (0.03) | (0.03) | (3.1) | (3.2) | (3.6) | (4.5) |

However, the distribution of anomalies is centered around zero, with a left tail that is comparable in size to the right tail. Indeed, only 14 out of 204 anomalies produce $t$-stats > 2.0 in absolute value, not far from the 10 implied by a model in which there is no predictability ($\sigma_{SR} = \mu_{SR} = 0$). As a result, luck can account for most of the heterogeneity in post-pub05 performance, and Bayesian formulas imply that exceptional mean returns should be shrunk to be close to zero (equation (11)).

## V. Zeroing In on Combination Strategies

Our main results examine trading on single anomalies. This restriction is required to make sharp inferences in the short post-information technology sample. It also dramatically simplifies the analysis, as there are an extremely large number of ways to trade on multiple anomalies. For example, there are 3 billion ways to choose 5 anomalies from the 204 in our data set.

In this section, we zero in on the expected returns of strategies that combine many anomalies. These strategies sort stocks on the expected gross return implied by some linear models and then apply cost mitigation. Though these combinations are simple, we will see that they perform comparably with more sophisticated algorithms in previous papers.

### A. Data Handling for Combining Anomalies

We restrict our analysis to anomalies that were published in 2005 or earlier. This restriction reduces the number of anomalies from 204 to 103, but ensures that our post-2005 results do not inherit look-ahead bias from the original studies. Notably, the existing literature on anomaly combination does not apply this restriction (Green et al. (2017), DeMiguel et al. (2020), and Freyberger et al. (2020)).

We also drop anomalies that are discrete or dominated by missing values at the stock level. In particular, we require the anomaly characteristic is observed for 50% of stocks with market cap observations in Jan. 1975. Dropping discrete anomalies reduces the list to 79 anomalies, and the missing value screen reduces the list to 58 anomalies. DeMiguel et al. (2020) also find that requiring nonmissing observations requires dropping many anomalies.

To focus on moderate-to-high liquidity stocks, we drop stocks with market cap below the 20th percentile in the current month. This screen is also used by DeMiguel et al. (2020) and Brandt, Santa-Clara, and Valkanov (2009). It also eases comparison with Green et al. (2017) and Freyberger et al. (2020), who use a similar screen in a subset of their results.

After these screens, we transform each anomaly characteristic by ranking stocks within each month, dividing by the number of stocks, and then subtracting 0.5. This transformation implies that each characteristic lies in the interval $[-0.5, +0.5]$ and eliminates sensitivity to outliers. This normalization is also used in Kelly et al. (2019), and a similar normalization is used in Freyberger et al. (2020).[19]

[19]In contrast, Green et al. (2017) and DeMiguel et al. (2020) winsorize all variables and then standardize to have mean zero and unit standard deviation. We did not choose this approach as some anomaly characteristics are sometimes transformed with logs (e.g., B/M in Fama and French (1992)) and it is unclear if a log or other transformation should be used.

We also impute missing values as zero, which is equivalent to imputing the raw characteristics with the cross-sectional medians (after imposing the missing value and other screens).

## B. Methods for Combining Anomalies

Our combination strategies use a linear model of expected gross returns

$$\mathbb{E}_t(r_{i,t+1}) = \beta_0 + \sum_{j=1}^{J} \beta_j x_{i,j,t}, \tag{13}$$

where $r_{i,t+1}$ is the gross return of stock $i$ in month $t+1$, $J$ is the total number of predictors, $\beta_j$ is the coefficient for predictor $j$, and $x_{i,j,t}$ is the rank of the $j$th anomaly characteristic for stock $i$ in month $t$.

We examine four ways of fitting equation (13):

1. Fama–Macbeth regressions of future returns on characteristics: This well-understood method serves as our benchmark anomaly combination.

A potential weakness of Fama–Macbeth is that it may overfit expected returns as the number of characteristics we use is large. This concern may be minor in the pre-2005 sample, but the fact that predictability has declined sharply in recent data suggests that regularization *might* improve performance.

2. Weighted Average Rank: We assume $\mathbb{E}_t(r_{i,t+1})$ is the weighted average of the characteristic rank $x_{i,j,t}$ across anomalies $j$, where the weight for anomaly $j$ is the mean gross long-short return from sorting on anomaly $j$ over the past 120 months of data.[20]

This assumption can be thought of as a form of model averaging, where each portfolio sort is thought of as a single model of expected returns. Taking the weighted average of ranks is then equivalent to taking the simple average of each individual-anomaly expected returns. Model averaging has been shown to improve equity premium forecasting and can be thought of as an easy-to-use regularization method (Rapach, Strauss, and Zhou (2010), Rapach and Zhou (2013)).

3. Instrumented Principle Component Analysis (IPCA): This method, proposed by Kelly et al. (2019), combines equilibrium factor models with characteristics-based expected return measurement. It jointly estimates stock-level expected returns, stock-level factor exposures, and latent factors, while using characteristics data as instruments.

We use the version of their model that imposes the equilibrium condition that all cross-sectional variation in expected returns is due to factor exposure. This approach may outperform Fama–Macbeth if an equilibrium pricing restriction holds. Moreover, as we assume the number of latent factors is only 5, this method can also be thought of as a regularization method.

[20] The past 120 month requirement is imposed for consistency with the other fitting methods. We find that using a longer sample leads to better performance for this algorithm, bringing its performance closer to the other algorithms, but the overall results are similar.

4. LASSO regression: Weighted average rank and IPCA can be thought of as dense regularization methods, as they draw on information found throughout all anomalies. But if the true structure of predictability is focused on select anomalies, then a method that imposes this structure will be more statistically efficient. LASSO regularization is a technique that imposes this structure.

We choose the LASSO penalty (often called $\lambda$) to minimize the MSE estimated by 5-fold cross-validation (James, Witten, Hastie, and Tibshirani (2013)).

Further details of the model fitting are in our replication code, provided at https://github.com/velikov-mihail/Chen-Velikov.

Using these four fitting methods, we form "out-of-sample" trading strategies. For each month beginning in Jan. 1985, we fit equation (13) using the past 120 months of data. We then sort stocks on the fitted expected return $\sum_{j=1}^{J}\beta_j x_{i,j,t}$ and form a long-short portfolio that is held for 1 month.

We examine two methods for forming the long-short portfolio. The first method simply goes long stocks in the top decile of $\widehat{\mathbb{E}}_t(r_{i,t+1})$ and short stocks in the bottom decile, with equal weighting in each leg. This method aids comparison with Green et al. (2017) and Freyberger et al. (2020), who also examine equal-weighted long-short decile portfolios formed on fitted expected returns with small cap screens.

