The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXVIII, NO. 3 • JUNE 2023

Model Comparison with Transaction Costs

ANDREW DETZEL, ROBERT NOVY-MARX, and MIHAIL VELIKOV*

ABSTRACT

Failing to account for transaction costs materially impacts inferences drawn when evaluating asset pricing models, biasing tests in favor of those employing high-cost factors. Ignoring transaction costs, Hou, Xue, and Zhang (2015, *Review of Financial Studies*, 28, 650–705) *q*-factor model and Barillas and Shanken (2018, *The Journal of Finance*, 73, 715–754) six-factor models have high maximum squared Sharpe ratios and small alphas across 205 anomalies. They do not, however, come close to spanning the *achievable* mean-variance efficient frontier. Accounting for transaction costs, the Fama and French (2015, *Journal of Financial Economics*, 116, 1–22; 2018, *Journal of Financial Economics*, 128, 234–252) five-factor model has a significantly higher squared Sharpe ratio than either of these alternative models, while variations employing cash profitability perform better still.

THE FINANCE LITERATURE DOCUMENTS HUNDREDS of cross-sectional anomalies in stock returns (see, e.g., Harvey, Liu, and Zhu (2016)). In response, researchers have proposed numerous asset pricing models, which has in turn generated multiple studies dedicated to choosing from among these models using progressively more sophisticated statistical techniques.[1]

*Andrew Detzel is at Baylor University Hankamer School of Business and University of Denver. Robert Novy-Marx is at University of Rochester Simon School of Business and NBER. Mihail Velikov is at Pennsylvania State University Smeal College of Business. For helpful comments and suggestions, we thank Stefan Nagel (Editor); two anonymous referees; an anonymous associate editor; Hank Bessembinder; Andrew Chen (discussant); Victor DeMiguel (discussant); Vincent Fardeau (discussant); Wayne Ferson; Andrei Gonçalves (discussant); Michael Halling (discussant); Avi Kamara; Raymond Kan; Lukas Kremens; Dong Lou; Claudia Moise; Alessio Saretto; Stephan Siegel; Yang Song; Fabio Trojani (discussant); Raman Uppal; Chen Xue; and Irina Zviadadze along with conference and seminar participants at the Adam Smith Asset Pricing Workshop, Federal Reserve Bank of Richmond, Midwest Finance Association, Northern Finance Association, Paris December Finance Meetings, Baylor University, University of Oklahoma, and University of Washington. Novy-Marx provides consulting services to Dimensional Fund Advisors, an investment firm headquartered in Austin, Texas with strong ties to the academic community. The thoughts and opinions expressed in this article are those of the authors alone, and no other person or institution has any control over its content. Andrew Detzel and Mihail Velikov have read *The Journal of Finance* disclosure policy and have no conflicts of interest to disclose.

Correspondence: Robert Novy-Marx, Simon School of Business, University of Rochester, Box 270100, Rochester, NY 14627-0100; e-mail: robert.novy-marx@simon.rochester.edu.

[1] See, for example, Fama and French (1993, 2015), Hou, Xue, and Zhang (2015), Barillas and Shanken (2018), Stambaugh and Yuan (2017), Fama and French (2018), Hou et al. (2019), Barillas

DOI: 10.1111/jofi.13225

A model's ability to price assets ultimately reflects how close its factors come to spanning the efficient frontier. If a test asset generates abnormal returns relative to a model, then that asset improves the investment opportunity set. If some combination of the model's factors are ex post mean-variance efficient (MVE), then no other asset can be used to improve performance, and the model prices everything. As a result, factor models are now typically evaluated based on how close their factors come to spanning the ex post efficient frontier, most commonly using the maximum squared Sharpe ratio criterion of Barillas and Shanken (2017).

Unfortunately, this methodology produces misleading results, at least as typically applied, regarding how close a model comes to spanning the efficient frontier. This fact is best illustrated with a simple example. A single factor based on one-month industry-relative reversals, constructed using only stocks with below median volatility, as in Novy-Marx and Velikov (2016), has a squared Sharpe ratio that is more than nine times the maximum squared Sharpe ratio of the Fama and French (1993) three-factor (FF3) model, and almost four times that of the Fama and French (2015) five-factor (FF5) model (4.39 vs. 0.42 and 1.26, respectively). However, the single-factor model based on low-volatility industry-relative reversals (LV-IRR) is not a credible asset pricing model, despite its apparent superior performance to other candidate models under the criterion of ex post efficiency.[2]

The basic problem is how prior studies define the investment opportunity set, without regard to implementation costs. The theoretical underpinning of linear factor models, Ross's (1976) Arbitrage Pricing Theory, is based on the idea that investment opportunities that generate abnormal returns attract arbitrage capital until they are eliminated. These opportunities only attract arbitrage capital in practice, however, if they generate abnormal returns net of costs. While the LV-IRR has enormous gross alpha relative to standard factor models, it is also extremely expensive to trade. As a result, the strategy does not actually represent an attractive trading opportunity; its gross alpha does not indicate that the strategy actually expands the achievable investment frontier.

This paper starts from the premise that arbitrage capital can only be expected to eliminate true abnormal trading opportunities, that is, those that can be exploited profitably net of costs. This is also the view taken by Fama (1991, p. 1575) when arguing that an "[economically] sensible version of the efficiency hypothesis says that prices reflect information to the point where the marginal benefits of acting on information (the profits to be made) do not

et al. (2020), Daniel, Hirshleifer, and Sun (2020), Feng, Giglio, and Xiu (2020), Kozak, Nagel, and Santosh (2020), Lettau and Pelger (2020), and Bryzgalova, Huang, and Julliard (2023).

[2] This strategy's high gross Sharpe ratio also makes it attractive to machine learning algorithms designed to select significant factors. Kozak, Nagel, and Santosh (2020) estimate a stochastic discount factor using 50 anomalies that puts more weight on this strategy than any other. The second- and third-largest weights are assigned to factors based on "industry momentum reversal" (p. 284) and "[ignoring vol] industry-relative reversal," (p. 284) related factors that are also expensive to trade.

exceed the marginal costs." Starting from this premise lowers the bar on what we should expect from an asset pricing model because even the right model may allow for strategies with large gross alphas. There is no reason to expect even the correct performance evaluation model to price strategies that cannot be traded profitably; the best models are those that parsimoniously span the achievable efficient frontier, fully describing the true investment opportunity set as simply as possible.

This paper reevaluates the performance of asset pricing models accounting for transaction costs. Accounting for these costs materially impacts results, reversing several major conclusions drawn by the prior literature. Ignoring transaction costs biases prior studies in favor of models with factors constructed to maximize gross returns, even when this increases trading costs more than gross performance. For example, the MVE portfolios of the factors in the Hou, Xue, and Zhang (2015) four-factor model and the FF5 model take similar positions in market, size, profitability, and investment factors. The size, investment, and profitability factors in the Hou, Xue, and Zhang (2015) model are rebalanced monthly instead of annually, however, and are constructed using a more complicated three-way sorting procedure that increases the weight put on small stocks. Both of these differences contribute to gross factor performance, but increase trading costs even more. As a result, while these differences improve gross performance, they diminish the performance an investor would actually realize. So while the Hou, Xue, and Zhang (2015) model has a maximum squared Sharpe ratio that is over 40% higher than that of the Fama and French (2015) model ignoring transaction costs, the maximum squared Sharpe ratio that the model's factors could actually have delivered in practice was almost 40% lower.

The Fama and French (2015) models also outperform the alternative candidate models pricing a broad cross section of 205 anomalies net of costs. Accounting for costs improves the apparent ability of all models to price this cross section because many of the anomalies with the most impressive gross returns are also expensive to trade, and thus have less anomalous net returns. The impact of diminished anomaly performance is partly offset by the lower net returns earned by the explanatory factors, but this effect is more acute for factors with high turnover, resulting in improved relative performance for models employing low-cost factors like those of Fama and French (2015).

We also find, consistent with Fama and French (2018) and Ball et al. (2016), that replacing the accruals-based operating profitability factor with one based on cash-based profitability significantly improves model performance. The net-of-costs maximum squared Sharpe ratios for variations of the Fama and French (2015) five- and six-factor models that employ a cash-based profitability factor are more than a third higher than those observed on the other candidate models we consider. These variations also span the other models in the sense that an investor already trading these models' factors cannot significantly expand their opportunity set by additionally trading another model's factors. Accounting for transaction costs, the cash-based variations on the Fama and French model also generally perform better than other models at pricing our set of 205 anomalies.

Overall, the evidence in this paper suggests caution when interpreting results from model comparison studies. Arbitrage capital only flows to opportunities that investors can actually exploit, so even the right factor model should, at best, explain returns that compensate for bearing risk, and not returns that merely reflect implementation frictions, no matter how elaborate the statistical techniques used to evaluate model performance. The model comparison landscape fundamentally changes once this insight is incorporated because models that appear to perform strongly ignoring costs often do so only because of high turnover that would be expensive in practice.

More generally, our results highlight serious issues associated with ignoring real-world concerns when doing financial research. Implementation has a first-order impact on the performance realized by investors. As argued by Harvey (2017), strong incentives to find positive results, which are far more likely to get published, may tempt researchers to design experiments that are more likely to "succeed." One of the easiest ways to do this is to ignore frictions. For example, most anomalies look stronger among the smallest stocks. Exploiting these anomalies using schemes that overweight these hard-to-trade firms can dramatically improve a strategy's "paper" performance, but generally does nothing to improve the investor's actual realized performance (see, e.g., Novy-Marx and Velikov (2022)). Ignoring transaction costs and using these strategies to evaluate performance is consequently misleading.

This is not to say that asset pricing results ignoring frictions are uninformative. Such results may, however, provide as much information about the costs incurred trading a model's factors as they do about the risk premia that market participants demand to hold them. In fact, the difference between a model's performance accounting for frictions and its performance ignoring frictions provides a useful metric that helps quantify the extent to which factor predictability can exist within arbitrage bounds. Suppose, for example, that some investors have a behavioral bias that induces autocorrelation in before-cost returns. Transaction costs could allow this correlation to persist in the presence of arbitrageurs, but do not pin down whether this correlation takes the form of momentum or reversal. So, even if the alpha of a strategy is zero after transaction costs, analyzing gross returns to determine whether there is momentum or reversals before transaction cost can provide economically useful information about how assets are priced in this market.

Transaction costs do, however, have a first-order impact on results when evaluating factor model performance, highlighting the importance of treating frictions seriously when doing financial research. While the purpose of our paper is to starkly highlight significant problems with standard performance evaluation tests that ignore transactions costs, our results provide some guidance on how to mitigate these problems, and how to properly test accounting for these costs. Simple linear factor models, which ignore frictions, are useful for quickly estimating betas, and remain our most important diagnostic tool for identifying tilts to common factors. When used for performance evaluation, however, they tend to overstate the true extent to which a test asset expands the achievable investment opportunity set. This is particularly true for strategies with high turnover, or those that hold small and illiquid stocks.

These issues are significantly mitigated for low-turnover value-weighted strategies, and the net performance of these strategies is often superior to that of high-cost strategies based on similar signals. Testing performance using low-cost strategies consequently provides a more accurate idea of the sort of performance an investor might realize in practice, especially when the model used to evaluate performance also employs low-cost factors. Ultimately, however, proper performance evaluation requires explicitly accounting for transaction costs, and statistical techniques that explicitly account for the fact that investors incur transaction cost regardless of whether they buy or sell, so the short version of a factor is not simply just the opposite of the long version. This methodology, which involves both constructing low-cost test strategies and accounting for the asymmetry between buying and selling a factor, is provided in this paper.[3]

Related Literature

Our paper is related to a growing literature that considers the asset pricing implications of transaction costs and implementation frictions. While the focus of our study is comparing asset pricing models, DeMiguel et al. (2020) investigate the role of transaction costs in optimal portfolio construction. They focus on which of many stock-level characteristics are jointly significant for an investor's optimal portfolio. They find that transaction costs increase the number of significant characteristics relative to the case when transaction costs are ignored. This increase results from what they call "trading diversification," the fact that stock purchases based on one characteristic frequently offset sales based on another characteristic, mitigating the costs incurred when trading both characteristics jointly in an "integrated" strategy relative to those incurred trading them separately as "siloed" strategies. When extending our main results to incorporate cost mitigation techniques, we find that exploiting this trading diversification effect yields improved net-of-costs Sharpe ratios for all the models we consider. This performance boost is particularly strong for models that employ high-cost factors, but the gains are not sufficiently large to alter our main conclusions.

