High on High Sharpe Ratios: Optimistically Biased Factor Model Assessments*

Sara Easterwood † Bradley Paye ‡

January 23, 2023

Abstract

Estimates of maximum obtainable Sharpe ratios for prominent multifactor models are too large to be compatible with risk-based economic models. We provide evidence that this arises due to optimistic bias driven by a combination of data snooping and publication-induced learning about mispricing. We argue that common 'out-of-sample' research designs do not adequately address this bias, and we propose alternative evaluation approaches that do so. Sharpe ratio estimates fall dramatically under these approaches, both for conventional models and for models distilled from large sets of cross-sectional return predictors using machine learning methods. Reassuringly, our reduced-bias Sharpe ratio estimates do not violate "good deal bounds." However, we also conclude that multifactor model improvements relative to the capital asset pricing model (CAPM) are far more modest than suggested in the literature.

Keywords: factor models, data snooping, Sharpe ratio, anomalies, machine learning

JEL Codes: G34, G12, G14

*We thank Roger Edelen and seminar participants at the University of North Texas and Virginia Tech for helpful comments.

†Pamplin College of Business, Virginia Tech; saraeast@vt.edu

‡Pamplin College of Business, Virginia Tech; bpaye@vt.edu

1 Introduction

Estimates of Sharpe ratios associated with popular multifactor models strain credulity. As a concrete example, the estimated maximum obtainable Sharpe ratio associated with the Fama and French (2018) six-factor model is around 1.2 on an annualized basis, a value that is roughly three times the Sharpe ratio for the U.S. equity market. Other recently proposed models, such as the factors extracted from a large set of anomaly portfolios by Lettau and Pelger (2020), produce even larger estimates of maximum obtainable Sharpe ratios. Moreover, similarly large Sharpe ratio estimates obtain even under "out-of-sample" research designs that guard against the possibility of overfitting tangency portfolio weights.

How should we interpret multifactor model Sharpe ratios that are several times larger than the market Sharpe ratio? An optimist might conclude that this evidence indicates extraordinary progress in our ability to explain cross-sectional return variation relative to the classic Capital Asset Pricing Model (CAPM) of Sharpe (1964), Lintner (1965) and Mossin (1966). However, in contrast to this view, MacKinlay (1995) argues that large increases in maximum Sharpe ratios relative to the market more likely reflect deviations from CAPM driven by non-risk-based explanations such as data snooping, market frictions, or investor irrationality, than risk-based explanations. This is because risk-based deviations from CAPM imply bounded increases in maximum obtainable Sharpe ratios, whereas maximum Sharpe ratios under deviations attributable to other sources are not similarly bounded.[1] Therefore, implausibly large Sharpe ratios are a cause for concern, rather than celebration, at least insofar as additional non-market factors are purported to capture priced risks.

In this paper, we consider the question of what explains seemingly excessive Sharpe ratios associated with popular multifactor models. We identify several potential sources of upward bias in conventional estimates of maximum Sharpe ratios associated with factor models. We

[1] Ross (1976) bounds asset pricing theory residuals under the assumption that no portfolio can have more than twice the market Sharpe ratio. Similarly, Cochrane and Saa-Requejo (2000) compute bounds on asset prices under a 'good deals' restriction that rules out (annualized) Sharpe ratios in excess of one. MacKinlay (1995) posits a lower maximum achievable Sharpe ratio of around 0.6.

then test whether evaluation methods that mitigate these biases produce lower Sharpe ratio estimates that fall within conventional notions of good deal bounds. We show that estimates of (maximum) Sharpe ratios obtained for popular models fall dramatically upon adopting alternative methods that mitigate upward bias driven by the effects of data snooping (Lo and MacKinlay (1990), Harvey et al. (2016)) and publication-driven learning about mispricing (McLean and Pontiff (2016)). Somewhat paradoxically, we interpret these lower Sharpe ratio estimates as 'good news' for multifactor models, in the sense that the lower estimates reflect plausible risk-based explanations of deviations from the CAPM. The corresponding Sharpe ratios are generally well within standard good deal bounds and are plausibly consistent with maximum Sharpe ratios for calibrated versions of benchmark economic models. However, we also find that reduced-bias Sharpe ratio estimates across a wide array of popular multifactor models are roughly similar to one another and to the market Sharpe ratio. Thus, at least from a *forward-looking* perspective, our results challenge the accumulated perception that increasingly sophisticated multifactor models are clearly superior to simpler multifactor models such as the three-factor model of Fama and French (1993), or even the classic CAPM.