The second method applies constrained cost-optimization following Section II.C. We examine 21 cost-mitigated implementations building on Novy-Marx and Velikov (2019) and then select the technique that produces the highest net return in the 1985–2005 sample. This cost optimization is applied for each of the four model-fitting methods, allowing the cost mitigation to depend on the type of regularization used. Importantly, we do not use post-2005 data in this optimization, implying that the mean returns post-2005 should be free from look-ahead bias.

## C. Results from Combining Anomalies

Table 5 shows the results. Each panel shows the performance of one fitting method, broken down into the 1985–2005 and 2006–2020 subsamples.

Panel A of Table 5 begins with equal-weighted decile sorts based on Fama–Macbeth regressions. This simple approach produces an extremely large gross return of 374 bps per month in the 1985–2005 sample, despite the fact that we exclude stocks below the 20th percentile of market cap. This gross return exceeds the 280 bps gross return found by Green et al. (2017) using similar methods on a similar sample, suggesting that the Chen and Zimmermann (2022) data set contains more predictive power than the Green et al. (2017) data.

Investors should not expect 374 bps per month, however, as this return accounts for neither trading costs nor stale data. Trading costs reduce the net return to 188 per month, even with cost mitigation. Then removing the stale data from 2005 or earlier, the net return drops to only 21 bps per month.

Other fitting methods show the same patterns, as seen in Panels B–D of Table 5. Whether we use the weighted average rank, IPCA, or LASSO, cost mitigation nets around 180 bps per month in the 1985–2005 sample, but this net return drops to around zero post-2005. Indeed, the 21 bps per month obtained

TABLE 5

Performance of Long–Short Strategies that Combine Anomalies

Table 5 sorts stocks on the expected gross return implied by various models using the 58 predictors that are published pre-2006 and satisfy availability and continuity conditions. Fits use the past 120 months of data, and stocks below the 20th percentile market cap are dropped. Combination strategies net nearly 200 bps pre-2006, but net returns are at best around 20 bps per month post-2005. Gross and net returns are bps per month. Turnover is 2-sided (% monthly). Parentheses denote standard errors.

| | Gross | Turnover | Net |
|---|---|---|---|
| *Panel A. Fama–MacBeth* | | | |
| Equal-weighted decile sorts on fitted expected returns | | | |
| 1985–2005 | 374 | 64 | 86 |
| | (35) | | (35) |
| 2006–2020 | 80 | 55 | 31 |
| | (35) | | (35) |
| With cost mitigation selected to maximize net return in 1985–2005 | | | |
| 1985–2005 | 246 | 13 | 188 |
| | (34) | | (34) |
| 2006–2020 | 31 | 12 | 21 |
| | (35) | | (35) |
| *Panel B. Average Rank* | | | |
| Equal-weighted decile sorts on fitted expected returns | | | |
| 1985–2005 | 276 | 28 | 149 |
| | (47) | | (47) |
| 2006–2020 | 4 | 26 | −22 |
| | (42) | | (42) |
| With cost mitigation selected to maximize net return in 1985–2005 | | | |
| 1985–2005 | 240 | 14 | 177 |
| | (46) | | (46) |
| 2006–2020 | −27 | 15 | −41 |
| | (40) | | (40) |
| *Panel C. IPCA* | | | |
| Equal-weighted decile sorts on fitted expected returns | | | |
| 1985–2005 | 379 | 63 | 94 |
| | (39) | | (39) |
| 2006–2020 | 46 | 44 | 3 |
| | (36) | | (36) |
| With cost mitigation selected to maximize net return in 1985–2005 | | | |
| 1985–2005 | 242 | 12 | 185 |
| | (37) | | (37) |
| 2006–2020 | 12 | 10 | 2 |
| | (34) | | (34) |
| *Panel D. LASSO* | | | |
| Equal-weighted decile sorts on fitted expected returns | | | |
| 1985–2005 | 375 | 53 | 128 |
| | (37) | | (37) |
| 2006–2020 | 74 | 41 | 37 |
| | (29) | | (29) |
| With cost mitigation selected to maximize net return in 1985–2005 | | | |
| 1985–2005 | 243 | 12 | 187 |
| | (36) | | (36) |
| 2006–2020 | 28 | 11 | 18 |
| | (28) | | (29) |

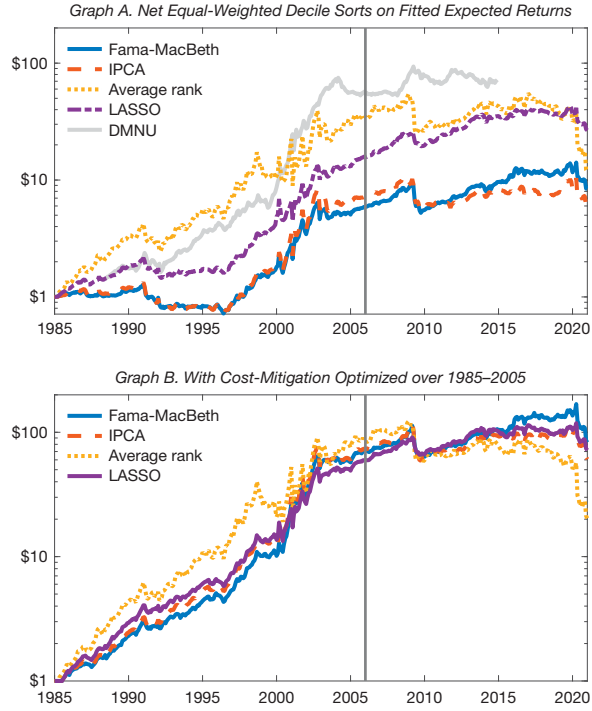
by Fama–Macbeth represents the highest mean return among the cost-mitigated combination strategies in the post-2005 data.

Interestingly, Table 5 shows that cost-mitigated strategies consistently underperform their unmitigated counterparts in the post-2005 sample. This underperformance is far from statistically significant, however, at roughly 10 bps per month, representing

FIGURE 8

Cumulative Return of $1 Invested in Combination Strategies

Figure 8 sorts stocks on the expected gross return implied by various models using 58 predictors that are published pre-2006 and satisfy availability and continuity conditions. Fits use the past 120 months of data and stocks below the 20th percentile market cap are dropped. Graph A shows results without cost mitigation. Graph B optimizes costs using data from 1985–2005. For comparison, "DMNU" shows the market-neutral component of DeMiguel et al.'s (2020) regularized out-of-sample portfolio, scaled to have the same volatility as our Fama–Macbeth strategy. For all strategies, impressive gains flatten out around 2003, and the 2005 cutoff matters little.




only about one quarter of a standard error. Nevertheless, this underperformance is striking because cost mitigation far *out*performs in the 1985–2005 sample. These results are consistent with the idea that arbitrageurs trade more on the liquid anomaly strategies, disproportionately reducing the net returns for liquid stocks.

Figure 8 takes a closer look by plotting the performance of a $1 investment in the combination strategies beginning in 1985. The figure shows that the impressive gains in the earlier sample flatten out post-2005. The inflection point occurs around 2003, suggesting that the precise choice of the 2005 cutoff matters little.