Li, DeMiguel, and Martín-Utrera (2020) expand on the motivation we use in this paper, focusing on how incorporating the price impact of large trades alters inferences regarding model comparison. They show that when trading impacts prices, the maximum squared Sharpe ratio criterion for model comparison is problematic because the metric differs, even for a single model, across investors that trade with different levels of capital. Incorporating price impact therefore helps identify the right benchmark model for a given investor. Kan, Wang, and Zheng (2019) also argue that comparing models based on maximum Sharpe ratios is problematic, even ignoring implementation costs, because investors can never achieve the ex post tangency portfolio. They find that several models

[3] The following is a link to the replication code for this paper: https://github.com/velikov-mihail/Detzel-Novy-Marx-Velikov.

that easily outperform the market in terms of in-sample (IS) Sharpe ratios fail to do so based on out-of-sample (OS) Sharpe ratios. While it is not the primary focus of our paper, many of our model comparison inferences are based on comparisons of OS Sharpe ratios, and therefore account for estimation error and transaction costs simultaneously.

Our cost mitigation results also take steps toward closing the gap between how academic studies construct asset pricing factors and how institutions implement versions of these factors in a cost-effective way. Frazzini, Israel, and Moskowitz (2015) argue that when trading anomalies sophisticated institutions incur transaction costs that are significantly lower than popular academic estimates because they design their trading strategies more efficiently. Incorporating cost mitigation strategies allows us to better approximate the strategies professional traders would use based on the academic factors. We find that models with factors based on similar characteristics, even when they have different gross returns, tend to perform similarly net of cost when traded in a manner designed to minimize implementation costs.

The rest of this paper proceeds as follows. Section I describes the problems of ignoring transaction costs when comparing factor models based on their maximum squared Sharpe ratios. Section II introduces the factor models we compare. Section III presents our main results comparing models after accounting for transaction costs. Section IV examines how transaction costs impact comparisons of models based on their performance explaining anomalies. Section V expands on our main comparisons by incorporating cost-mitigation techniques in factor construction and Section VI concludes.

I. Factor Model Comparison

Factor models are often judged by how well they price test assets. Admitting only small abnormal returns relative to the model is the hallmark of success, indicating that the factors come close to spanning the MVE frontier. Conversely, if a test asset has a large abnormal return relative to the model, then an investor trading a model's factors can significantly expand their investment opportunity set by additionally trading the test asset.

Comparing factor models by how well they price test assets is problematic, however, for several reasons. First, the procedure generally lacks a formal statistical criterion for model comparison. Given even a moderately challenging set of test assets, formal statistical tests typically reject all models. Models that are rejected less emphatically are consequently deemed "better" than those that are rejected more emphatically, even when tests do not actually directly compare models. The formal statistical tests used to test each individual model also perversely reward those that explain *less* test asset return variation. Greater residual variation reduces the precision with which the test assets' factor model alphas are estimated, and thus their significance, which makes rejection less likely (or at least less emphatic).

Comparing factor models using test assets is also sensitive to which anomalies are used to test the models, often yielding contradictory answers when

employing different sets of test assets. Using "anomalies" to test asset pricing models also always raises prima facie selection bias concerns, particularly when comparing alternative candidate models to the canonical FF3 model and its variations. Anomalies are usually defined by their high abnormal returns relative to the Fama and French (1993) model, biasing tests that employ anomalies as test assets against this model.

A. Maximum Squared Sharpe Ratio Test

In response to these issues, Barillas and Shanken (2017) introduce a summary statistic of model quality, the maximum squared Sharpe ratio (SR^2). This metric quantifies how close the span of a factor model is to the ex post MVE frontier of *all* assets. Higher SR^2 indicates smaller alphas relative to the model, in a precise sense that is strictly true for the appropriate combination of all test asset alphas.

Gibbons, Ross, and Shanken (1989) quantify the gains (ignoring transaction costs) from adding test assets to a set of factors using the increase in the maximum squared Sharpe ratio. They show that, given a set of tradable factors with excess returns f, then adding test assets with excess returns Π yields an increase in the maximum squared Sharpe ratio of

$$SR^2(\Pi, f) - SR^2(f) = \alpha' \Sigma^{-1} \alpha, \tag{1}$$

where $SR^2(X)$ denotes the maximum possible Sharpe ratio attainable from X, α denotes the vector of intercepts from regressions of the test assets' excess returns on the factor returns, and Σ is the covariance matrix of residuals from these regressions.

While this metric of investment frontier expansion can be sensitive to the choice of test assets (see, e.g., Lewellen, Nagel, and Shanken (2010) and Fama and French (2018)), Barillas and Shanken (2017) note that if Π is the entire universe of excess returns, or at least contains the factors of all relevant models, then $SR^2(\Pi, f) = SR^2(\Pi)$ for any choice of f. In this case, the model with the lowest mispricing is precisely the one with the highest maximum squared Sharpe ratio. They consequently use the maximum squared Sharpe ratio as a model comparison criterion, and this metric has seen wide adoption in the literature (see, e.g., Barillas and Shanken (2018), Fama and French (2018), Ferson, Siegel, and Wang (2019), and Barillas et al. (2020)). We use the criterion here, but, unlike earlier studies, we do so accounting for transaction costs.

B. Issues with Tests Based on Maximum Squared Sharpe Ratio

Figure 1 graphically illustrates the problems that can arise from ignoring transaction costs when using maximum squared Sharpe ratios to compare model performance. The figure shows both the before- and after-costs maximum squared Sharpe ratios of three-factor models: the capital asset pricing model (CAPM), the FF3 model, and a single-factor model based on IRR constructed using only stocks with volatility below the NYSE median (LV-IRR).

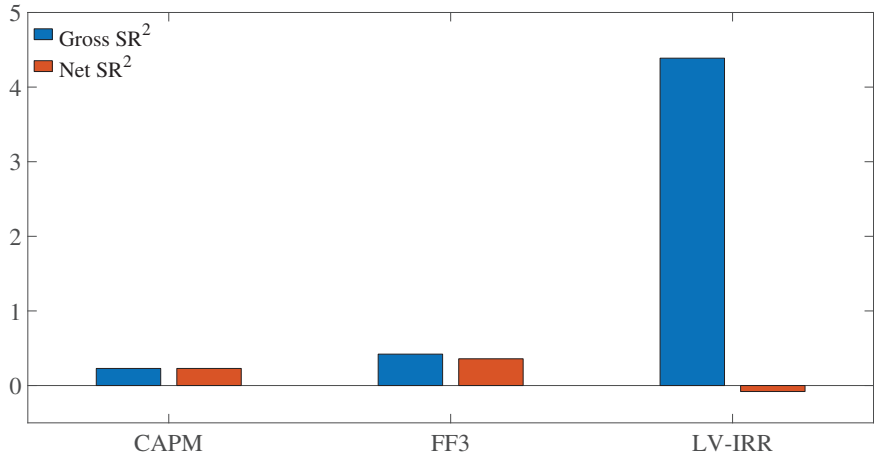

Figure 1. Maximum squared Sharpe ratios of the CAPM, the Fama-French three-factor (FF3) model, and the low-volatility industry-relative-reversal factor (LV-IRR). The blue (left) bars, "Gross SR^2," use factor returns that ignore transaction costs. The red (right) bars, "Net SR^2," use factor returns that account for transaction costs. A negative squared Sharpe ratio indicates that the corresponding Sharpe ratio is negative before squaring. The sample period is January 1972 to December 2021. (Color figure can be viewed at wileyonlinelibrary.com)

The CAPM and FF3 have squared Sharpe ratios, before costs, of 0.23 and 0.42, respectively, while the gross squared Sharpe ratio of LV-IRR is an astounding 4.39. That is, despite the fact that hundreds of papers attempt to price assets using the FF3 model, and none do so using the LV-IRR model, the single-factor model consisting of LV-IRR absolutely dominates the FF3 model under the maximum squared Sharpe ratio criterion—at least ignoring costs. The reader is therefore forced to either accept the implausible conclusion that an LV-IRR factor by itself represents a better asset pricing model than the canonical FF3 model, or to conclude that the current practice of comparing models based on gross squared Sharpe ratios is deeply flawed.

The figure provides a simple resolution for this tension. The LV-IRR factor is so expensive that, after accounting for trading costs, it has a negative Sharpe ratio, while the FF3 achieves a squared Sharpe (0.36) that is more than 50% greater than that of the market factor alone (0.23). Overall, Figure 1 shows that ignoring costs biases comparisons based on squared Sharpe ratios in favor of models with high before-costs performance, even if this performance is unachievable by investors, and thus does not represent a meaningful asset pricing benchmark.

II. Candidate Factor Models

The candidate models used in this paper are taken from Barillas and Shanken (2018) and Fama and French (2018). Barillas and Shanken (2018)

compare the FF5 model, sometimes augmented with a momentum factor (FF6), to the four-factor "q-theory" model of Hou, Xue, and Zhang (2015) (HXZ4). They also construct their own preferred model using the strongest combination of factors from the other models (BS6). We analyze these four models and, following Fama and French (2018), additionally consider versions of the Fama and French five- and six-factor factor models modified by replacing the standard profitability factor, RMW, with a variation based on cash profitability, RMW_C, (FF5_C and FF6_C, respectively). Table I provides a simple summary of the factors employed in all six models. The table includes, for each factor, the stock-level characteristic used for portfolio construction, the basic construction methodology, and the rebalancing frequency. A more detailed description of the factors is provided in Appendix A.1.

Table II provides summary statistics for the factors. It reports each factor's average monthly returns, with t-statistics, both ignoring transaction costs ("gross") and net-of transaction costs ("net"), along with average monthly turnover (TO) and average monthly trading costs (TC).[4] All the factors except for SMB have statistically significant average gross returns over our sample, which begins in 1972, when the earnings announcement dates used in the construction of the HXZ4 ROE factor become widely available. The monthly average excess gross returns on the factors range from 0.14% to 0.62% per month (SMB and MOM, respectively). The turnover and transaction costs tend to be relatively high for factors that rebalance monthly, which also tend to be those with higher average gross returns. The turnover of the factors rebalanced monthly—ME, ROE, IA, HML(m), and MOM—ranges from 20.2% to 52.4% per month. This is an order of magnitude higher than the 3.6% to 10.6% average monthly turnover observed on the annually rebalanced factors, SMB, HML, RMW, CMA, and RMW_C. As a result, the net-of-costs average returns of the former are generally insignificant after costs. This extends prior evidence from Lesmond, Schill, and Zhou (2004) that momentum, while seemingly highly profitable before costs, is hardly profitable after costs.

III. Main Results

A. *Maximum Squared Sharpe Ratios*

In this section, we evaluate the impact of transaction costs on comparisons of asset pricing models based on the maximum squared Sharpe ratio criterion. Suppose a model consists of K factors with vector of gross returns $f_t = (f_{1t}, \ldots, f_{Kt})'$. Following Novy-Marx and Velikov (2016), the net return on the k^{th} factor in period t is given by $f_{kt}^{net} = f_{kt} - TC_t^{f_k}$, where $TC_t^{f_k}$ is the

[4] Transaction costs are calculated as in Novy-Marx and Velikov (2016), using the stock-level effective spread measure of Hasbrouck (2009). This measure captures the substantial cross-sectional heterogeneity in transaction costs and is broadly available over our whole sample. A more detailed description of the methodology is provided in Appendix A.2.