Empirically, we first consider a relatively prosaic explanation for excessive Sharpe ratios reported in the literature: estimates computed using standard 'plug-in' estimates of unknown factor means and covariances are upward biased in finite samples (Jobson and Korkie (1981)). We compute alternative estimates that correct for this bias. The evidence indicates that this source of bias is relatively small and is not the primary explanation for apparent violations of good deal bounds.[2]

Given that classic finite-sample bias does not resolve the Sharpe ratio puzzle, we consider other potential sources of bias. One involves data snooping with respect to factor model specification. Characteristics-based factors are often motivated by return patterns known to exist in prior data. Thus, when Sharpe ratios or other model metrics are computed using data that overlaps with samples examined in earlier studies, the resulting metrics are

[2]For example, alternative bias-reduced Sharpe ratio estimates for the Fama-French 6 factor model remain in excess of 1.1 on an annualized basis.

subject to optimistic bias. This reflects a form of 'overfitting,' such that model specification is based, at least in part, on known empirical outcomes, and is in the spirit of pre-test biases discussed in, e.g., Leamer (1978), and emphasized by Lo and MacKinlay (1990) in the context of evaluating asset pricing models. An additional source of upward bias relates to (bona fide) mispricing that is uncovered by academic research and subsequently reduced or eliminated by practitioners who trade to exploit the mispricing, as in McLean and Pontiff (2016). Sharpe ratios measured using factor returns that predate academic discovery of corresponding anomalous return patterns are optimistically biased because they embed mispricing that has been subsequently reduced or eliminated. To motivate the potential importance of these channels, we contrast pre- and post-publication average returns for 11 prominent characteristics-based factors. On a pooled basis, the average post-publication return falls by around 70% relative to pre-publication.

Researchers often evaluate factor models using 'out-of-sample' (OOS) empirical designs, and several recent studies report OOS Sharpe ratio estimates for multifactor models that are well in excess of one on an annualized basis (e.g., Fama and French (2018) and Lettau and Pelger (2020)). However, we show that popular OOS designs involve considerable overlap between the OOS analysis period and sample periods previously analyzed in studies that document anomalous return patterns associated with the characteristics featured in the model.[3] Thus, OOS returns analyzed in such studies are not free of biases related to data snooping and publication-induced learning. Although some researchers acknowledge potential optimistic bias even in 'out-of-sample' designs,[4] the extent of this bias is unclear.

In order to shed light on the potential magnitude of biases in Sharpe ratio estimates, we

[3]As a concrete example, an OOS evaluation of the Fama-French 6 factor model using a sample of factor returns beginning in 1963 and an initial estimation window of 10 years results in over 40% of OOS factor-month observations overlapping with samples analyzed in original papers documenting the underlying characteristics.

[4]Freyberger et al. (2020) estimate an annualized OOS Sharpe ratio in excess of 2.5 for a nonparametric model extracted from a large set of firm characteristics using machine learning methods. They note that "The characteristics we study are not a random sample, but have been associated with cross-sectional return premiums in the past. Therefore, we focus mainly on the comparison across models rather than emphasizing the overall magnitude of the Sharpe ratios." (p. 2328).

contrast pre-publication and post-publication Sharpe ratios for characteristics-based factors. We focus on characteristics for which sufficient post-publication data exists for reasonably precise estimates. Post-publication Sharpe ratios are typically over 50% lower than pre-publication sample Sharpe ratios for these factors, and are typically less than 0.25 on an annualized basis. We then compare post-publication estimates of maximum Sharpe ratios for multifactor models with full sample estimates for models with sufficient post-publication data. Similar to results for individual factors, post-publication Sharpe ratio estimates for these models are typically much lower than full sample estimates. The estimates are similar to the market Sharpe ratio over the same period, and we generally cannot reject the null hypothesis of equal Sharpe ratios.