The figure also shows an interesting change in cyclicality post-2005. While the combination strategies suffered little to no losses around the 2001 dot-com crash, all strategies suffered large losses just after the 2008–2009 financial crisis and the 2020 coronavirus market crash. These results echo Daniel and Moskowitz's (2016) finding that momentum crashes following market declines.

Readers familiar with the trading cost literature may wonder how our results compare with those of DeMiguel et al. (DMNU) (2020). In a seminal contribution, DMNU show how "trading diversification" can improve performance when combining anomalies. As trading diversification effects should be found whenever

trading signals are not perfectly correlated (DMNU's Proposition 3), these effects are surely found in our results too.

To take a closer look at this issue, Figure 8 compares the performance of our combination strategies to DMNU's regularized out-of-sample portfolio. Since DMNU's portfolio combines the CRSP index with a set of "long-short" weights that sum to zero, we subtract the CRSP index to make the DMNU returns comparable to ours. We also normalize the volatility of this portfolio to match that of our Fama–Macbeth based strategy. Since DMNU's portfolio starts in 1987, we start this portfolio with the same dollar value as the average value of our four strategies in 1987.[21]

The figure shows that DMNU's optimal portfolio far outperforms our unmitigated combination strategies in the 1987–2005 sample.[22] This outperformance does not persist after 2005 however. Like our combination strategies, DMNU's portfolio seems to reach an inflection point in the early 2000s, after which its gains flatten to a slope similar to our combination strategies.

A couple of caveats are required in this comparison. The first is that our combination strategies draw on the Chen and Zimmermann (2022) data set, which seems to have more predictive power than the Green et al. (2017) characteristics used by DMNU (see discussion above). The second is that the DMNU results use direct trading costs based on Hasbrouck's (2009) low-frequency Gibbs measure (see Brandt et al. (2009)), which tends to imply higher trading costs relative to our combined trading cost measure (Figure 2). Nevertheless, the fact that performance is much worse in more recent data is robust and helps square our findings with the large Sharpe ratio found by DMNU.

Overall, our results suggest that even combining anomalies will lead to little or no economic profits going forward. Combination strategies net nearly 200 bps per month in the early sample. But after the information technology revolution, the best combinations net only around 20 bps per month.

## VI. Reconciliation with Related Papers

In contrast to our results, some recent papers find that anomaly profits still exist after trading costs. In this section, we show that our results differ largely because we account for stale pre-2006 data.

### A. Reconciliation with Novy-Marx and Velikov (2016) Findings

In contrast to our findings, Novy-Marx and Velikov (2016) (NV) find that anomalies with low to moderate turnover generate significant net returns. Our studies differ, however, in that NV's study examines a smaller selection of the "best known, strongest performing anomalies" and does not aim to isolate the returns seen in more recent data. Table 6 shows that using a comprehensive set of anomalies or focusing on more recent data leads to lower net returns.

[21]We are grateful to Victor DeMiguel, Alberto Martin-Utrera, and Raman Uppal for sharing their data and making this analysis possible.

[22]Our cost-mitigated strategies exhibit look-ahead bias in the pre-2006 data, which makes them not comparable to DMNU's out-of-sample optimizations.

TABLE 6

Reconciliation with Novy-Marx and Velikov (2016)

In Table 6, we vary the anomaly selection, implementation, sample period, and direct cost measurement to reconcile with Novy-Marx and Velikov (NV) (2016). Returns are in bps per month. NV (2016) anomalies are 23 of the "best known, and strongest performing anomalies." CZ (Forth) anomalies are 204 anomalies covering the majority of the literature. Column 3 limits the CZ anomalies to continuous ones, in order to apply decile sorts. Low turnover is below 10%, mid turnover is between 10% and 50%, and high turnover is above 50% (1-sided, monthly). Anomaly selection and sample period both contribute to the lower net returns found in our results compared to NV (2016).

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Anomaly Selection | NV (2016) | NV (2016) | CZ (Forth) | CZ (Forth) | CZ (Forth) |
| Implementation | VW NYSE Deciles | | | Original Paper | |
| Sample | Full | Post-Pub05 | Full | Post-Pub05 | Post-Pub05 |
| Direct Costs | Gibbs | Gibbs | Gibbs | Gibbs | Combined |
| *Panel A. Mean Gross Return (bps monthly)* | | | | | |
| Low turnover | 36 | −20 | 20 | 11 | 11 |
| Medium turnover | 78 | 27 | 54 | 28 | 28 |
| High turnover | 96 | 45 | 58 | 30 | 30 |
| All | 68 | 15 | 34 | 19 | 19 |
| *Panel B. Mean Net Return (bps monthly)* | | | | | |
| Low turnover | 28 | −25 | 12 | 1 | 4 |
| Medium turnover | 34 | −1 | 12 | −16 | −3 |
| High turnover | −35 | −39 | −60 | −92 | −46 |
| All | 14 | −19 | 3 | −17 | −5 |

The table begins by reproducing the results of NV. Column 1 uses the anomaly characteristics listed in NV's Table 2 instead of the CZ data, as well as the portfolio implementation and trading cost measure used by NV. As in NV, this method leads to moderate net returns of 30–35 bps per month over the full sample for all but the highest turnover anomalies.

Column 2 shows that limiting the data to the post-publication and post-2005 sample renders these strategies unprofitable. In this more recent sample, low turnover anomalies have negative *gross* returns of −20 bps per month, and mid turnover anomalies gross only 27 bps per month. As a result, all turnover groups earn negative net returns in more recent data.

Column 3 shows that expanding the anomalies data to Chen and Zimmermann's (2022) comprehensive data set (CZ) also leads to weaker performance. Applying the methods used in NV to the 168 continuous CZ anomalies leads to gross returns of 20 and 54 bps per month for low and mid turnover anomalies, respectively, about 20 bps smaller than the gross returns for NV's anomalies. Accordingly, the net returns of the larger anomaly set are smaller by about 20 bps per month, leading to negligible or negative net returns for all turnover groups. Intuitively, the anomalies in the broader zoo are significantly weaker than the strongest anomalies studied in NV.

In contrast, the measure of effective spreads has relatively little effect. Comparing columns 4 and 5, we see that the Gibbs trading cost measure implies much more negative net returns for high turnover anomalies, but its effects on mid- and low-turnover anomalies are more limited. As a result, using the Gibbs measure still implies that anomalies have near-zero net returns.

## B. Reconciliation with Institutional Trading Cost Studies

Trading costs depend on the trader under consideration. This issue is highlighted in Frazzini et al. (2015) (FIM), who argue that size, value, and momentum are quite profitable for a large institutional investor that they study. Briere et al. (2019) (BLNR) find similar results using ANcerno data, which covers more than 500 global institutional investors. These authors argue that their measured profits are larger than those in studies that focus on average effective spreads (e.g., Lesmond et al. (2004), Novy-Marx and Velikov (2016)) because these spreads are likely higher than the spreads paid by large and sophisticated institutions.