Table I
Factor Models Employed in Tests

For each of the asset pricing models we consider in this paper, this table lists the nonmarket factors used by the model, denoted by an "x" in the column below the model name, and three properties of the factors' construction: the primary characteristics used to form the factor, "Primary sorting characteristic," the frequency that each factor is updated, "Rebalance frequency," and the sorting method used to form each factor, "Portfolio construction." With the exception of SMB, factors with a portfolio construction of 2×3 are based on the six value-weighted portfolios resulting from the intersections of independent sorts of stocks into two groups based on size and three groups based on the primary sorting characteristic. With the exception of ME, factors with portfolio construction of 2×3×3 are based on the 18 value-weighted portfolios obtained from the intersections of independent sorts of stocks into two size groups, three groups based on a secondary characteristic besides size, and three groups based on the primary sorting characteristic. Breakpoints for size are based on the median market capitalization of NYSE stocks at the time of rebalancing and breakpoints for all other characteristics are based on the 30th and 70th percentiles of NYSE stocks. In all models, the factor returns are obtained as the equal-weighted average of the returns on the portfolios with high (or low) values of the primary sorting characteristic minus the equal-weighted average of the portfolios with low (or high) values. SMB returns are given by the simple average of the returns on all portfolios with low size minus the average of the returns on all portfolios with large size in three independent 2×3 sorts of stocks on size and each of the following characteristics: book-to-market ratio, growth in book assets, and operating profitability. ME returns are given by the simple average of the returns on all portfolios with low size minus those with large size in 2×3×3 sorts on size, growth in book assets, and return on equity. All models employ a market factor, MKT, which is the return on the CRSP value-weighted index in excess of the return on the 30-day Treasury bill. FF5 and FF6 denote the Fama and French (2015, 2018) five- and six-factor models, respectively. HXZ4 denotes the Hou, Xue, and Zhang (2015) four-factor q-model. BS6 denotes the Barillas and Shanken (2018) six-factor model. $FF5_C$ and $FF6_C$ denote versions of the FF5 and FF6, respectively, that use cash-based operating profitability instead of accruals operating profitability.

				Models that Employ the Factor					
Factor	Sorting Sgnal	Rebalancing	Construction	FF5	FF6	HXZ4	BS6	$FF5_C$	$FF6_C$
SMB	Market capitalization	annual	2×3	x	x		x	x	x
HML	Book-to-market	annual	2×3	x	x			x	x
RMW	Accruals operating profitability	annual	2×3	x	x				
CMA	Growth in book assets	annual	2×3	x	x			x	x
MOM	Prior year's stock performance, excluding most recent month	monthly	2×3		x		x		x
ME	Market capitalization	monthly	2×3×3			x			
IA	Growth in book assets	monthly	2×3×3			x	x		
ROE	Quarterly returns-on-equity	monthly	2×3×3			x	x		
HML(m)	Book-to-market	monthly	2×3				x		
RMW_C	Cash operating profitability	annual	2×3					x	x

transaction cost of the factor and is defined in Appendix A.2. The corresponding short version of the factor is defined by $f_{kt}^{S,net} = -f_{kt} - TC_t^{f_k}$.[5] Letting f_t^{net} denote the $2K$ vector of both the long and short factor returns, $\mu = E(f^{net})$ and

[5] Transaction costs act as a drag on performance regardless of whether an investor takes a long or short position in a factor. These costs act like first-order penalties on portfolio weights when doing portfolio optimization. Similar to least absolute shrinkage and selection operator techniques, this introduces sparsity in the results. While classical mean-variance portfolio optimization almost

Table II
Factor Summary Statistics

For each candidate asset pricing factor, this table presents average monthly returns and t-statistics, both gross and net of transaction costs, along with average monthly turnover, TO, and transaction costs, TC. MKT, SMB, HML, RMW, and CMA denote the Fama and French (2015) market, size, value, profitability, and investment factors, respectively. MOM denotes the Fama and French (2018) momentum factor. RMW_C denotes the Fama and French (2018) cash profitability factor, which is constructed similarly to RMW but using cash-based, not accruals-based, operating profitability. ME, ROE, and IA denote the Hou, Xue, and Zhang (2015) size, profitability, and investment factors, respectively. HML(m) denotes the monthly updated value factor of Asness and Frazzini (2013). The units for average returns, TO, and TC are % per month. The sample period is January 1972 to December 2021.

	Average Monthly Excess Return (%)					
	Gross	t-Statistic	Net	t-Statistic	TO	TC
MKT	0.63	3.38	0.63	3.38	0.0	0.00
SMB	0.14	1.15	0.11	0.89	3.6	0.03
HML	0.27	2.23	0.22	1.74	6.0	0.06
MOM	0.62	3.48	0.15	0.86	52.4	0.47
RMW	0.30	3.19	0.24	2.56	6.0	0.06
CMA	0.29	3.63	0.19	2.39	10.6	0.09
ME	0.24	1.93	0.08	0.61	20.2	0.17
ROE	0.53	4.85	0.19	1.75	38.2	0.33
IA	0.34	4.33	0.11	1.41	26.2	0.23
HML(m)	0.29	1.96	0.08	0.57	20.9	0.21
RMW_C	0.37	4.84	0.30	3.81	8.1	0.08

$\Sigma = \text{cov}(f^{net})$, the net maximum squared Sharpe ratio is given by

$$\text{SR}^2(f^{net}) = \left(\max_{\theta \in \mathbb{R}^{2K}} \frac{\theta' \mu}{\sqrt{\theta' \Sigma \theta}} \right)^2 \tag{2}$$

subject to the constraints that $\theta_k \geq 0 \ \forall k$ and $\sum_k \theta_k = 1$. Intuitively, one can think of problem (2) as finding the MVE portfolio of funds that replicate the long and short positions of each factor. In the case in which transaction costs are zero, it is straightforward to see that problem (2) is equivalent to finding the maximum Sharpe ratio of the gross factors without nonnegativity constraints on the weights since a positive weight on a short factor becomes economically equivalent to a negative weight on the long factor.

Figure 2 shows the maximum squared Sharpe ratios for the six candidate factor models, both ignoring and accounting for transaction costs. It graphically depicts a key result of our paper that the ranking of model performance ignoring transaction costs differs dramatically from the ranking accounting for transaction costs. The HXZ4 and BS6 are both constructed to get closer to the ex post mean-variance frontier than their predecessors. However, they are

surely yields nonzero positions in all uncorrelated assets, marginal factors are completely excluded from the optimal portfolio after accounting for transaction costs.

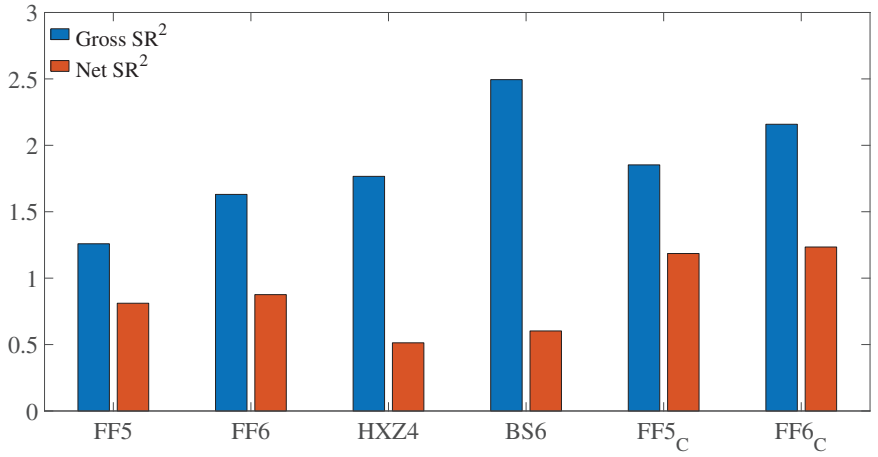

Figure 2. Maximum squared Sharpe ratios of factor models. This figure presents maximum squared Sharpe ratios from the factors in the models listed on the *x*-axis. The blue (left) columns, "Gross SR^2," use factor returns that ignore transaction costs. The red (right) columns, "Net SR^2," use factor returns that account for transaction costs. The sample period is January 1972 to December 2021. (Color figure can be viewed at wileyonlinelibrary.com)

also constructed without any concern for transaction costs. As a result, both of these models have higher gross squared Sharpe ratios than the Fama and French models, despite actually moving an investor farther from the achievable investment opportunity set.

Moving across the dark (blue) bars, Figure 2 shows an upward-sloping pattern. The FF5 model's gross SR^2 is 1.26. Adding a momentum factor (FF6) increases this to 1.63. The HXZ4 model's gross SR^2 is 1.77, and the BS6 model's is 2.49. These squared Sharpe ratios are illusory, however, because they do not actually represent what an investor could have achieved. The achievable investment opportunity set is what is relevant for directing arbitrage capital.

The pattern for the net squared Sharpe ratios, which directly reflect the best risk-reward trade-off that an investor could have achieved in practice, looks very different. While the HXZ4 and BS6 models have higher gross squared Sharpe ratios than the Fama and French models, these models' factors are more expensive to trade, so their spans are actually farther from the achievable efficient frontier. The Fama and French five- and six-factor models' net SR^2 of 0.81 and 0.88 exceed the HXZ4 and BS6 models' 0.51 and 0.60, respectively.[6]

The performance of the variations of the Fama and French model that includes the cash profitability factor are more impressive still. While the gross SR^2 of the $FF5_C$ is 1.85, slightly lower than that of the HXZ4 model, its net SR^2

[6] Accounting for taxes would further diminish the performance of the factors that rebalance monthly relative to those that rebalance annually because short-term capital gains are taxed at a relatively high rate. Similarly, we do not account for short selling fees, which are only available for a short period of time and a limited number of stocks. These fees would also reduce the net Sharpe ratios of all models relative to those in Figure 2.

Table III
Ex Post Mean-Variance Efficient Portfolios

For each of the asset pricing models specified by the row headings, this table presents the weights of each factor in the portfolio consisting of the model's factors that maximize the ex post squared Sharpe ratio, SR^2. Panel A uses factor returns that ignore transaction costs. Panel B uses factor returns that account for transaction costs. The sample period spans January 1972 through December 2021.

Panel A: Results Ignoring Transaction Costs

	Optimal Factor Weight (% of Holdings in the Ex Post MVE Portfolio)											
	MKT	SMB	HML	MOM	RMW	CMA	ME	ROE	IA	HML(m)	RMW_C	SR^2
FF5	18	11	−6		31	47						1.26
FF6	17	9	1	12	25	36						1.63
HXZ4	16						12	31	40			1.77
BS6	13	11		14				29	9	24		2.49
$FF5_C$	16	13	−5			31					45	1.85
$FF6_C$	16	11	0	8		25					39	2.16

Panel B: Results Accounting for Transaction Costs

	Optimal Factor Weight (% of Holdings in the Ex Post MVE Portfolio)											
	MKT	SMB	HML	MOM	RMW	CMA	ME	ROE	IA	HML(m)	RMW_C	SR^2
FF5	21	10	0		32	37						0.81
FF6	20	9	0	7	28	35						0.88
HXZ4	28						5	28	39			0.51
BS6	20	11		9				27	17	16		0.60
$FF5_C$	18	13	0			23					46	1.19
$FF6_C$	18	12	0	4		23					43	1.23

is 1.19, more than twice that of the HXZ4 model. Adding the momentum factor only marginally increases the net SR^2 of the $FF6_C$, to 1.23. The $FF6_C$ model dominates all other models in the sense that adding all of the other models' factors does not expand the investment opportunity set. The net squared Sharpe ratio of the 11-factor model that includes all the factors from all four models is the same 1.23 as that of the $FF6_C$ model alone.