Some popular models incorporate characteristics that were documented in the anomaly literature only recently. This implies that post-publication samples are very short, leading to highly imprecise Sharpe ratio estimates. In order to circumvent this constraint, we evaluate 'adaptive' versions of models. Such models begin as the CAPM, and gradually evolve into the corresponding destination models (e.g., the Fama-French 5 factor model) as characteristics motivating additional non-market factors are documented in academic literature. This feature ensures that factor and tangency portfolio returns associated with the adaptive model do not include data analyzed in original papers that document the underlying characteristics in the anomaly literature. The Sharpe ratio for the adaptive version of a model can be interpreted as an estimate of the Sharpe ratio that a real-time investor could achieve, assuming the investor is sophisticated in the sense of closely following academic literature and quickly incorporating newly proposed factors, but not unnaturally prescient in the sense of being able to identify characteristics that predict returns in the cross section well before they were identified in academia.[5] These Sharpe ratios are informative concerning good deal

[5]Consider, for example, models that include the size factor proposed by Fama and French (1993). Under such a model, the mean-variance efficient tangency portfolio will take a position in the SMB factor, in addition to the market (and potentially other factors). However, adopting a portfolio of this form during the 1960s and 1970s implicitly assumes an extraordinary degree of investor prescience, as the size effect had yet to be documented and the seminal single factor CAPM had only recently been proposed. The notion of evaluating adaptive versions of models therefore follows naturally from considering factor models

bounds, because such bounds presumably relate to Sharpe ratios that could plausibly be achieved by real-time investors.

We compare conventional Sharpe ratio estimates with estimates based on the adaptive version of the corresponding model.[6] Sharpe ratios based on standard in-sample tangency portfolio returns exceed the Sharpe ratio of the market factor and often exceed one on an annualized basis. Sharpe ratio estimates continue to violate good deals bounds under a conventional OOS research design or using simple equal-weights for factors. However, Sharpe ratio estimates fall precipitously upon considering adaptive versions of the models. Strikingly, we find that Sharpe ratios for adaptive versions of the models are often roughly equal to or below that of the market Sharpe ratio. Similarly, we show that alternative asset pricing metrics characterising the magnitude of pricing errors with respect to a set of test assets become substantially less optimistic under the adaptive models.

In contrast to traditional ad-hoc characteristics-based models in the spirit of Fama and French (1993), several recent studies distill factors from large sets of characteristics or related portfolios by applying forms of regularization or related machine learning methods.[7] Estimates of Sharpe ratios reported in the literature for these approaches are large, even using OOS designs. The underlying 'basis' characteristics that serve as inputs for these methods are typically chosen to be characteristics from the accumulated literature on cross-sectional return anomalies. This creates the potential for optimistically biased assessments of the resulting model performance similar to that for ad-hoc characteristics-based models.

As an illustration, we consider the performance of models based on principal components analysis (PCA) applied to portfolios constructed as sorts on a large number of characteristics

as representing real-time proxies for an efficient portfolio.

[6]Operationally, we add new factors starting in the publication year of the original paper documenting the characteristic that underlies a particular factor, or alternatively in the ending year of the data sample analyzed in that paper. For example, original documentation of the size effect in returns is generally attributed to Banz (1981). Our treatment is conservative, in the sense that we permit the SMB factor to be included in adaptive models starting in 1975 (sample year end criterion) or 1981 (publication year criterion) based on Banz (1981) rather than approximately a decade later when the SMB factor was formally proposed by Fama and French (1993).

[7]Selected examples include Kelly et al. (2019), Freyberger et al. (2020), Kozak et al. (2020), and Lettau and Pelger (2020).

from the anomaly literature, as considered in, e.g., Kozak et al. (2018), Kozak et al. (2020), and Haddad et al. (2020). We replicate large Sharpe ratios for statistical factors constructed in this way, both in-sample and using standard out-of-sample designs. We then expand the basis set of portfolios used to extract factors to include industry portfolios and portfolios that reflect exposure to macroeconomic factors in the spirit of, e.g., Chen et al. (1986), in addition to standard anomaly characteristics. This produces even higher in-sample Sharpe ratios. However, OOS estimates fall on the order of 50% for models with 4–6 factors and are well below one on an annualized basis.[8] Estimates fall further if the set of included characteristics is limited to those documented in published literature at the time.