Extending this argument, one might conjecture that our results would change significantly if we measured trading costs following FIM and BLNR. In this subsection, we present evidence suggesting that our results would not change significantly. In particular, we show our trading costs turn out to be quite similar to those found by FIM and BLNR, once the sample period and anomaly selection are accounted for.

Table 7 examines the performance of size, B/M, and momentum as measured by our methods ("Cost-Mitigated") and compares it to FIM and BLNR. Importantly, we subset performance measurement based on two subsamples: 2006–2020 and 1998–2013. The 2006–2020 sample is equivalent to the post-publication and

TABLE 7

Reconciliation with Institutional Trading Cost Studies

In Table 7, returns are in bps per month. Parentheses denote standard errors. The "Cost-Mitigated" column uses our calculations. FIM (2015) is from Table IV of Frazzini et al. (2015), BLNR (2019) is from Table 7 of Briere et al. (2019), and PW (2020) is from Table 2 ($N_S = 100$) of Patton and Weller (2020). These studies measure costs with institutional executions rather than the aggregate TAQ data we use. Performance is similar across studies if the sample period is controlled for, however, performance is much worse in the 2006–2020 sample. Individual anomaly performance is highly sensitive to the sample period, and thus we need many anomalies to estimate expected returns post-2005, as we do in Tables 2 and 4.

*Panel A. Size*

| Return | Cost-Mitigated | | FIM (2015) | BLNR (2019) | PW (2020) |
|---|---|---|---|---|---|
| (bps p.m.) | 2006–2020 | 1998–2013 | 1998–2013 | 1999–2011 | 1993–2016 |
| Gross | −10.9 | 55.1 | 66.5 | 44.8 | 16.3 |
| | (30.2) | (37.9) | (22.1) | (28.1) | (20.2) |
| Net | −14.5 | 48.2 | 54.3 | 43.4 | 18.3 |
| | (30.1) | (37.9) | (21.9) | | (19.9) |

*Panel B. B/M*

| Return | Cost-Mitigated | | FIM (2015) | BLNR (2019) | PW (2020) |
|---|---|---|---|---|---|
| (bps p.m.) | 2006–2020 | 1998–2013 | 1998–2013 | 1999–2011 | 1993–2016 |
| Gross | −3.7 | 58.7 | 40.5 | 25.3 | 45.3 |
| | (20.6) | (28.6) | (36.2) | (31.2) | (23.4) |
| Net | −8.8 | 49.0 | 29.3 | 23.3 | 19.3 |
| | (20.8) | (28.8) | (36.6) | | (23.2) |

*Panel C. Momentum*

| Return | Cost-Mitigated | | FIM (2015) | BLNR (2019) | PW (2020) |
|---|---|---|---|---|---|
| (bps p.m.) | 2006–2020 | 1998–2013 | 1998–2013 | 1999–2011 | 1993–2016 |
| Gross | 6.1 | 22.4 | 18.8 | 45.9 | 50.1 |
| | (67.4) | (74.6) | (47.1) | (47.4) | (31.3) |
| Net | 0.3 | 11.5 | −6.4 | 23.1 | 14.4 |
| | (67.4) | (74.4) | (45.8) | | (32.0) |

post-2005 sample for these anomalies, as all of them are published before 2006. The 1998–2013 sample period is selected to match FIM's sample, and is similar to BLNR's 1999–2011 sample period.

The table shows that size, B/M, and momentum are all unprofitable in the 2006–2020 sample, even before trading costs. The *gross* returns of these anomalies are −11, −4, and + 6 bps per month, respectively, over this recent sample. Thus, the TAQ effective spreads we use have a minimal effect on our conclusion about the poor expected returns of these classic anomalies.

Performance looks much different in FIM's 1998–2013 sample period, however. Here, size and B/M have notable gross returns, and thus our calculations produce net returns of 48 and 49 bps per month for these anomalies, respectively. This measurement is quantitatively close to FIM's finding of 54 and 29 bps per month for size and B/M, respectively, and close to BLNR's estimates of 43 and 23 bps per month. Our estimates find momentum nets only 12 bps per month between 1998 and 2013, but FIM also find that momentum performs poorly here too. Indeed, they find negative returns over this sample period.

The similarity in net returns is likely due to offsetting effects from trading cost measurement. Our costs use the average effective spread, which are likely higher than spreads used by FIM and BLNR, but we omit price impact, leading to opposite effects. Which effect dominates depends on details including the assumed size of the trades. It turns out that the size of trades assumed by FIM and BLNR lead to net returns that are quite similar to what we find – as long as the sample period is taken into account.

A caveat in this comparison is that FIM and BLNR's results suggest that a sophisticated institution may still be able to profitably trade on anomalies by strategically using limit orders, if the institution is satisfied with a smaller portfolio than the ones assumed by FIM and BLNR.

Table 7 also compares our net returns for size, B/M, and momentum to those measured by Patton and Weller (2020) (PW). We focus on their 1993–2016 results for closer comparability with the sample period used by FIM and BLNR. In this sample period, PW's net returns are typically smaller than those found by FIM And BLNR, but they remain positive. Accordingly, PW's net returns are also smaller than those found in our 1998–2013 sample, perhaps because mutual funds seem to avoid short positions (An, Huang, Lou, and Shi (2021)).

Overall, these results highlight the importance of isolating the sample period and accounting for many anomalies when trying to make inference on expected returns. The performance of individual anomalies can change dramatically between older and more recent samples, and the standard error on an individual anomaly is on the order of 50 bps per month over a 15 year sample. However, by aggregated information across many anomalies, one can make more precise estimates, as we do in our main results.

## VII. Conclusion

We zero in on the expected returns of anomalies by accounting for trading costs and the staleness of historical data. Net of these effects, the expected return on

even the best anomalies is effectively zero. Indeed, we find that even combining anomalies lead to meager net returns in recent data.

These results come from applying data-mining adjustments to data that includes high-frequency trading costs and a very large set of anomalies. High-frequency data is necessary as low-frequency spreads are biased upward in recent years. A large set of anomalies is required as individual anomaly returns are very noisy after excluding stale data. Finally, data-mining adjustments are required to control for the bias that comes from selecting the best anomalies. Compared to Frazzini et al. (2015), Novy-Marx and Velikov (2016), and DeMiguel et al. (2020), our study is unique in combining these data sets and methods.

Taken with other recent papers, our results provide a complete accounting for the average return on the anomaly zoo. Previous papers show that about 15% of the average gross return is due to publication bias (McLean and Pontiff (2016), Chen and Zimmermann (2020)). We find that trading costs account for another 40% and that the remaining net returns (45%) are traded away over time, consistent with the idea that mispricing is removed as information proliferates and technology improves (Chordia et al. (2014), McLean and Pontiff (2016)).