Table III reports portfolio weights in ex post MVE portfolios constructed using factors from the six candidate asset pricing models, that is, the holdings of the portfolios that yield the squared Sharpe ratios shown in Figure 2. Panel A gives the weights an investor would like to have held in each factor, provided they could have rebalanced the factors costlessly. It shows that the HXZ4 model improves on the gross squared Sharpe ratio of the FF5 model essentially by swapping out the Fama and French size, profitability, and investment factors for their own versions of these factors, which rebalance more frequently and are constructed using a more complicated three-way sorting procedure that further overweights small capitalization stocks. Similarly, the BS6 model improves on the gross squared Sharpe ratio of the FF6 model

essentially by swapping out the Fama and French profitability factor for the HXZ4 ROE factor, and shifting the weight on the Fama and French investment factor to a value factor that rebalances monthly, largely in response to stock price performance, yielding a larger negative correlation with momentum. The Fama and French model variations that use cash profitability improve on the standard versions essentially by moving all the of weight on RMW, and a third of the weight on CMA, to RMW_C, which has a higher Sharpe ratio than the accrual-based profitability factor.

Panel B reports results accounting for transaction costs. While the portfolio holdings are similar, the net factor returns are lower by the trading costs they incur, and the squared Sharpe ratios that could actually have been realized are consequently reduced by 36% to 76% (FF5 and BS6, respectively). This reduction is greater for models employing high turnover factors. As a result, the FF5 and FF6 models have higher net squared Sharpe ratios than the HXZ4 and BS6 models (0.81 and 0.88 vs. 0.51 and 0.60), despite having gross squared Sharpe ratios that are significantly lower. The net squared Sharpe ratios of the FF5 model variation that employs cash profitability is much higher, 1.19. For this model, including momentum only yields marginal performance improvements, increasing the net squared Sharpe ratio to 1.23.[7]

B. Comparing SR^2 Using OS Tests

The previous section provides strong evidence that the most popular method for comparing factor model performance, the maximum squared Sharpe ratio tests, gives very different results accounting for frictions. Even the tests accounting for transaction costs are biased, however because the empirical tests compare ex post maximum Sharpe ratios, while the supporting theory concerns ex ante maximum Sharpe ratios. The ex post maximum squared Sharpe ratios are biased strongly upward because the MVE portfolio weights are chosen to maximize the IS Sharpe ratios using information from the full sample. The simplest way to reduce this bias is to choose MVE portfolio weights using only information available at the time of portfolio formation.

Figure 3 shows the time-series performance, net of costs, of the "ex ante optimal" portfolios constructed using the factors from each of the six candidate models. For each candidate factor model, MVE portfolio weights are estimated at the start of each month using only the factors' past performance, that is, information available at the time of portfolio formation. Initial weights are calculated from the first 10 years of data, and our return series consequently starts 10 years later than the data used in most of our other tests. While this procedure significantly mitigates the look-ahead bias present in the IS results, the results are not completely free from all potential biases, as the factors

[7] HML also adds little to the Fama and French model variation based on cash profitability because CMA and HML are positively correlated and thus play similar roles. The four-factor model employing MKT, SMB, CMA, and RMW_C has a squared Sharpe ratio of 1.19, statistically indistinguishable from the 1.19 on the five-factor model that additionally includes HML.

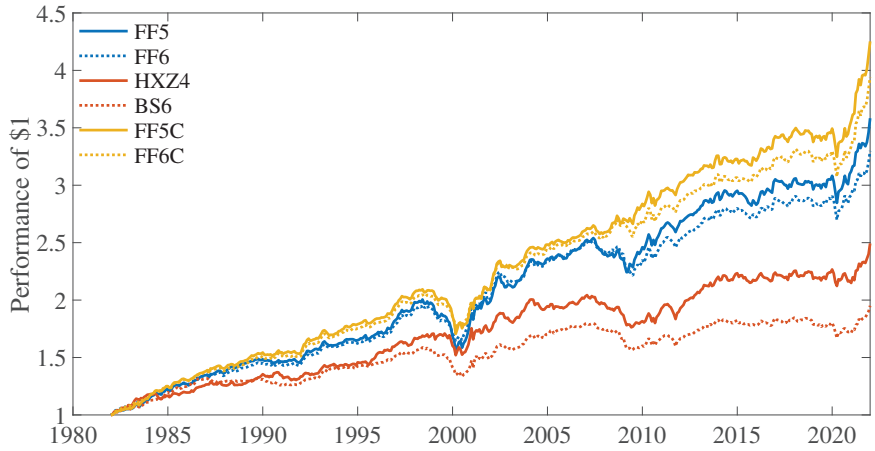

Figure 3. Performance of ex ante optimal portfolios of candidate model factors. The figure shows the performance over time, net of transaction costs, of $1 invested in portfolios constructed using factors from each of the six candidate models (FF5, FF6, HXZ4, BS6, $FF5_C$, and $FF6_C$). Each month, a portfolio holds each factor in proportion to the optimal weights estimated using only returns prior to portfolio formation in expanding windows. The data cover January 1972 through December 2021, but we require a minimum of 10 years to estimate portfolio weights, so the returns begin in January 1982. Appendix B provides details of our calculation of cumulative returns. (Color figure can be viewed at wileyonlinelibrary.com)

under consideration were themselves identified with the benefit of hindsight in studies that use the full-sample data.

Figure 3 shows, for each of the six candidate models, the cumulative performance of these "ex ante optimal" factor portfolios, and yields inferences similar to those in Figure 2, though with some important differences. The five- and six-factor Fama and French models with cash profitability again exhibit the strongest realized performance after accounting for transaction costs (squared Sharpe ratios of 1.03 and 0.99), while the HXZ4 and BS6 have the weakest net performance (squared Sharpe ratios of 0.34 and 0.25). The underperformance of these latter two models is particularly acute here because the HXZ4 IA factor included in both models exhibits by far its strongest relative performance in the first 10 years of our full sample. This yields large initial overweighting on this factor for the HXZ4 and BS6 models, and the factor performs relatively poorly over the remainder of the sample.

Perhaps the most striking difference between these results and the full-sample results presented in Figure 2 is the impact of momentum. Unlike the full-sample results, including a momentum factor is generally detrimental to the performance of these ex ante optimal portfolios. For the full-sample results, the weight on momentum in the ex post MVE portfolios is reduced over the whole sample by the momentum crash of 2009. The OS results, which do not benefit from this look-ahead bias, do not see the crash coming. They consequently overweight momentum relative to the ex post optimal weights when the crash comes, and are thus punished more when it does. As a result,

including a momentum factor in the model slightly reduces overall model performance. The FF5 models have marginally higher squared Sharpe ratios than their six-factor counterparts (0.65 vs. 0.63 for the standard models and 1.03 vs. 0.99 for the ones employing the cash-based profitability factor).

While Figure 3 shows economically large differences in OS performance across models, it does nothing to quantify the statistical significance of these differences. To assess the statistical significance of differences in OS performance, we use bootstraps simulations, following the methodology of Fama and French (2018). These bootstraps first split the 600 months of our sample period, January 1972 to December 2021, into 300 adjacent pairs: months (1, 2), (3, 4),…, (599, 600). Each bootstrap run then randomly selects one month from the pair to be IS, the other to be OS, and then draws 300 random pairs with replacement. For each model, we estimate MVE portfolio weights as the weights on each factor that yield the maximum net-of-costs Sharpe ratio over the 300 IS months, and then compute the Sharpe ratio of the portfolio that puts those weights on each factor over the 300 OS months. It is well known that ex post maximum Sharpe ratios overstate what investors can expect to realize in practice because they overweight (underweight) assets that outperformed (underperformed) in sample. In contrast, OS Sharpe ratios provide a better approximation of the performance an investor might expect to realize going forward.

In the case of nested models like the FF5 and FF6, which differ only by MOM, it is impossible for the nested model to have a higher IS Sharpe ratio than the nesting model, even if the factors missing from the nested model do not improve the mean-variance frontier and therefore should be omitted. For our empirical results, we decide ties in favor of parsimony. That is, when two nested models have the same Sharpe ratio in a given run, we break the tie in favor of the model constructed using fewer factors. For example, when MOM is not used in the optimal IS portfolio for FF6 in a given simulation run, the FF6 Sharpe ratio must equal that of the FF5, and in such cases we declare the latter model the "winner." Results reported in our tables are based off 100,000 runs.

Table IV presents the bootstrap comparison of the differences between the net-of-costs Sharpe ratios of each model. Panel A presents IS results. These bootstrap IS SR^2 are biased upward, even more than in the full-sample results because the overfitting is more acute in the shorter sample, so their level should be interpreted with caution. With that caveat, the average IS net SR^2 is higher for the FF5 and FF6 models than for the HXZ4 and BS6 models. Moreover, the FF6 model outperforms the HXZ4 in over 97% of the samples, and the BS6 model in over 87% of the samples. The $FF5_C$ and $FF6_C$ perform better still, with one of these two models having the greatest Sharpe ratio of all models in 95% of runs.

Panel B presents OS results. The reported squared Sharpe ratios, which are not as inflated by IS overfitting, are strikingly lower, generally little more than half the corresponding IS estimates. These average OS squared Sharpe ratios are remarkably consistent with the performance of the ex ante optimal portfolios of the candidate factor models shown in Figure 3. For the OS bootstraps,

Table IV

Bootstrapped Net-of-Costs Maximum Squared Sharpe Ratios

For each model we consider, the first column presents average net-of-costs maximum squared Sharpe ratios, SR^2, from 100,000 in-sample (IS) or out-of-sample (OS) simulation runs. IS and OS simulations split the 600 sample months of our sample period, January 1972 through December 2021, into 300 adjacent pairs: months (1, 2), (3, 4), … (599, 600). A simulation run draws a random sample with replacement of 300 pairs. The IS simulation run chooses a month randomly from each pair in the run, reusing the same month if the pair is drawn more than once. We calculate IS SR^2 for all models on that sample of months using equation (2) and then apply the corresponding portfolio weights in the unused months of the simulation pairs to produce the corresponding OS estimate of the Sharpe ratio for the IS tangency portfolio. The six columns labeled by model names, "FF5" to "$FF6_C$," present the percentage of bootstrap simulations in which the squared Sharpe ratio of the model defined by the row heading is greater than that of the model defined by the column heading. The last column, "Best," presents the percentage of bootstrap simulation runs in which the model specified by the row heading has the highest squared Sharpe ratio among all models in the run. Panel A presents IS results and Panel B presents OS results.