The much smaller Sharpe ratio estimates that we obtain in this paper might seem like bad news for popular multifactor models. However, we suggest a more positive interpretation. It is reassuring that popular models do *not* violate standard notions of good deal bounds in the sense that real-time investors who 'factor invest' using these models after they are proposed do not achieve exorbitant Sharpe ratios. Our results suggest that improvements offered by popular models relative to the CAPM are much smaller than suggested by existing literature. But this is precisely what one should expect under risk-based alternatives, whereas unnaturally large Sharpe ratios are evidence against risk-based explanations (MacKinlay (1995)). Admittedly, our results also pose challenges. In particular, many models perform similarly upon adopting evaluation methods that mitigate pre-test biases and the effects of publication-induced learning. This exacerbates the already challenging problem of factor model specification, and leaves ample room for yet more work on this time-honored topic.

Related Literature

Our paper contributes to a literature documenting secular reductions in average returns or premia associated with anomaly characteristics following publication. Related papers include Schwert (2003), McLean and Pontiff (2016), and Smith and Timmermann (2021).

[8] Early literature on multifactor models focuses on both macroeconomic and industry-based factors, and thus the expanded set of basis portfolios better reflects ex ante agnosticism concerning whether industry-based, macroeconomic, or characteristics-based factors will better explain the cross-section of returns.

This tendency for average returns to decline may be driven, at least in part, by data snooping, in the spirit of, e.g., Lo and MacKinlay (1990), Harvey et al. (2016), Harvey (2017), Linnainmaa and Roberts (2018), Harvey and Liu (2021), and Lopez-Lira and Roussanov (2021). Much of the data snooping literature focuses on the properties of statistical tests for the validity of models or the relevance of particular individual characteristics.[9] In contrast, but related to this work, we focus on obtaining reduced-bias estimates of Sharpe ratios and other performance metrics for multifactor models.

Bessembinder et al. (2021) report evidence of significant time-variation in the number of factors required to explain the cross-section of returns. Related to our study, they document economically large OOS Sharpe ratios for factor models constructed as principal components from a large set of characteristics-based hedge portfolios. They find that many factors are statistically significant in periods before and after the sample studied by original authors. We emphasize post-publication outcomes, which incorporates effects of publication-induced learning as in McLean and Pontiff (2016) as well as traditional data snooping. With regard to PCA-type factors, we highlight the role of potential hindsight bias in the choice of underlying portfolios by demonstrating how Sharpe ratios fall when the ex ante portfolio set is less oriented around anomaly characteristics. Bessembinder et al. (2022) show that even the specific composition of anomaly-based portfolios can have a significant impact on Sharpe ratios for PCA-based factors.

Other related studies show that many anomaly hedge portfolios take positions in relatively small, illiquid firms, and/or entail significant transactions costs. For example, Novy-Marx and Velikov (2016) study the extent to which transactions costs erode the profitability of anomaly strategies and find that few strategies with turnover greater than 50% per month yield significant return spreads, even after efforts to mitigate transactions costs. Hou et al. (2018) examine a larger set of anomalies and find that many do not remain significant after

[9] The extent to which data snooping drives apparent anomalous return patterns remains debated in the literature. For example, Chen and Zimmermann (2022) apply alternative false discovery rate estimators and conclude that most claimed results in the literature are true. Chordia et al. (2020) estimate the proportion of false discoveries in the absence of multiple testing corrections to be around 45%.

excluding microcaps and value-weighting returns.[10] Our paper shows that Sharpe ratios and other factor model performance metrics erode substantially upon adopting methods that mitigate biases even before considering transaction costs. Explicitly accounting for such costs would further reduce the maximum Sharpe ratios obtainable for the models.

Pesaran and Timmermann (1995) consider predictability in excess US stock market returns from a real-time perspective. They point out that approaches described as "out-of-sample" nevertheless condition upon the relevance of a particular predictive model or set of models. This "inevitably raises the possibility that the choice of the model could have been made with the benefit of hindsight." We adopt a similar real-time perspective concerning factor models and propose evaluation methods to mitigate the problem. From a theoretical perspective, Da et al. (2022) show how learning about investment characteristics (mispricing in their context) creates a wedge between ex post Sharpe ratios measured by an econometrician versus feasible Sharpe ratios for real-time investors. Finally, our focus on addressing hindsight bias distinguishes our paper from others that focus on econometric issues in comparisons of factor models including, e.g., Kan et al. (2013), Barillas and Shanken (2018), and Fama and French (2018).