This decomposition paints a picture of a dynamic equilibrium process, but one more in line with Lo's (2004) adaptive market hypothesis or "efficiently inefficient" markets (Grossman and Stiglitz (1980), Gârleanu and Pedersen (2018)) than standard dynamic equilibrium models (e.g., Campbell and Cochrane (1999)). Every month, researchers find imperfections in the existing market equilibrium. As information about predictability diffuses and trading technology improves, the net returns of these imperfections are traded away, leading to a new equilibrium.

## Appendix A. Details of High-Frequency Data

The HF effective spread for the $k$th trade of a given stock is

$$[\text{EFFECTIVE\_SPREAD}]_k = 2|\log(P_k) - \log(M_k)|, \tag{A-1}$$

where $P_k$ is the price of the $k$th trade and $M_k$ is the midpoint of the matched consolidated best bid and offer (BBO) quote.

Daily TAQ (DTAQ) and Monthly TAQ (MTAQ) spreads come from daily dollar-weighted average spreads from WRDS Intraday Indicators (WRDS IID). DTAQ observations follow Holden and Jacobsen (2014) (EFFECTIVESPREAD_PERCENT_DW). For MTAQ observations before 1999, we assume a 1 second quote delay (ESPREADPCT_VW1), and a 0 second delay (ESPREADPCT_VW0) otherwise. We do not use WRDS IID's interpolated MTAQ spreads because we found that they have much lower pooled correlations with DTAQ spreads, perhaps because we examine all stocks rather than the volume-stratified sample used in Holden and Jacobsen (2014).

WRDS applies its own data screens following Holden and Jacobsen (2014), but we found that some outliers remain. To remove these remaining outliers, we remove firm-day effective spreads if the effective spread exceeds 40%, the daily average quoted spread exceeds 40%, or the effective spread exceeds four times the daily average quoted

spread. For data availability reasons, we use time weighting for the daily quoted spreads (QSPREADPCT_TW_M for MTAQ and QUOTEDSPREAD_PERCENT_TW for DTAQ).

We coalesce DTAQ and MTAQ spreads at the firm-day level, favoring DTAQ when it is available. The daily spreads are then converted to monthly by simple averaging across daily observations.

ISSM daily spreads use our own calculations from the raw ISSM intraday data. We screen the data following Holden and Jacobsen (2014). Quotes are excluded if any of the following hold:

- Time is before 9:00 am or after 4:00 pm
- if mode in (C, D, F, G, I, L, N, P, S, V, X, Z)
- BID>OFR and BID>0 and OFR > 0
- BID>0 and OFR = 0
- OFR-BID>5 and BID>0 and OFR > 0
- OFR $\leq$ 0 or missing
- BID $\leq$ 0 or missing
- OFRSIZE $\leq$ 0 or missing
- BIDSIZE $\leq$ 0 or missing.

NASDAQ listed stocks from 1987 to 1989 and NYSE listed stocks in 1986 are not subject to the size filters as they are all missing ofrsize and bidsize. Trades are kept if all of the following hold

- TIME is after 9:30 am and before 4:00 pm
- PRICE > 0
- TYPE = T
- COND not in (C, L, N, R, O, Z) and SIZE > 0
- From TAQ and correction field is zero

For ISSM, we use a 2-second quote delay. As with the TAQ spreads, we compute dollar-weighted averages within each day, and then simple average across days to covert to monthly.

## Appendix B. Details of Low-Frequency Spreads

Three of our four proxies build off of Roll's (1984) classic microstructure model. The Roll model assumes that the true value of a stock follows a random walk, and that the observed trade prices deviate from the true value by the effective spread. The fourth proxy uses a completely different framework: the Kyle and Obizhaeva (2016) microstructure invariance hypothesis. All 4 proxies have been shown to be highly correlated with HF spreads.

The LF proxies we use are as follows:

- *Hasbrouck's (2009) Gibbs sampler estimate of the Roll model (Gibbs)*

  Hasbrouck (2009) estimates the Roll model using Bayesian methods (Gibbs sampler) and daily closing prices. Identification comes from the "bid–ask bounce" – the phenomenon in which buyer-initiated trades tend to occur at higher prices than seller-initiated trades. Bid–ask bounce induces a negative serial

correlation in transaction prices, which is stronger for stocks that are more expensive to trade. The Bayesian approach ensures that the measured serial correlation is negative, and thus the estimated spread is well defined. Our Gibbs proxy is estimated using annual samples, following the approach recommended by Hasbrouck (2009).

Gibbs forms the basis for transaction costs in several other studies of portfolio returns, including Brandt et al. (2009), Hand and Green (2011), Novy-Marx and Velikov (2016), DeMiguel et al. (2020), and Freyberger et al. (2020).

- *Corwin and Schultz's (2012) High-Low Spread (HL)*

  Corwin and Schultz (2012) estimate the Roll model from daily high and low prices (hence, HL) that are available in CRSP. Identification comes from the fact that the daily high-low ratio reflects both spreads and return volatility, but these two components decay at different rates. Thus, the comparison of 1-day and 2-day price ranges provides information about the effective spread.

  HL is used in many studies including Karnaukh et al. (2015), Koch et al. (2016), McLean and Pontiff (2016), and Chen and Zimmermann (2020).

- *Abdi and Ranaldo's (2017) Close-High-Low (CHL)*

  Abdi and Ranaldo's (2017) CHL proxy estimates the Roll model using daily closing prices as well as the daily high and low (hence, CHL). Abdi and Ranaldo's identification builds off the insight that the average of the daily high and low prices (the midpoint) contains important information about the true price. Abdi and Ranaldo (2017) show that CHL outperforms both Gibbs and HL using a number of empirical tests.

- *Volume-over-Volatility (VoV), based on Kyle and Obizhaeva's (2016) microstructure invariance hypothesis*

  Our last LF proxy takes a rather different approach. Rather than build off of Roll (1984), VoV is based on Kyle and Obizhaeva's (2016) microstructure invariance hypothesis. In particular, we use Fong et al.'s (2017) (FHT's) implementation:

$$[\text{VoV}]_{i,t} = \frac{8.0[\text{STD\_DEV\_OF\_DAILY\_RETURNS}]^{\frac{2}{3}}}{[\text{MEAN\_REAL\_DAILY\_DOLLAR\_VOLUME}]^{\frac{1}{3}}}, \tag{B-1}$$

where $[\text{VoV}]_{i,t}$ is the proxy for effective spread for stock $i$ in month $t$, the $\frac{2}{3}$ and $\frac{1}{3}$ exponents are predictions of Kyle and Obizhaeva's (2016) invariance hypothesis, and the 8.0 coefficient was chosen by FHT to fit the average monthly TAQ effective spread in their U.S. sample. Nominal dollar volume is converted to real dollar volume using the CPI.