Panel A: In-Sample Bootstrap Results

		Probability (%) that the Row Model Performs Better than the Column Model						
	Mean-SR^2	FF5	FF6	HXZ4	BS6	$FF5_C$	$FF6_C$	Best
CAPM	0.30	0.0	0.0	0.1	0.0	0.0	0.0	0.0
FF5	1.13		16.0	90.3	68.3	2.5	1.4	0.5
FF6	1.29	84.0		97.5	87.6	19.8	3.2	2.4
HXZ4	0.77	9.7	2.5		0.4	1.9	0.4	0.0
BS6	1.00	31.7	12.4	99.6		9.2	2.3	2.1
$FF5_C$	1.53	97.5	80.2	98.1	90.8		20.6	19.6
$FF6_C$	1.67	98.6	96.8	99.6	97.7	79.4		75.3

Panel B: Out-of-Sample Bootstrap Results

		Probability (%) that the Row Model Performs Better than the Column Model						
	Mean-SR^2	FF5	FF6	HXZ4	BS6	$FF5_C$	$FF6_C$	Best
CAPM	0.30	17.7	18.1	31.9	33.9	6.7	7.5	5.3
FF5	0.60		46.1	82.1	81.3	2.8	7.7	1.2
FF6	0.61	53.9		83.8	87.5	7.3	3.1	1.5
HXZ4	0.39	17.9	16.2		43.0	4.1	4.1	1.9
BS6	0.39	18.7	12.5	57.0		3.6	2.2	1.1
$FF5_C$	0.94	97.1	92.7	95.9	96.4		54.2	48.6
$FF6_C$	0.93	92.3	96.9	95.9	97.8	45.8		40.4

these average squared Sharpe ratios are 0.60, 0.61, 0.39, 0.39, 0.94, and 0.93 for the six candidate models, while the squared Sharpe ratios realized in Figure 3 are 0.65, 0.63, 0.34, 0.25, 1.03, and 0.99, respectively.[8] Also consistent

[8] The HXZ4 and BS6 models exhibit somewhat stronger performance in the bootstraps than in the ex ante portfolio tests. The bootstraps include the first 10 years of the sample, which are

with Figure 3, including a momentum factor in the OS bootstraps does not yield significant gains. The average OS net-of-costs SR^2 of the FF5 and FF6 models are almost identical, and both significantly higher than those of the HXZ4 and BS6 models. The net-of-costs OS SR^2 of the FF5 and FF6 are greater than those of the HXZ4 and BS6, respectively, in roughly 83% of the bootstrap samples. Overall, the evidence in Figure 2 and Tables III and IV shows that the FF5 and FF6 models tend to dominate the HXZ4 and BS6 on a net basis because of the latter models' high transaction costs. This finding essentially reverses the comparisons of Hou et al. (2019) and Barillas and Shanken (2018).

The Fama and French model variations that employ cash-based profitability have stronger OS performance still, with average squared Sharpe ratios of nearly one, more than 50% higher than their counterparts employing the accruals-based profitability factor. In fact, in 89% of the runs, the $FF5_C$ or $FF6_C$ model is the best-performing model. The single-factor market model (CAPM) is the best-performing model in nearly half of the remaining 11% of runs, while none of the other four models perform best even 2% of the time. The performance of the two models employing cash profitability is also surprisingly similar. The $FF5_C$ has a slightly higher average OS squared Sharpe ratio than the six-factor version that also includes momentum, and with ties breaking in favor of the more parsimonious model, the $FF5_C$ outperforms the $FF6_C$ more often than it underperforms it.

IV. Model Performance Explaining Anomalies

This section takes an alternative throwback approach to factor model comparison. Instead of studying the ability of the factor models to price each other's factors, this alternative approach compares models' ability to price various "test assets." For these test assets, we use the 205 anomaly long/short portfolios from Chen and Zimmermann (2022).[9]

For each model M taken from the six candidates, and for each test asset A taken from the 205 anomalies, we compute the maximum squared Sharpe ratios attainable from the model's factors alone, $SR^2(M)$, and from the model's factors augmented with the anomaly, $SR^2(M, A)$. We then compute the extent to which adding the anomaly expands the model's ex post mean-variance frontier, $\%\Delta SR^2(M, A) = SR^2(M, A)/SR^2(M) - 1$. We do this using both gross- and net-of-costs returns. In each case, we then rank the anomalies for each model based on this measure.

used to estimate the initial portfolio weights for the ex ante optimal portfolios, and over which the HXZ4 IA factor included in both models had by far its strongest performance.

[9] Chen and Zimmermann (2022) is an open source asset pricing project. This project shares code and data to reproduce these 205 cross-sectional predictors. We refer to all of these predictors as anomalies for simplicity. We use the March 2021 data release, downloaded from www.openassetpricing.com. We follow Chen and Velikov (2021) and use the original paper anomaly construction, as defined in the anomaly summary file provided by Chen and Zimmermann (2022). The anomalies are constructed from the usual data sources. More than half of the predictors focus on Compustat data, and about 30% use purely price data. Most of the remainder use analyst forecasts, though several focus on institutional ownership data, trading volume, or specialized data.

Figure 4 plots the extent to which the anomalies improve each model's squared Sharpe ratio. Anomalies that improve a model's maximum SR^2 the least are shown on the left, while those that yield the greatest improvements are shown to the right. If one model's curve is below another, then across the distribution, the anomalies expand the frontier of the former model less than they do the second, indicating that the former better prices anomalies.

The left side of the figure (Panel A), which shows results ignoring transaction costs, is consistent with the work of Hou, Xue, and Zhang (2015). It shows that the anomalies generally expand the frontier for the FF5 model more than they do for the HXZ4 or BS6 models, and that these latter models perform similarly to the Fama and French six-factor model.

The right side of the figure (Panel B), which shows results net of transaction costs, looks very different. First, more than half the anomalies contribute nothing to frontier expansion, even for the worst-performing model. That is, only half of anomalies seem at all anomalous after accounting for the cost of trading them. In the right half of the distribution, where the anomalies do expand the investment frontier, the lines for the Fama and French five- and six-factor models are always below those for the HXZ4 and BS6 models, suggesting that the latter two models, presented as improvements on the standard models, actually are worse at explaining achievable investment performance. This is particularly remarkable, given that our test assets are selected in large part on the basis of their anomalous performance relative to the Fama and French models, and consequently biased against these models. Perhaps most strikingly, the lines corresponding to the Fama and French model variations that employ cash profitability are well below those for the other models, admitting much smaller frontier expansions across the whole distribution. For each pair of models, {HXZ4, BS6}, {FF5, FF6}, and {$FF5_C$, $FF6_C$}, the line for the model that includes momentum is nearly indistinguishable from the line for the model that omits momentum, suggesting that momentum has little role in expanding the achievable investment frontier.

In the Internet Appendix, we replicate this exercise calculating the frontier expansion for each anomaly using only data that follow its publication, which helps control for any data snooping biases (Figure IA.1).[10] With the caveat that the anomalies have a median sample period of 15 post-publication years, the difference in performance across all models becomes small before costs, consistent with the finding that many anomalies and asset pricing factors experienced a reduction in returns in the post-2000 sample (see, e.g., Chordia, Subrahmanyam, and Tong (2014), Ball et al. (2016), and Chen and Velikov (2021)). Consistent with Figure 4, however, transaction costs still have a significant impact, with anomalies expanding the achievable frontier the most for the higher turnover, higher cost HXZ4, and BS6 models.

[10] The Internet Appendix may be found in the online version of this article.

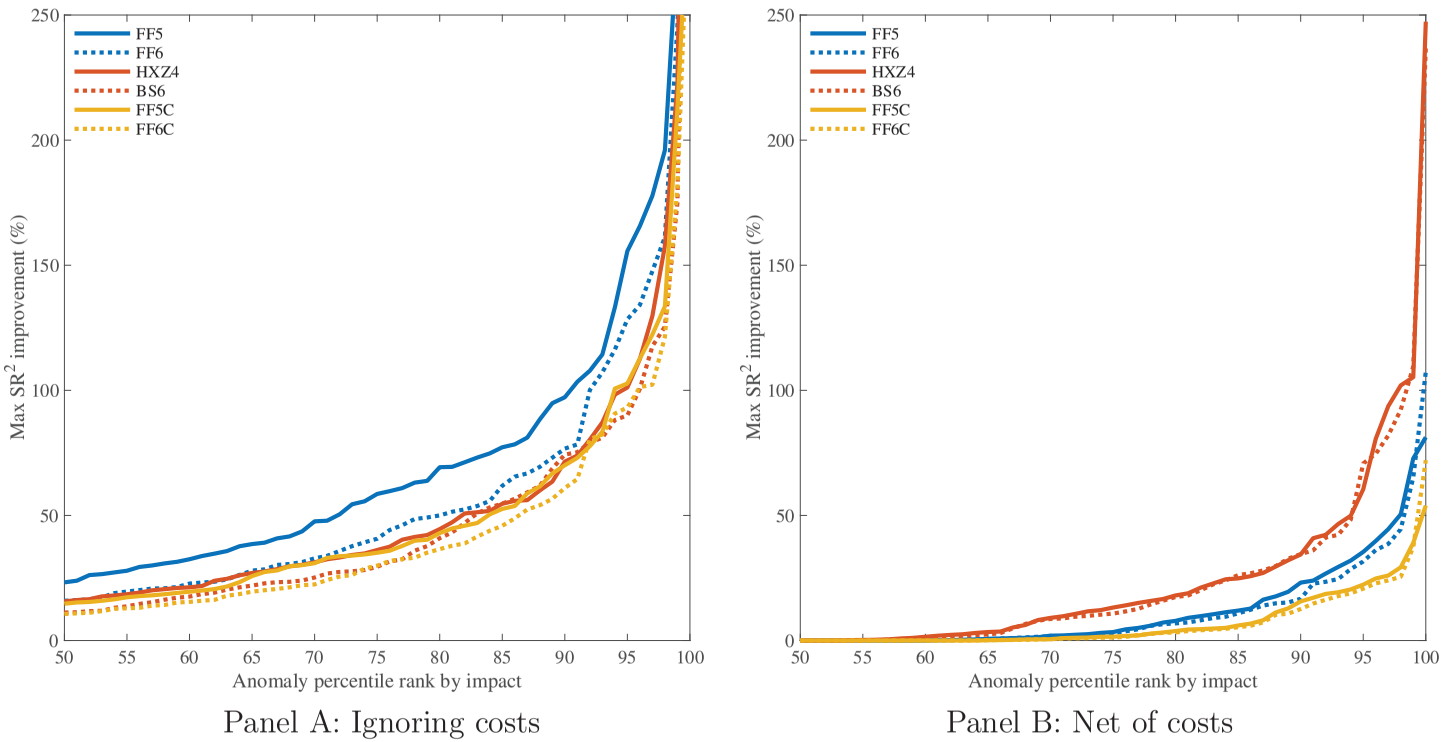

Figure 4. Frontier expansion (squared Sharpe ratio improvement) from adding anomalies to asset pricing models. For each asset pricing model we consider, M, and each of the 205 anomalies, A, described in Section IV, we compute the maximum ex post squared Sharpe ratio attainable from the model's factors, $\text{SR}^2(M)$, along with the maximum squared Sharpe ratio attainable from the model's factors and the anomaly, $\text{SR}^2(M, A)$. We then compute the squared Sharpe ratio improvement, $\%\Delta\text{SR}^2(M, A) = \text{SR}^2(M, A)/\text{SR}^2(M) - 1$. This figure depicts plots of the percentiles from the distribution of the 120 $\%\Delta\text{SR}^2(M, A)$ for each model specified by the plot legend. Panel A (B) ignores (accounts for) transaction costs. For ease of visualization, the y-axis is truncated at 250%. (Color figure can be viewed at wileyonlinelibrary.com)

Table V
Summary Statistics of Cost-Mitigated Factors

This table presents average monthly returns and t-statistics, both gross and net of transaction costs, along with average monthly turnover, TO, and transaction costs, TC, of versions of the asset pricing factors defined in Table II that employ the 20% banding cost mitigation strategy described in Section V. The units for average returns, TO, and TC are percent per month. The sample period is January 1972 to December 2021.