2 Factor Model Sharpe Ratios and Optimistic Bias

We aim to evaluate the performance of popular asset pricing models based on factors constructed from (U.S.) stocks. What differentiates our paper from previous studies is the evaluation method we adopt and contrast with conventional methods. This section discusses sources of bias in Sharpe ratios and other metrics associated with factor models, highlighting potentially important biases related to model specification and academic research patterns. We then consider alternative evaluation methods that mitigate optimistic bias.

[10] Related to this work, Bessembinder et al. (2022) compare the set of 207 anomalies studied in Chen and Zimmermann (2022) to the set of 153 anomalies studied in Jensen et al. (2022) and find that value-weighting portfolios and defining portfolios based on less extreme values of the underlying characteristics both lead to substantially smaller Sharpe ratio estimates.

2.1 The issue with high Sharpe ratios

A high maximum obtainable Sharpe ratio would seem to be a positive attribute for a factor model. Indeed, if a factor model is correctly specified in the sense that it spans the tangency portfolio for the full investment universe, then the factors maximize the attainable Sharpe ratio. Furthermore, Barillas and Shanken (2017) show that, in comparing the relative pricing performance of models with traded factors, test-asset returns are irrelevant and the model that achieves the higher (maximum) Sharpe ratio is preferred. This discussion refers to unobserved population Sharpe ratios. Of course, it is possible to estimate population Sharpe ratios and perform statistical inference. However, the resulting estimates can be biased for several reasons, as discussed further below.

MacKinlay (1995) argues that Sharpe ratio improvements relative to the CAPM or other benchmark factor models are bounded under the premise that additional factors capture risk-based variation in expected returns. In contrast, improvements that reflect alternatives such as data snooping, market frictions, or investor irrationality are not similarly bounded. Thus, excessive Sharpe ratio estimates are more consistent with non-risk-based alternatives than risk-based alternatives. The essence of MacKinlay (1995)'s argument can be appreciated via the Hansen-Jagganathan bound (Hansen and Jagannathan (1991)), which relates maximum Sharpe ratios to a lower bound on the volatility of any stochastic discount factor (SDF) that can price the corresponding assets. Specifically, let $SR^2(f)$ denote the maximum squared Sharpe ratio obtainable from a set of traded factors f. Then the Hansen-Jagganathan bound can be written as

$$\sigma(m) \geq (1/R_f)\,|SR(f)|, \tag{1}$$

where m denotes any SDF that prices the factors f and R_f equals the gross risk-free rate. The key implication of Eq. (1) is that a high obtainable Sharpe ratio associated with a factor model imposes a large minimum volatility for for any SDF that is able to price the factors. As $SR(f)$ becomes very large, it becomes difficult for plausible calibrations of leading economic

SDF models, such as, e.g., the habit model of Campbell and Cochrane (1999), to satisfy the bound. Calibrations of most popular models imply maximum annualized Sharpe ratios substantially below one. Thus, estimated annualized Sharpe ratios near or above one are problematic in the sense that they seem incompatible with risk-based explanations given existing models in the literature.

2.2 Potential sources of bias

We focus on the popular class of linear models for the SDF m:

$$0 = E(m_{t+1} R^e_{t+1}), \tag{2}$$

$$m_{t+1} = 1 - b'[f_{t+1} - E(f)], \tag{3}$$

where R^e is an arbitrary excess return, f_{t+1} denotes a $K \times 1$ vector of excess returns on a set of tradeable factors, and b is a $K \times 1$ vector of factor loadings.[11] The factor loadings vector b is proportional to the tangency portfolio weights for the factors:

$$b = \text{Cov}(f)^{-1} E(f), \tag{4}$$

where $\text{Cov}(f)$ denotes the $K \times K$ factor covariance matrix. The maximum squared obtainable Sharpe ratio associated with the factors equals $SR^2(f) = E(f)' \, \text{Cov}(f)^{-1} E(f)$.