The invariance hypothesis is that the distribution of transaction costs is the same across assets and time periods when expressed in terms of "business time," that is, the speed with which "bets" arrive at the market. This hypothesis leads to the prediction that the constant term in trading costs (alternatively, the bid–ask spread) is proportional to the RHS of equation (B-1). Fong et al. (2017) find that VoV is the best-performing LF proxy among many proxies in terms of correlations and RMSE with respect to TAQ spreads.

We set to missing LF spreads if the required CRSP data is commonly missing in the exchange-year (see CRSP documentation). We set to missing Gibbs and VoV spreads for NASDAQ stocks before 1983 because this sample lacks volume data. We set to missing HL and CHL spreads for NASDAQ stocks before 1993 because daily high and low prices are missing for NASDAQ SmallCap Market. Note that in CRSP, the askhi and bidlo fields are not missing but negative for this subsample, indicating that the askhi is an ask rather than a high price. Last, we set all LF spreads to missing for AMEX stocks prior to 1963 because volume, daily high, and daily low are all predominantly missing prior to this date.

HL and CHL, we use the most recent observation of high and low prices for days in which stocks do not trade. As noted in Corwin and Schultz (2012) and Abdi and Ranaldo (2017), on days in which stocks do not trade CRSP provides closing quoted spreads, and closing quoted spreads are very highly correlated with effective HF spreads in the recent sample. In these cases, we do not use the closing quoted spread in order to make interpretation of our LF proxy average simple.

The LF proxies require multiple firm-day observations to compute a spread for a given firm-month. We follow the original papers and do not compute the proxy if the data is insufficient. Specifically, HL requires 12 daily observations, CHL requires 12 eligible days following the definition in Abdi and Ranaldo (2017), VoV requires 5 positive volume and 11 nonzero return observations, and Gibbs requires the sampler to converge.

We compute an LF average if we have at least one LF proxy with data. In 12.24% of observations, all LF and HF spreads are missing data. These missing observations have little effect on our main results, however, as only 0.27% of post-1993 observations are missing, and 90% of our anomalies are published after 1993. If ISSM, TAQ, and the LF spreads are all missing, we match the firm to the nearest firm with available data in terms of Euclidean distance of market equity rank and idiosyncratic volatility rank. If idiosyncratic volatility is missing, we use just the market equity rank. This data-filling procedure follows Novy-Marx and Velikov (2016).

## Appendix C. Details of Cost Optimization

Figures C1 and C2 show that our cost mitigation is effective in-sample. The figures show the distribution of in-sample net returns before (Figure C1) and after (Figure C2) cost mitigation.

Figure C1 shows that net returns before cost mitigation feature a long left tail. While most anomalies have positive net returns ranging between 0 and 60 bps per month, many anomalies have very negative net returns of −50 to −300 bps. Averaging across all anomalies leads to near-zero net return in Table 2.

Anomalies with above-median turnover are shown in bold. These high turnover anomalies occupy the vast majority of the left tail of net returns. These high turnover anomalies include many momentum anomalies, but also include a variety of unrelated anomalies like real estate and order backlog. Persistent anomaly signals like B/M and size are little affected by bid–ask spreads and occupy the right tail of this distribution.

Cost mitigation should be very helpful with this left tail of net returns. Indeed, Figure C2 shows that the left tail is essential gone after cost optimization, and net returns center around 40–50 bps per month.

The best net returns are often reached with equal weighting, however. In Figure C2, these are shown in the nonbold font, while value-weighted anomalies are shown in bold. Thus, in our main results, we examine two types of cost-optimization, one that allows for equal-weighting and another that enforces value-weighting.

FIGURE C1

Distribution of Net Returns: In-Sample, Before Cost Optimization

In Figure C1, we adjust anomaly returns for effective bid–ask spreads (Figure 3). Portfolios are implemented following the original papers as described in Chen and Zimmermann (2022). Anomalies with 2-sided turnover >30% per month are shown in bold. Hash marks indicate larger bins. Academic anomaly strategies are on average unprofitable even in sample, due in part to a long left tail in net returns.

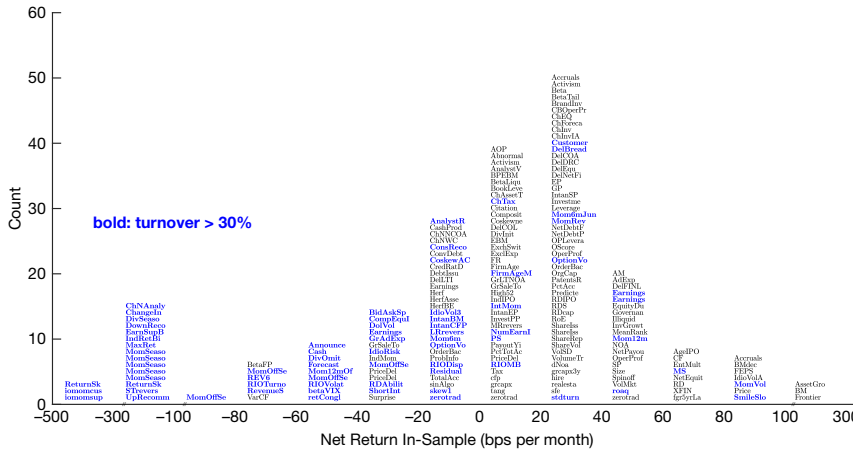



FIGURE C2

Cost Optimization Results: Distribution of Net Returns: In-Sample

For each anomaly in Figure 2, we examine up to 21 cost-mitigated implementations based on restricting trading to liquid stocks, reducing rebalancing, and buy/hold spreads (see Section II.C). We then select the implementation that produces the highest in-sample net return, including the original implementation if the original is best. Bold anomalies are value-weighted. Cost mitigation leads to notably positive net returns in-sample, though this performance sometimes requires equal-weighting.