	Average Monthly Excess Return (%)				%/mo.	
	Gross	t-Statistic	Net	t-Statistic	TO	TC
MOM	0.72	3.97	0.43	2.40	29.4	0.28
ME	0.21	1.73	0.07	0.55	16.2	0.14
ROE	0.57	4.89	0.31	2.67	27.2	0.26
IA	0.39	4.64	0.20	2.31	20.8	0.19
HML(m)	0.25	1.70	0.15	1.02	9.0	0.10

V. Cost Mitigation

The previous results clearly show that transaction costs materially impact the performance of prominent asset pricing models in the literature. However, these models' factors were designed largely without considering costs. Institutional investors dedicate entire departments to executing trades in a cost-effective way and can implement alternative versions of the strategies on which asset pricing factors are based less expensively than those estimated in Table II (see, e.g., Frazzini, Israel, and Moskowitz (2015), Novy-Marx and Velikov (2019), and DeMiguel et al. (2020)). In this section, we compare the performance of models when factors are constructed incorporating two prominent cost mitigation techniques, "banding" and "netting."

Intuitively, cost mitigation should reduce unnecessary trading while maintaining exposure to a given strategy's underlying signal. Banding seeks to accomplish this by employing a stricter criterion to enter than to exit a position. Even a small buy/hold spread significantly reduces turnover without significantly reducing the quality of the underlying signal.[11] Novy-Marx and Velikov (2016) find banding to be the single most effective cost mitigation method for individual factors, generally yielding significant net performance gains for strategies that rebalance at monthly or higher frequencies.

Table V gives summary statistics for factors constructed using the banding technique, which we employ for factors that rebalance monthly, including average monthly returns, turnover, and transaction costs. Banding dramatically

[11] The academic factors typically hold/short stocks in the top/bottom 30% of NYSE stocks ranked on the primary sorting variable. Our banded versions introduce a 20% spread centered at these cutoffs, so do not establish long/short positions in stocks until they enter the top/bottom 20% of the primary sorting variable, but maintain established positions until the stocks fall out of the top/bottom 40%, with both thresholds calculated using NYSE breaks. For market capitalization, which typically uses the NYSE median as a breakpoint, we buy (short) stocks entering the top (bottom) 40% and continue to hold current positions until they leave the top (bottom) 60%.

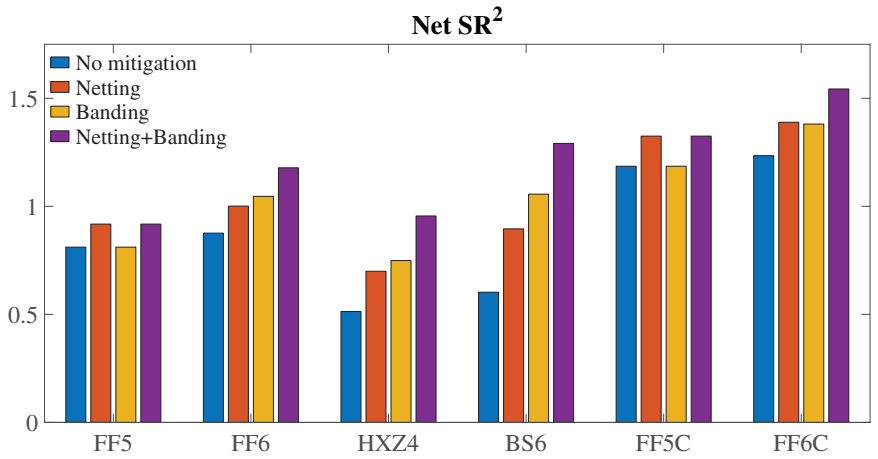

Figure 5. Net-of-costs maximum squared Sharpe ratios with cost mitigation.This figure presents maximum squared Sharpe ratios, accounting for trading costs, for versions of the six candidate factor models constructed to be transactionally efficient. The blue columns, which serve as a baseline, use the standard academic factors, which are designed without any specific concerns for minimizing transaction costs. The red columns assume that traders can realize the full benefits of trading diversification. The yellow columns are based on factors that use the banding strategy with a 20% buy/hold spread described in Section V. The purple columns employ both cost mitigation strategies. The sample period is January 1972 to December 2021. (Color figure can be viewed at wileyonlinelibrary.com)

improves these factors' net performance. It more than doubles the average net returns of MOM (from 0.15 with $t = 0.86$ to 0.43 with $t = 2.40$). It also increases the net returns of ROE and IA by about 10 bps per month and renders these returns significant at the 5% level, while the corresponding returns of the unmitigated factors are insignificant.

Investors trading multiple factors can also reduce turnover by netting trades across strategies, that is, by letting purchases based on one factor offset sales based on another. As a result, trading individual stocks based on the portfolio weights of multiple factors (the "integrated strategy") is generally less expensive than the sum of the costs associated with investing in each factor in isolation (the "siloed strategy").[12] For example, Frazzini, Israel, and Moskowitz (2015) observe that value and momentum trades tend to offset each other and institutional investors trading both strategies use netting to take advantage of the resulting cost savings. DeMiguel et al. (2020) refer to the cost reduction of netting as "trading diversification" and show that it significantly increases the number of characteristic-based factors that contribute to the optimal portfolio.

Figure 5 shows the impact of employing cost mitigation techniques on asset pricing model performance. It shows each candidate model's maximum net-of-costs squared Sharpe ratios without cost mitigation, employing banding and netting each alone, and employing the two mitigation techniques together.

[12] Formulas for transaction costs when using netting can be found in Appendix A.2.

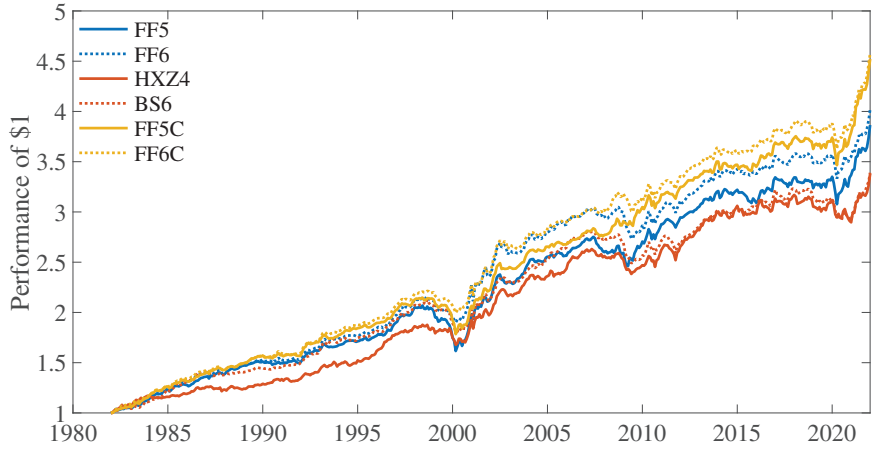

Figure 6. Performance of transactionally efficient ex ante optimal portfolios. The figure shows the performance over time, net of transaction costs, of portfolios constructed from factors, using both netting and banding, for each of the six candidate models (FF5, FF6, HXZ4, BS6, $FF5_C$, and $FF6_C$). Portfolios are constructed each month using only returns prior to portfolio formation in expanding windows. The data cover January 1972 through December 2021, but we require a minimum of 10 years to estimate portfolio weights, so the returns begin in January 1982. (Color figure can be viewed at wileyonlinelibrary.com)

Trading cost mitigation increases the maximum squared Sharpe ratios for all the models, but these gains are most dramatic for the models with factors that rebalance monthly. For these models, the combined effect of both techniques roughly doubles the squared Sharpe ratios of the HXZ4 and BS6 models. The squared Sharpe ratios of the cost-mitigated HXZ4 and BS6 models are even slightly higher than those of the FF5 and FF6, respectively. Despite these dramatic improvements, however, no model comes close to realizing the before-costs squared Sharpe ratios of the corresponding "unmitigated" models from Table III, which are as much as 114% higher (BS6). The relatively large performance improvement for the high-turnover factors also results in less performance difference across models with factors constructed using similar sorting variables. This result suggests that, while the academic literature dedicates many studies to comparing asset pricing models, sophisticated investors trading efficiently perceive little difference between different models based on the same underlying economic criteria.

A. OS Model Performance Comparison with Cost Mitigation

The Sharpe ratios presented in Figure 5 are maximized over the full sample and are thus also strongly biased upward. The simplest way to attempt to correct for this bias is to again choose MVE portfolio weights using only information available at the time of portfolio formation. Figure 6 shows the time-series performance, net of costs, of the "ex ante optimal" portfolios

constructed using both netting and banding for each of the six candidate models, similar to Figure 3, for the models that employ the simply constructed academic versions of the factors.[13] For each candidate factor model, MVE portfolio weights are estimated at the start of each month using only information available at the time of portfolio formation. Initial weights are calculated from the first 10 years of data, and our return series consequently starts 10 years later.

Figure 6 yields similar inferences to Figure 5, again with some important differences. The five- and six-factor Fama and French models with cash profitability again exhibit the strongest realized performance after accounting for transaction costs (squared Sharpe ratios of 1.17 and 1.25), and momentum generally yields small but significant performance enhancements when traded in a transactionally efficient way. Unlike the IS results, however, the FF5 and FF6 significantly outperform the HXZ4 and BS6 OS. While the FF5 and FF6 have slightly lower IS maximum squared Sharpe ratios than the HXZ4 and BS6, respectively, the four models' realized ex ante optimal squared Sharpe ratios are 0.74, 0.86, 0.63, and 0.72, respectively. Again, the exceptional performance of the HXZ4 investment factor IA over the first 10 years of the sample yields ex ante portfolios of the HXZ4 and BS6 factors initially strongly tilted to IA, and this factor's relatively poor subsequent performance negatively impacts portfolio performance. Despite this, there is much less dispersion in model performance here than is observed in Figure 3, which compares net performance of the ex ante optimal model portfolios before cost mitigation. That is, consistent with Figure 5, designing factors to be transactionally efficient yields smaller performance difference across models with factors constructed using similar sorting variables.

Table VI presents bootstrap Sharpe ratio comparisons, analogous to those provided in Table IV, using the factors constructed using both netting and banding. Panel A tabulates the maximum net-of-costs squared Sharpe ratios and the gains from each cost mitigation technique. Panel B presents the IS bootstrap results, while Panel C provides OS results. Overall, these results are generally consistent with the takeaways from Figures 5 and 6. After cost mitigation, the average OS Sharpe ratios of the HXZ4 and BS6 rise dramatically from the 0.39 and 0.39 reported in Table IV to 0.78 and 1.00, respectively. These figures also now exceed those observed on the FF5 and FF6 models, 0.70 and 0.87, respectively. The cost-mitigated HXZ4 and BS6 also have higher OS Sharpe ratios than the FF5 and FF6 in 58% and 68% of bootstrap runs, respectively, while the corresponding proportion for these models' unmitigated counterparts are less than 20% in Table IV. While trading the momentum factor designed without regard for reducing transaction costs diminishes overall OS portfolio performance almost half the time in Table IV, after cost mitigation, all three models that include momentum (FF6, BS6, and $FF6_C$) beat their closest counterparts that exclude it (FF5, HXZ4, and $FF5_C$) out of sample in

[13] The Internet Appendix provides a similar analysis using the two mitigation techniques individually (Figures IA.2 and IA.3) as well as Ledoit and Wolf (2008) heteroskedasticity- and autocorrelation-robust tests for statistical significance of the differences in the Sharpe ratios between different models' ex ante optimal portfolios (Table IA.V).

Table VI

Bootstrapped Maximum Squared Sharpe Ratios of Models Employing Cost Mitigation Techniques

Panel A presents net-of-costs full-sample maximum squared Sharpe ratios for versions of the six candidate models constructed using two transaction cost mitigation techniques, netting and banding, and the gains from using each technique individually and together. The first column of Panels B and C presents average net-of-costs maximum squared Sharpe ratios, SR^2, from 100,000 IS or OS simulation runs. IS and OS simulations split the 600 months of our sample period, January 1972 to December 2021, into 300 adjacent pairs: months (1, 2), (3, 4), … (599, 600). A simulation run draws a random sample with replacement of 300 pairs. The IS simulation run chooses a month randomly from each pair in the run. We calculate IS SR^2 on that sample of months and then apply the corresponding portfolio weights in the unused months of the simulation pairs to produce the corresponding OS estimate of the Sharpe ratio for the IS tangency portfolio. The six columns in Panels B and C labeled with model names present the percentage of bootstrap simulations in which the squared Sharpe ratio of the model defined by the row heading is greater than or equal to that of the model defined by the column heading. The last column of Panels B and C, "Max," presents the percentage of bootstrap simulation runs in which the model specified by the row heading has the highest squared Sharpe ratio of all models in the run. Panel B presents IS results and Panel C presents OS results.