A natural and popular approach to estimating $SR^2(f)$ 'plugs-in' sample analogs of the first and second moments for the factors, e.g.,

$$\widehat{SR^2(f)} = \widehat{E(f)}' \, \widehat{\text{Cov}(f)}^{-1} \widehat{E(f)}, \tag{5}$$

where $\widehat{\text{Cov}(f)}$ and $\widehat{E(f)}$ denote sample analogs of the corresponding population quantities.

[11] Note that we normalize the intercept of the SDF to one in this equation, which reflects our focus on pricing excess returns for risky assets and not the risk-free rate.

This estimator is consistent, but upward biased in finite samples see, e.g., Jobson and Korkie (1981)). Thus, Sharpe ratio estimates reported in the literature that appear to violate good deal bounds could simply be attributable to severe finite sample bias. We consider various reduced-bias estimators to assess whether this potential explanation is compelling.[12]

A second potential source of optimistic bias involves 'overfitting' in the process of factor selection. Since the seminal work of Fama and French (1992) and Fama and French (1993), it has become common to specify factors as long-short hedge portfolios constructed by sorting firms according to 'anomaly' characteristics documented in academic literature. This approach to specifying factors is subject to what Lo and MacKinlay (1990) term "data-instigated pretest biases discussed in Leamer (1978)."[13] The overfitting problem is prominent in many areas of data science. A standard approach to obtaining unbiased model assessments is to validate model performance using a 'holdout' dataset that has not been used for model selection or estimation. However, this certainly does not describe typical "in-sample" estimates of Sharpe ratios for factor models in the literature and therefore these estimates are subject to bias associated with data snooping.

Researchers cognizant of potential overfitting often report Sharpe ratios and other factor model metrics using "out-of-sample" designs. Common OOS designs estimate tangency portfolio weights at each point in time using a historical window of factor data that would have been available to practitioners in real time. Tangency portfolio weights are typically re-estimated each period using a specified window of historical data. The return associated with the (estimated) tangency portfolio for the subsequent month is then recorded. Proceeding over time in this fashion generates an OOS time series of tangency portfolio returns used to estimate Sharpe ratios or other metrics associated with the model.

[12]For example, assuming Gaussian returns, Kan and Zhou (2007) derive an unbiased estimate of SR^2, as well as an alternative reduced bias estimator that is guaranteed to be positive. Explicit details appear in the Internet Appendix.

[13]MacKinlay (1995) notes that "the Fama and French approach to building the extra factors will tend to create a portfolio like the optimal orthogonal portfolio independent of the explanation for the CAPM deviations." In other words, constructing factors from identified anomaly characteristics will generate apparent deviations from CAPM even if these result from data snooping.

OOS research designs of the variety discussed above address optimistic bias related to the use of ex post optimized tangency portfolio weights. However, in general they do not address optimistic bias associated with selecting factors based on documented properties of data analyzed in prior studies. This is because OOS Sharpe ratios are based on a time series of tangency portfolio returns and it is common to analyze a relatively long time series of such returns in order to obtain reasonably precise estimates. This implies that the 'out-of-sample' analysis period overlaps considerably with sample periods analyzed in prior literature establishing cross-sectional return anomalies. In the following sections, we provide empirical evidence that the degree of overlap is significant for standard OOS designs, especially for recently proposed models such as the Fama-French five- or six- factor models.

A third source of optimistic bias involves secular declines in average anomaly returns driven by investor learning and related valuation effects. Investors who discover cross-sectional return patterns believed to reflect mispricing have an incentive to trade against this mispricing. One means by which investors might discover mispricing is via academic research output. McLean and Pontiff (2016) test for decay in the average returns of long-short hedge portfolios following the publication of academic articles documenting characteristics associated with anomalous return patterns. They document a post-publication decay in average anomaly returns and attribute this to investors learning about signals from academic literature and trading to exploit mispricing. In contrast to the data snooping case, large historical average returns and Sharpe ratios under this scenario are not the result of fitting noise. Instead, they reflect mispricing that has been reduced or eliminated after discovery via the academic research process.[14] Sharpe ratios estimated using pre-publication data are optimistically biased in the sense that an investor who forms a portfolio from the factors