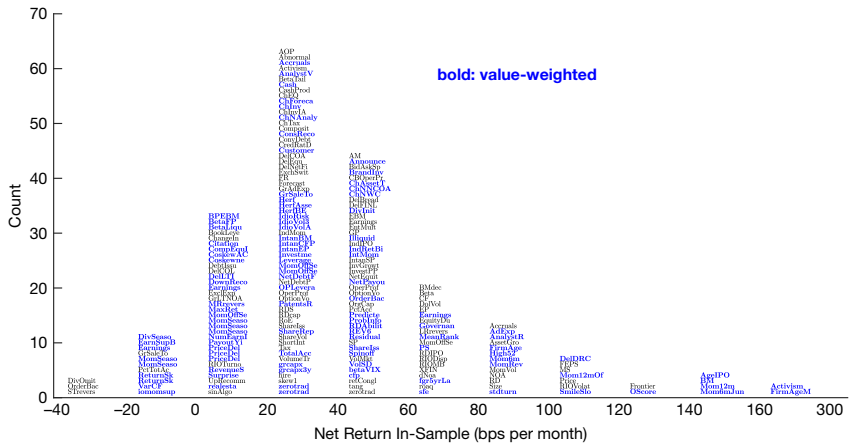
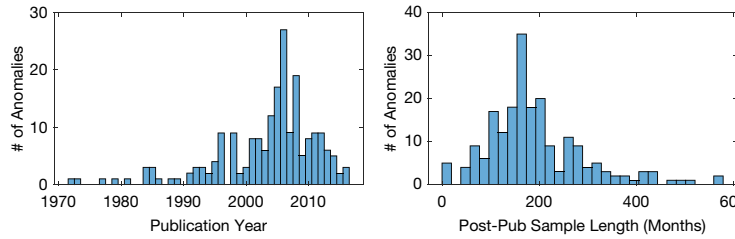


FIGURE C3

Distribution of Publication Years

Figure C3 plots a histogram of the publication years for the 120 anomalies we use and a histogram of the post-publication sample length in months.




# References

Abdi, F., and A. Ranaldo. "A Simple Estimation of Bid-Ask Spreads from Daily Close, High, and Low Prices." *Review of Financial Studies*, 30 (2017), 4437–4480.

An, L.; S. Huang; D. Lou; and J. Shi. "Why Don't Most Mutual Funds Short Sell?" Available at SSRN 3813790 (2021).

Asquith, P.; P. A. Pathak; and J. R. Ritter. "Short Interest, Institutional Ownership, and Stock Returns." *Journal of Financial Economics*, 78 (2005), 243–276.

Azevedo, E. M.; A. Deng; J. L. M. Olea; and E. G. Weyl. "Empirical Bayes Estimation of Treatment Effects with Many A/B Tests: An Overview." *AEA Papers and Proceedings*, 109 (2019), 43–47.

Ball, R.; S. P. Kothari; and J. Shanken. "Problems in Measuring Portfolio Performance: An Application to Contrarian Investment Strategies." *Journal of Financial Economics*, 38 (1995), 79–107.

Barber, B. M.; T. Odean; and N. Zhu. "Do Retail Trades Move Markets?" *Review of Financial Studies*, 22 (2008), 151–186.

Barberis, N. "Psychology-Based Models of Asset Prices and Trading Volume." In *Handbook of Behavioral Economics – Foundations and Applications 1*, Vol. 1, B. D. Bernheim, S. DellaVigna, and D. Laibson, eds. Amsterdam, Netherlands: Elsevier (2018), 79–175.

Bates, J. M., and C. W. J. Granger. "The Combination of Forecasts." *Journal of the Operational Research Society*, 20 (1969), 451–468.

Brandt, M. W.; P. Santa-Clara; and R. Valkanov. "Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns." *Review of Financial Studies*, 22 (2009), 3411–3447.

Briere, M.; C.-A. Lehalle; T. Nefedova; and A. Raboun. "Stock Market Liquidity and the Trading Costs of Asset Pricing Anomalies." Available at SSRN 3380239 (2019).

Campbell, J. Y., and J. H. Cochrane. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *Journal of Political Economy*, 107 (1999), 205–251.

Chen, A. Y. "Most Claimed Statistical Findings in Cross-Sectional Return Predictability are Likely True." Available at SSRN 3912915 (2021).

Chen, A. Y., and T. Zimmermann. "Publication Bias and the Cross-Section of Stock Returns." *Review of Asset Pricing Studies*, 10 (2020), 249–289.

Chen, A. Y., and T. Zimmermann. "Open Source Cross Sectional Asset Pricing." *Critical Finance Review*, 27 (2022), 207–264.

Chordia, T.; A. Subrahmanyam; and Q. Tong. "Have Capital Market Anomalies Attenuated in the Recent Era of High Liquidity and Trading Activity?" *Journal of Accounting and Economics*, 58 (2014), 41–58.

Cochrane, J. H. "Macro-Finance." *Review of Finance*, 21 (2017), 945–985.

Cohen, L.; K. B. Diether; and C. J. Malloy. "Supply and Demand Shifts in the Shorting Market." *Journal of Finance*, 62 (2007), 2061–2096.

Corwin, S. A., and P. Schultz. "A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices." *Journal of Finance*, 67 (2012), 719–760.

Daniel, K., and T. J. Moskowitz. "Momentum Crashes." *Journal of Financial Economics*, 122 (2016), 221–247.

Dawid, A. P. "Selection Paradoxes of Bayesian Inference." In Lecture Notes-Monograph Series (1994), 211–220.

DeMiguel, V.; L. Garlappi; and R. Uppal. "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?" *Review of Financial Studies*, 22 (2009), 1915–1953.