Panel A: Full-Sample Maximum-Squared Sharpe Ratios and Gains Relative to No Mitigation

	FF5	FF6	HXZ4	BS6	$FF5_C$	$FF6_C$
SR^2 (Netting + Banding)	0.92	1.18	0.96	1.29	1.32	1.54
Gain from Netting + Banding	13%	35%	86%	114%	12%	25%
Gain from Netting	13%	14%	36%	49%	12%	13%
Gain from Banding	NA	19%	46%	75%	NA	12%

Panel B: In-Sample Bootstrap Results

		Probability (%) that the Row Model Performs Better than the Column Model						
	Mean-SR^2	FF5	FF6	HXZ4	BS6	$FF5_C$	$FF6_C$	Max
CAPM	0.30	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FF5	1.26		3.6	53.4	15.1	2.6	0.8	0.1
FF6	1.64	96.4		85.1	37.2	40.4	3.9	2.0
HXZ4	1.24	46.6	14.9		0.5	16.3	3.5	0.1
BS6	1.74	84.9	62.8	99.5		52.2	24.5	24.0
$FF5_C$	1.69	97.4	59.6	83.7	47.8		5.1	4.2
$FF6_C$	2.03	99.2	96.1	96.5	75.5	94.9		69.5

Panel C: Out-of-Sample Bootstrap Results

		Probability (%) that the Row Model Performs Better than the Column Model						
	Mean-SR^2	FF5	FF6	HXZ4	BS6	$FF5_C$	$FF6_C$	Max
CAPM	0.30	13.2	9.0	10.7	6.8	4.7	3.8	2.1
FF5	0.70		19.9	40.3	20.1	2.6	5.0	0.4
FF6	0.87	80.1		61.5	32.4	25.9	4.1	1.4
HXZ4	0.78	59.7	38.5		17.6	23.5	14.2	5.4
BS6	1.00	79.9	67.6	82.4		43.5	27.1	21.1
$FF5_C$	1.07	97.4	74.1	76.5	56.5		23.8	19.4
$FF6_C$	1.20	95.0	95.9	85.8	72.9	76.2		50.1

more than three quarters of the bootstrap runs. While the $FF6_C$ is still the best-performing model 70% of the time in sample, and 50% out of sample, the BS6 model is a stronger contender than its unmitigated counterpart in Table IV, performing best in sample 24% of the time and out of sample about 21% of the time.[14]

Overall, the results in Figure 5 and Table VI show that cost mitigation techniques can substantially improve the performance of asset pricing models after accounting for costs. Trading efficiently also closes the performance gaps between models with similar factors. Considering the large impact of cost mitigation techniques, researchers should consider incorporating these techniques in future generations of models, especially when factors rebalance monthly or more frequently.[15]

VI. Conclusion

In a world with implementation costs, even the "right" risk-based asset pricing model should not completely explain the cross section of before-cost expected returns, but only the portion that arises as compensation for risk. We investigate which asset pricing models explain the cross section of returns taking transaction costs into account. Our results show that correcting for costs fundamentally alters the outcome of model comparison exercises. Prior model-selection studies tend to pick factors that update at a monthly frequency because they account for the benefits of frequent updating while ignoring the costs incurred when doing so. This gives these models an unrealizable and illusory advantage over models with factors that update only annually. After costs are considered, models with less frequent rebalancing tend to outperform.

More generally, these results highlight serious problems associated with ignoring real-world concerns when researching financial markets. Accounting for frictions can completely reverse conclusions obtained when ignoring them. The performance that investors realize is often far inferior to that promised by active fund managers, who support their claims using back tests that fail to adequately account for slippage. Implementation issues are not just an annoyance; they have first-order impacts on outcomes. Compounding these issues are strong incentives to find positive results, in both academia and industry. Negative results rarely get published; mandates are won by promising spectacular returns. These incentives can bias researchers, even unintentionally, toward experimental designs that are more likely to yield "positive" results. Ignoring implementation issues often yields stronger results. Many high-turnover strategies look attractive when ignoring trading costs, but are impossible to profitably trade. Inferences drawn from tests ignoring frictions often cannot be

[14] Tables IA.II and IA.III tabulate results similar to those presented in Table VI, but using factors that incorporate only either banding or netting, respectively.

[15] In the Internet Appendix, we also consider the possibility that investors could closely approximate the trades of the baseline factors used in Section III, but at only a fraction of our estimated effective spreads. Table IA.IV reports the "break even" costs, as a percentage of the full estimated costs, required to eliminate the differences in squared Sharpe ratios across models.

generalized into real economic insights. To be most useful, financial research must incorporate real-world concerns to the greatest extent possible.

Initial submission: March 18, 2021; Accepted: October 14, 2022
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

Appendix A: Data and Methods

This section describes our data sources and methods.

A.1. Factor Models

We compare the performance before and after trading costs of six prominent empirical asset pricing models: the Fama and French (2015) five-factor (FF5) model, the Fama and French (2018) six-factor (FF6) model, the Hou, Xue, and Zhang (2015) four-factor (HXZ4) model, the Barillas and Shanken (2018) six-factor model, and versions of the two Fama and French models that employ a cash profitability factor in place of the one based on accruals used in the baseline models.

- FF5: The five-factor model of Fama and French (2015) has factors MKT, SMB, HML, RMW, and CMA. MKT denotes the return on the Center for Research in Securities Prices (CRSP) value-weighted stock market index in excess of the risk-free rate. SMB is a size factor that is long stocks with small market capitalization and short stocks with large market capitalization, HML is a value factor that is long stocks with high book-to-market ratios and short stocks with low book-to-market ratios. RMW is a profitability factor that is long stocks with high (robust) operating profitability and short stocks with low (weak) operating profitability. CMA is an investment factor that is long stocks with low (conservative) investment and short stocks with high (aggressive) investment.
- FF6: The six-factor model of Fama and French (2018) constructed as FF5 augmented by MOM, the momentum factor of Carhart (1997). The momentum factor is rebalanced monthly and is long (short) stocks with the highest (lowest) returns over the prior 12 months excluding the prior month.
- HXZ4: The four-factor q-model of Hou, Xue, and Zhang (2015) has factors MKT, ME, IA, and ROE. ME is a size factor that is long stocks with low market capitalization and short stocks with high market capitalization, IA is an investment factor that is long stocks with low investment and short stocks with high investment, and ROE is a factor that is long stocks with high profitability (from the most recent quarterly earnings) and short stocks with low profitability.
- BS6: The six-factor model of Barillas and Shanken (2018). Barillas and Shanken (2018) use a Bayesian technique to compare the factors of FF6, HXZ4, and HML(m), which is the monthly-updated value factor of Asness

and Frazzini (2013). They conclude that the dominant model includes six factors: MKT, SMB, IA, ROE, MOM, and HML(m).

- $FF5_C$: The five-factor model of Fama and French (2015) in which the RMW factor is replaced by a profitability factor RMW_C that is constructed similarly, but replaces operating profitability with cash operating profitability, following Ball et al. (2016).
- $FF6_C$: The six-factor model of Fama and French (2015) in which the RMW factor is replaced by a profitability factor RMW_C that is constructed similarly, but replaces operating profitability with cash operating profitability, following Ball et al. (2016).

All the characteristic-based factors in FF5 (SMB, HML, RMW, and CMA) are constructed by rebalancing the underlying portfolios once a year, at the end of each June. The MOM factor, as well as the characteristic-based factors in HXZ4 (ME, IA, ROE) and the value factor in BS6 (HML(m)), are rebalanced monthly. The HXZ4 factors use a "2×3×3" construction, which sorts the universe of stocks into two groups based on market capitalization (size), three based on profitability, and three based on growth in book assets. All other factors are based on a "2×3" construction, which sorts stocks into two groups based on size, and three based on the other characteristic of the factor. RMW is based on operating profitability derived from the Compustat annual files, while ROE is based on return-on-book-equity derived from the Compustat quarterly files. Based on the availability of the quarterly earnings announcement data (RDQ), our sample period is January 1972 through December 2021.

We obtain before-cost returns on each Fama-French factor except RMW_C from the Kenneth French's website, https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. We replicate the before-cost returns on the RMW_C factor using CRSP and Compustat data following Fama and French (2018). We obtain before-cost returns to the HXZ4 factors from Lu Zhang's website, http://global-q.org/, and the HML(m) factor returns from AQR's website, https://www.aqr.com/Insights/Datasets. We also replicate all these factors and find correlations with the externally obtained versions ranging from 0.97 to 1.00.[16]

A.2. Transaction Costs

A.2.1. Stock-Level Trading Costs: To measure trading costs, we follow Novy-Marx and Velikov (2016) and use the effective bid-ask spread estimator from Hasbrouck (2009).[17] Spreads are estimated using a Bayesian-Gibbs sampler on a generalized Roll (1984) model of stock price dynamics:

$$V_t = V_{t-1} + \epsilon_t, \tag{A.1}$$
$$P_t = V_t + cQ_t,$$

[16] See the Internet Appendix for formal replication statistics.

[17] This procedure is also used in a growing literature. See, for example, Detzel and Strauss (2018), Detzel, Schaberl, and Strauss (2019), Chen and Velikov (2021), Barroso and Detzel (2021), and Novy-Marx and Velikov (2022).

where V_t is the "efficient value" of a (log) stock price; P_t is the trade price, $Q_t = +1$ (-1) if the trade was a buy (sell); ϵ_t is a random public shock to the efficient value; and c is the effective one-way transaction cost. It follows from equation (A.1) that:

$$\Delta P_t = c\Delta Q_t + \epsilon_t. \tag{A.2}$$

Hasbrouck estimates c via Bayesian methods applied to an augmented daily-frequency version of equation (A.2):

$$\Delta P_t = c\Delta Q_t + \beta r_{m,t} + \epsilon_t, \tag{A.3}$$

where $r_{m,t}$ denotes the market return on day t.

Hasbrouck (2009) documents that the effective spreads estimated using this procedure achieve a 96.5% correlation with the effective spread estimated using the Trade and Quote (TAQ) intraday data. While this measure does not account for the price impact of large trades, its estimation does not rely on intraday or proprietary data and is available for the full sample of publicly traded companies.[18] The asset pricing factors implicitly assume market orders because they transact at exactly the moment of closing. Thus, our net-of-costs factors can be interpreted as the net-of-costs return on the marginal dollar's worth of investment in the literal factors. This is arguably the relevant correction for traders using the benchmark factors as a proxy for the opportunity cost of investing in a given test asset.

The Hasbrouck procedure yields a number of missing observations, and these need to be filled in to compute trading-strategy costs. To do so, in each month t we assign to each stock i missing an estimate of effective spread that of the stock j with the closest match in terms of market cap and idiosyncratic volatility, which are the main observable correlates of transactions costs.[19] Specifically, each month we rank all stocks' market values and idiosyncratic volatilities, referring to the ranks as $rankME_i$, and $rankIVOL_i$, respectively. Then, we assign to stock i the estimated spread of the stock j with the smallest value of

$$\sqrt{(rankME_i - rankME_j)^2 + (rankIVOL_i - rankIVOL_j)^2}.$$

A.2.2. Portfolio-Level Trading Costs: For each asset pricing factor, f, we compute turnover (TO) and transaction costs (TC), following Novy-Marx and Velikov (2016), as:

$$TO_t^f = \frac{1}{2}\sum_{i=1}^{N_t} |w_{i,t} - \tilde{w}_{i,t-}|, \text{ and} \tag{A.4}$$

[18] SAS code to estimate the effective bid-ask spreads is available on Joel Hasbrouck's website at http://pages.stern.nyu.edu/ jhasbrou/.