[14] In a related vein, Martin and Nagel consider the problem of testing for market efficiency in a setting in which a high-dimensional set of characteristics potentially contain information concerning firm cash flows. They show that a 'factor zoo' arises even in the absence of data mining. Although Martin and Nagel (2021) focus on a rational pricing paradigm, whereas McLean and Pontiff (2016) focus on potential mispricing, both studies emphasize that cross-sectional return patterns identified in historical data may be difficult or impossible to exploit by current investors.

would achieve a lower Sharpe ratio on a going-forward basis.[15]

2.3 Alternative approaches

In order to avoid biases related to data snooping or publication-induced learning, it is natural to measure performance using factor returns that post-date the introduction of the corresponding factor(s) in academic literature. Therefore, we undertake this exercise for characteristics-based factors and factor models for which there exists sufficient post-publication data in order to obtain reasonably precise Sharpe ratio estimates.

The rapidly evolving nature of the empirical literature makes it challenging to apply this 'gold standard' approach comprehensively. Many currently popular models incorporate characteristics from relatively recent empirical literature. As a concrete example, the five-factor model of Fama and French (2015) analyzes a sample of data that ends in 2013, and incorporates factors based on profitability and investment that were linked with cross-sectional return patterns in papers appearing in the mid 2000s (see Table 1).

As an alternative approach that is more widely applicable, we evaluate the performance of 'adaptive' versions of factor models. The concept is quite simple: the adaptive version of a factor model begins as the CAPM, as virtually all return-based factor models include a market factor, and subsequently evolves to incorporate additional factors as the characteristics that underlie these factors are documented in the academic literature. By construction, adaptive versions of models do not incorporate factors during periods that overlap with the original sample periods that establish the anomaly characteristics underlying the (non-market) factors. This feature mitigates optimistic bias.

The adaptive model approach mimics the real-time evolution of thought of an investor who eventually adopts the 'destination' model, e.g., the Fama-French five-factor model. The

[15] Martin and Nagel (2021) assume that *all* potentially relevant variables are considered at all times by investors, and thus under their assumptions a standard OOS research design provides unbiased inference. Almost invariably in practice, only a subset of potentially relevant variables are actually examined and the selection of these variables is motivated by prior research outcomes. In this case, a standard OOS design will not generally be unbiased.

Sharpe ratio produced under the adaptive version of a model is informative concerning good deal bounds because it reflects the performance of an investor who is sophisticated in the sense of rapidly incorporating factors into the model once they are proposed, but who is not unnaturally prescient in the sense of investing in non-market factors well before there is (academic) empirical evidence that the underlying characteristics predict returns in the cross-section.

The discussion to this point focuses on ad-hoc characteristics-based models. However, other approaches to estimating factors exist. In particular, statistical factor models extract factors from a high-dimensional set of returns. Classic versions of this approach seek to implement the arbitrage pricing theory of Ross (1976) by applying principal components analysis or related methods to individual stock or portfolio returns (see, e.g., Connor and Korajczyk (1993)). Recently, several studies aim to 'tame the factor zoo' by constructing a factor model or SDF from a large set of anomaly characteristics or portfolios sorted with respect to such characteristics using principal components or machine learning methods.

Machine learning approaches are designed to consider very large information sets in which much information may be irrelevant. Thus, the second approach we consider permits investors to consider the entire 'zoo' of characteristics at all points in time, but explores the role of the composition of the zoo itself. From an ex ante perspective, there are many types of characteristics or information that might plausibly relate to cross-sectional differences in returns based on theoretical considerations. These include, e.g., industry membership information, covariances or betas with respect to macroeconomic factors constructed as in Chen et al. (1986), and a massive number of additional firm variables that might include, e.g., rich text information as well as standard price and accounting data. Our second approach therefore explores how OOS Sharpe ratios for proposed machine learning approaches vary as we increase the set of ex ante characteristics (or sorted portfolios) that are plausibly relevant such that this set is no longer dominated by characteristics associated with return anomalies ex post. We argue that Sharpe ratios obtained in such a setting more accurately reflect plausible real-time investment strategies that do not benefit from hindsight by tilting the composition of the information set toward anomaly characteristics from the literature.