DeMiguel, V.; A. Martin-Utrera; F. J. Nogales; and R. Uppal. "A Transaction-Cost Perspective on the Multitude of Firm Characteristics." *Review of Financial Studies*, 33 (2020), 2180–2222.
Detzel, A. L.; R. Novy-Marx; and M. Velikov. "Model Selection with Transaction Costs." Available at SSRN 3805379 (2021).
Drechsler, I., and Q. F. Drechsler. "The Shorting Premium and Asset Pricing Anomalies" (2016).
Efron, B. "Tweedie's Formula and Selection Bias." *Journal of the American Statistical Association*, 106 (2011), 1602–1614.
Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Vol. 1. Cambridge: Cambridge University Press (2012).
Fama, E. F., and K. R. French. "The Cross-Section of Expected Stock Returns." *Journal of Finance* (1992), 427–465.
Fong, K.; C. Holden; and O. Tobek. "Are Volatility Over Volume Liquidity Proxies Useful For Global Or US Research?" (2017).
Frazzini, A.; R. Israel; and T. J. Moskowitz. "Trading Costs of Asset Pricing Anomalies" (2015).
Freyberger, J.; A. Neuhierl; and M. Weber. "Dissecting Characteristics Nonparametrically." *Review of Financial Studies*, 33 (2020), 2326–2377.
Gârleanu, N., and L. H. Pedersen. "Efficiently Inefficient Markets for Assets and Asset Management." *Journal of Finance*, 73 (2018), 1663–1712.
Green, J.; J. R. M. Hand; and X. F. Zhang. "The Supraview of Return Predictive Signals." *Review of Accounting Studies*, 18 (2013), 692–730.
Green, J.; J. R. M. Hand; and X. F. Zhang. "The Characteristics that Provide Independent Information about Average us Monthly Stock Returns." *The Review of Financial Studies*, 30 (2017), 4389–4436.
Grossman, S. J., and J. E. Stiglitz. "On the Impossibility of Informationally Efficient Markets." *American Economic Review*, 70 (1980), 393–408.
Hand, J. R. M., and J. Green. "The Importance of Accounting Information in Portfolio Optimization." *Journal of Accounting, Auditing & Finance*, 26 (2011), 1–34.
Hanna, J. D., and M. J. Ready. "Profitable Predictability in the Cross Section of Stock Returns." *Journal of Financial Economics*, 78 (2005), 463–505.
Harvey, C. R.; Y. Liu; and H. Zhu. "… and the Cross-Section of Expected Returns." *Review of Financial Studies*, 29 (2016), 5–68.
Hasbrouck, J. "Trading Costs and Returns for US Equities: Estimating Effective Costs from Daily Data." *Journal of Finance*, 64 (2009), 1445–1477.
Holden, C. W., and S. Jacobsen. "Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions." *Journal of Finance*, 69 (2014), 1747–1785.
Hou, K.; S. Kim; and I. M. Werner. "(Priced) Frictions" (2016).
Hou, K.; C. Xue; and L. Zhang. "Replicating Anomalies." *Review of Financial Studies*, 33 (2020), 2019–2133.
Huang, J.-Z., and Z. J. Huang. "Real-Time Profitability of Published Anomalies: An Out-of-Sample Test." *Quarterly Journal of Finance*, 3 (2013), 03n04.
Jacobs, H., and S. Müller. "Anomalies Across the Globe: Once Public, No Longer Existent?" *Journal of Financial Economics*, 135 (2020), 213–230.
Jahan-Parvar, M., and F. Zikes. "When Do Low-Frequency Measures Really Measure Transaction Costs?" (2019).
James, W., and C. Stein. "Estimation with Quadratic Loss." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1 (1961), 361–379.
James, G.; D. Witten; T. Hastie; and R. Tibshirani. *An Introduction to Statistical Learning*, Vol. 112. New York, NY: Springer (2013).
Karnaukh, N.; A. Ranaldo; and P. Soderlind. "Understanding FX Liquidity." *Review of Financial Studies* 28 (2015), 3073–3108.
Kelly, B. T.; S. Pruitt; and Y. Su. "Characteristics Are Covariances: A Unified Model of Risk and Return." *Journal of Financial Economics*, 134 (2019), 501–524.
Knez, P. J., and M. J. Ready. "Estimating the Profits from Trading Strategies." *Review of Financial Studies*, 9 (1996), 1121–1163.
Koch, A.; S. Ruenzi; and L. Starks. "Commonality in Liquidity: A Demand-Side Explanation." *Review of Financial Studies*, 29 (2016), 1943–1974.
Korajczyk, R. A., and R. Sadka. "Are Momentum Profits Robust to Trading Costs?" *Journal of Finance*, 59 (2004), 1039–1082.
Kyle, A. S., and A. A Obizhaeva. "Market Microstructure Invariance: Empirical Hypotheses." *Econometrica*, 84 (2016), 1345–1404.
Lesmond, D. A.; M. J. Schill; and C. Zhou. "The Illusory Nature of Momentum Profits." *Journal of Financial Economics*, 71 (2004), 349–380.

Liu, L.; H. R. Moon; and F. Schorfheide. "Forecasting with Dynamic Panel Data Models." *Econometrica*, 88 (2020), 171–201.
Lo, A. W. "The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective." *Journal of Portfolio Management*, 30 (2004), 15–29.
Marquering, W.; J. Nisser; and T. Valla. "Disappearing Anomalies: A Dynamic Analysis of the Persistence of Anomalies." *Applied Financial Economics*, 16 (2006), 291–302.
McLean, R. D. "Idiosyncratic Risk, Long-Term Reversal, and Momentum." *Journal of Financial and Quantitative Analysis*, 45 (2010), 883–906.
McLean, R. D., and J. Pontiff. "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance*, 71 (2016), 5–32.
Novy-Marx, R., and M. Velikov. "A Taxonomy of Anomalies and Their Trading Costs." *Review of Financial Studies*, 29 (2016), 104–147.
Novy-Marx, R., and M. Velikov. "Comparing Cost-Mitigation Techniques." *Financial Analysts Journal*, 75 (2019), 85–102.
Patton, A. J., and B. M. Weller. "What You See Is Not What You Get: The Costs of Trading Market Anomalies." *Journal of Financial Economics*, 137 (2020), 515–549.
Perold, A. F. "The Implementation Shortfall: Paper Versus Reality." *Journal of Portfolio Management*, 14 (1988), 4.
Pontiff, J., and M. Schill. "Long-Run Seasoned Equity Offering Returns: Data Snooping, Model Misspecification, or Mispricing? A Costly Arbitrage Approach" (2001).
Rapach, D. E.; J. K. Strauss; and G. Zhou. "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy." *Review of Financial Studies*, 23 (2010), 821–862.
Rapach, D. E., and G. Zhou. "Forecasting Stock Returns." In *Handbook of Economic Forecasting*, Vol 2, A Timmerman and G. Elliott, eds. Amsterdam, Netherlands: Elsevier (2013), 328–383.
Roll, R. "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market." *Journal of Finance*, 39 (1984), 1127–1139.
Schultz, P. "Transaction Costs and the Small Firm Effect: A Comment." *Journal of Financial Economics*, 12 (1983), 81–88.
Schwert, G. W. "Anomalies and Market Efficiency." In *Handbook of the Economics of Finance*, Vol. 1, G. Constantinides, M. Harris, and R. Stulz, eds. Amsterdam, Netherlands: Elsevier (2003), 939–974.
Senn, S. "A Note Concerning a Selection "Paradox" of Dawid's." *American Statistician*, 62 (2008), 206–210.
Stigler, S. M. "The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators." *Statistical Science*, 5 (1990), 147–155.
Stoll, H. R. "Market Microstructure." In *Handbook of the Economics of Finance*, Vol. 1, G. M. Constantinides, M. Harris, and R. M. Stulzv, eds. New York, NY: Elsevier (2003), 553–604.
Stoll, H. R., and R. E. Whaley. "Transaction Costs and the Small Firm Effect." *Journal of Financial Economics*, 12 (1983), 57–79.
Timmermann, A. "Forecast Combinations." In *Handbook of Economic Forecasting*, Vol. 1, G. Elliott, C. Granger, and A. Timmermann, eds. New York, NY: Elsevier (2006), 135–196.
Wooldridge, J. M. "Estimation and Inference for Dependent Processes." In *Handbook of Econometrics*, Vol. 4, R. Engle and D. McFadden, eds. New York, NY: Elsevier (1994), 2639–2738.
WRDS: Center for Research in Security Prices. "CRSP/Compustat Merged Database." Wharton Research Data Services, http://www.whartonwrds.com/datasets/crsp/.
Xie, X.; S. C. Kou; and L. D. Brown. "SURE Estimates for a Heteroscedastic Hierarchical Model." *Journal of the American Statistical Association*, 107 (2012), 1465–1479.