[19] See Novy-Marx and Velikov (2016) for more details. Idiosyncratic volatility is defined as the standard deviation of residuals from an FF3 regression based on the prior 90 days of a given stock's returns.

$$TC_t^f = \sum_{i=1}^{N_t} |w_{i,t} - \tilde{w}_{i,t-}| \cdot c_{i,t}, \tag{A.5}$$

where $c_{i,t}$ (r_{it}) is the one-way transaction cost (return) of stock i at time t, N_t = number of stocks at time t, $w_{i,t}$ is the weight of stock i in f at time t after rebalancing, and $\tilde{w}_{i,t-} = w_{i,t-1}(1 + r_{it})$ is the weight of stock i in f at time t before rebalancing. The net-of-costs long return on f, f_t^{net}, is

$$f_t^{net} = f_t^{gross} - TC_t^f, \tag{A.6}$$

and the corresponding short return is

$$f_t^{S,net} = -f_t^{gross} - TC_t^f, \tag{A.7}$$

where f_t^{gross} denotes the return ignoring costs. We estimate the w_{it} of the portfolios by replicating the factors following the respective studies (specifically, Asness and Frazzini (2013), Fama and French (2015, 2018), and Hou, Xue, and Zhang (2015)).

In Section V, we consider the effects of trading diversification on transaction costs, following DeMiguel et al. (2020). Suppose a portfolio $\theta = (\theta_1, \ldots, \theta_K)'$ consists of K factors $f = (f_1, \ldots, f_K)'$ and w_{it}^k denotes the weight of stock i in factor k. Without trading diversification (the baseline case in Sections III and IV), the transaction costs of this portfolio are given by the weighted average of those in equation (A.5):

$$\begin{aligned} TC_t(\theta) &= \sum_{k=1}^{K} |\theta_k| TC_t^{f_k} \\ &= \sum_{i=1}^{N_t} \sum_{k=1}^{K} |\theta_k| \left| w_{it}^k - \tilde{w}_{i,t-}^k \right| c_{it}. \end{aligned} \tag{A.8}$$

With trading diversification, individual stock trades net across factors such that the transaction costs of the portfolio become:

$$TC_t^{\text{with TD}}(\theta) = \sum_{i=1}^{N_t} \left| \sum_{k=1}^{K} \theta_k (w_{it}^k - \tilde{w}_{i,t-}^k) \right| c_{it}. \tag{A.9}$$

Because of the triangle inequality, equation (A.9) will always be bound above by equation (A.8).

A.2.3. Transaction Costs and Maximum Sharpe Ratios: Throughout the paper, we estimate portfolios of factors with the highest Sharpe ratio net of trading costs:

$$\max_{\theta \in \mathbb{R}^K} \frac{E\left(r_{pt}(\theta)\right)}{\sigma\left(r_{pt}(\theta)\right)}, \tag{A.10}$$

where, following the definitions in equations (A.8) and (A.9),

$$r_{pt}(\theta) = f_t'\theta - TC_t(\theta), \tag{A.11}$$

when traders do not practice netting to benefit from trading diversification, and

$$r_{pt}(\theta) = f_t'\theta - TC_t^{\text{with TD}}(\theta), \tag{A.12}$$

when they do. By convention, we constrain the elements of θ to sum to one although the choice of normalization does not affect the Sharpe ratio. Problem (A.10) with equation (A.11) can easily be seen to be equivalent to problem (2) by letting a negative value of θ_k in equation (A.10) indicate a positive weight on the short version of factor k in equation (2). Thus, problem (A.10) is more general than problem (2) since it accommodates an arbitrary transaction cost function.

Two important implementation issues arise in the optimization problem (A.10). First, the short version of a given factor is not equal to the negative of the long factor (because both subtract transaction costs in equations (A.11) and (A.12), and $TC_t(-\theta) = TC_t(\theta)$). As a result, the weights for the optimal portfolio will not be the same as the well-known unconstrained tangency portfolio formula with weights proportional to $\Sigma^{-1}\mu$, where μ and Σ are the mean and covariance matrix of the factors. The second important subtlety in equation (A.10) is that the absolute values in equations (A.8) and (A.9) render the transaction costs functions not differentiable. So, in general, no closed-form formula exists for net-of-costs efficient portfolios and we must solve for them numerically. Barillas et al. (2020) propose a closed-form test statistic based on the delta method to perform pairwise comparisons of maximum Sharpe ratios, but only when they take the form $\mu'\Sigma^{-1}\mu$, which obtains when portfolio weights are proportional to $\Sigma^{-1}\mu$. This test and any other based on the delta method (see, e.g., that of Li, DeMiguel, and Martín-Utrera (2020)) require Sharpe ratios to be smooth functions of the underlying parameters, which is ruled out in general by the nondifferentiability of transaction costs. Thus, of the available tools to compare maximum Sharpe ratios, it is necessary in general to use a bootstrap design, such as that of Fama and French (2018).

Appendix B: Cumulative Return Calculations

We define a cumulative return on a long-short portfolio as: $\prod_{s=0}^{t}(1 + r_{L,s} - r_{S,s})$, where $r_{L,s}$ and $r_{S,s}$ are the monthly returns to the portfolios held long and short, respectively, and t ranges from January 1973 to December 2021. This is the cumulative performance of a strategy that initially buys \$1 of the portfolio held long and short-sells \$1 of the portfolio held short, and, motivated by Regulation T, each month posts cash collateral equal to 50% of the total equity positions in a noninterest bearing margin account. Alternatively, this may be conceptualized as the performance of the book of a trader following the strategy when the trader's margin account earns the risk-free rate but her firm charges her for the use of their capital at that same rate. A common alternative

is to assume that investors do not pay fees and therefore additionally earn the risk-free rate (see, e.g., Daniel and Moskowitz (2016)). However, we prefer our formulation, because it more accurately reflects economic profitability and does not arbitrarily reward performance due to high-inflation environments. Said differently, our choice does not impact the *relative performance* of the portfolios we compare, but does make it easier to visually discern the time variation in the performance of interest that is driven by the excess returns on the portfolios we compare.

REFERENCES

Asness, Clifford S., and Andrea Frazzini, 2013, The devil in HML's details, *Journal of Portfolio Management* 39, 49–68.

Ball, Ray, Joseph Gerakos, Juhani T. Linnainmaa, and Valeri Nikolaev, 2016, Accruals, cash flows, and operating profitability in the cross section of stock returns, *Journal of Financial Economics* 121, 28–45.

Barillas, Francisco, Raymond Kan, Cesare Robotti, and Jay Shanken, 2020, Model comparison with Sharpe ratios, *Journal of Financial and Quantitative Analysis* 55, 1840–1874.

Barillas, Francisco, and Jay Shanken, 2017, Which alpha? *Review of Financial Studies* 30, 1316–1338.

Barillas, Francisco, and Jay Shanken, 2018, Comparing asset pricing models, *Journal of Finance* 73, 715–754.

Barroso, Pedro, and Andrew Detzel, 2021, Do limits to arbitrage explain the benefits of volatility-managed portfolios? *Journal of Financial Economics* 140, 744–767.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard, 2023, Bayesian solutions for the factor zoo: We just ran two quadrillion models, *Journal of Finance* 78, 487–557.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Chen, Andrew, and Mihail Velikov, 2021, Zeroing in on the expected returns of anomalies, *Journal of Financial and Quantitative Analysis* (forthcoming).

Chen, Andrew Y., and Tom Zimmermann, 2022, Open source cross-sectional asset pricing, *Critical Finance Review* 11, 207–264.

Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong, 2014, Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics* 58, 41–58.

Daniel, Kent, David Hirshleifer, and Lin Sun, 2020, Short- and long-horizon behavioral factors, *Review of Financial Studies* 33, 1673–1736.

Daniel, Kent, and Toby J. Moskowitz, 2016, Momentum crashes, *Journal of Financial Economics* 122, 221–247.

DeMiguel, Victor, Alberto Martín-Utrera, Francisco J. Nogales, and Raman Uppal, 2020, A transaction-cost perspective on the multitude of firm characteristics, *Review of Financial Studies* 33, 2180–2222.

Detzel, Andrew, Philipp Schaberl, and Jack Strauss, 2019, Expected versus ex post profitability in the cross-section of industry returns, *Financial Management* 48, 505–536.

Detzel, Andrew, and Jack Strauss, 2018, Combination return forecasts and portfolio allocation with the cross-section of book-to-market ratios, *Review of Finance* 22, 1949–1973.

Fama, Eugene F., 1991, Efficient capital markets: II, *Journal of Finance* 46, 1575–1617.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F., and Kenneth R. French, 2018, Choosing factors, *Journal of Financial Economics* 128, 234–252.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *Journal of Finance* 75, 1327–1370.

Ferson, Wayne E., Andrew F. Siegel, and Junbo L. Wang, 2019, Asymptotic variances for tests of portfolio efficiency and factor model comparisons with conditioning information, University of Southern California Working paper.

Frazzini, Andrea, Ronen Israel, and Tobias J. Moskowitz, 2015, Trading costs of asset pricing anomalies, AQR Capital Management Working paper.

Gibbons, Michael R., Stephen Ross, and Jay Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica* 57, 1121–1152.

Harvey, Campbell R., 2017, Presidential address: The scientific outlook in financial economics, *Journal of Finance* 72, 1399–1440.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, … and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Hasbrouck, Joel, 2009, Trading costs and returns for U.S. equities: Estimating effective costs from daily data, *Journal of Finance* 64, 1445–1477.

Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, 2019, Which factors? *Review of Finance* 23, 1–35.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting anomalies: An investment approach, *Review of Financial Studies* 28, 650–705.

Kan, Raymond, Xiaolu Wang, and Xinghua Zheng, 2019, In-sample and out-of-sample Sharpe ratios of multi-factor asset pricing models, University of Toronto Working paper.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.

Ledoit, Oliver, and Michael Wolf, 2008, Robust performance hypothesis testing with the Sharpe ratio, *Journal of Empirical Finance* 15, 850–859.

Lesmond, David A., Michael J. Schill, and Chunsheng Zhou, 2004, The illusory nature of momentum profits, *Journal of Financial Economics* 71, 349–380.

Lettau, Martin, and Markus Pelger, 2020, Factors that fit the time series and cross-section of stock returns, *Review of Financial Studies* 33, 2274–2325.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics* 96, 175–194.

Li, Sicong (Allen), Víctor DeMiguel, and Alberto Martín-Utrera, 2020, Which factors with price-impact costs? London Business School Working paper.

Novy-Marx, Robert, and Mihail Velikov, 2016, A taxonomy of anomalies and their trading costs, *Review of Financial Studies* 29, 104–147.

Novy-Marx, Robert, and Mihail Velikov, 2019, Comparing cost-mitigation techniques, *Financial Analyst Journal* 75, 85–102.

Novy-Marx, Robert, and Mihail Velikov, 2022, Betting against betting against beta, *Journal of Financial Economics* 143, 80–106.

Roll, Richard, 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* 39, 1127–1139.

Ross, Stephen A., 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–360.

Stambaugh, Robert F., and Yu Yuan, 2017, Mispricing factors, *Review of Financial Studies* 30, 1270–1315.

Supporting Information

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

Appendix S1: Internet Appendix.
Replication Code.