3 Data

We obtain monthly returns for the market factor, the size factor (SMB), the value factor (HML), the profitability factor (RMW), the investment factor (CMA), the momentum factor (UMD), and the short-term and long-term reversal factors (STR and LTR, respectively) from Ken French's website. Monthly returns for the 'betting against beta' (BAB) factor proposed by Frazzini and Pedersen (2014) and the 'quality minus junk' (QMJ) factor constructed by Asness et al. (2019) are obtained from AQR's data library. We obtain data on the investment and return on equity factors of Hou et al. (2015) from the Global-Q data library, for a total of eleven characteristics-based factors.

In addition to the aforementioned characteristics-based factors, we construct sets of factors that are distilled from various supersets of 'anomaly portfolios' following several recent papers. To this end, we obtain monthly returns for a set of 88 extreme decile portfolios associated with 44 different anomalies from Serhiy Kozak's website.

A key aspect of our analysis concerns the time at which anomalous return patterns were "discovered" and made public. We assess this primarily via published academic research. We obtain the year of publication for over 300 characteristics or "return signals" from the "Open Source Asset Pricing" website maintained by Andrew Y. Chen and Tom Zimmermann and related to Chen and Zimmermann (2022). We obtain the sample period used to establish the main cross-sectional return relation documented in the original paper proposing each signal from the "Open Source Asset Pricing" website as well.

Panel A of Table 1 lists the eleven prominent characteristics-based factors. The first two columns identify the factor. The third column shows the original academic paper credited with documenting the cross-sectional relevance of the characteristic underlying the factor, including the paper's publication year. The fourth column lists the time period (in years) spanned by the data sample from the original paper documenting the characteristic. The difference between the sample end year and the publication year for the original studies proposing the characteristics ranges from one year (value) to six years (size) and is around three years on average. Panel B of Table 1 summarizes the 44 characteristics associated with anomalies that are used to form sorted-portfolios analyzed in, e.g., Kozak et al. (2020).

The publication dates and sample periods listed in Table 1 reveal that, although a handful of characteristics are associated with papers published in the 1970s or early 1980s, most anomaly characteristics have been discovered relatively recently. Figure 1 summarizes academic research activity associated with several alternative sets of anomaly characteristics. For each set, Figure 1 shows the cumulative number of publicly available anomaly characteristics over time, where public status is defined using the publication year (solid blue line) and ending year of the data sample analyzed (dashed red line) in the original studies documenting the characteristics. Panel A presents results for characteristics underlying the eleven popular pricing factors considered in Panel A of Table 1, Panel B covers the 44 characteristics from the Kozak dataset, and Panel C covers over 300 signals considered in Chen and Zimmermann (2022).[16] Results in Figure 1 illustrate the secular increase in the number of published anomaly characteristics from the early 1980's to the late 2010's. Panels A and B show a fairly steady growth rate in the number of characteristics over the prior four decades, with a median publication year in the mid 1990s. The larger sample in Panel C indicates that there was an inflection point in the 2000's, with very high growth in the number of published anomaly characteristics during this decade.

[16] Publication and sample period data regarding the 326 signals considered in Chen and Zimmermann (2022) are also obtained from the "Open Source Asset Pricing" website.

4 Factor Model Sharpe Ratios: Empirical Results

Table 2 shows benchmark estimates of annualized Sharpe ratios for several prominent factor models and variations that add other selected factors. Columns (1)–(3) show alternative estimates of Sharpe ratios based on monthly tangency portfolio returns over the period 1963–2020. The first column reports standard plug-in Sharpe ratio estimates. The MKT factor annualized Sharpe ratio of around 0.44, similar to estimates reported elsewhere in the literature. The estimated Sharpe ratio for the Fama-French three-factor model equals around 0.62 and that for the Fama-French five-factor model equals 1.06. An eight-factor model that augments the Fama-French five-factor model with the UMD, BAB, and QMJ factors achieves a Sharpe ratio in excess of 1.5, as does a somewhat ad hoc model that includes the STR and LTR factors, the BAB factor, and QMJ along with the market.

The plug-in, full sample Sharpe ratios are so large as to strain credulity. A possible explanation is that these Sharpe ratios are in-sample optimized, in the sense of being based on ex post estimates of the means and covariance matrix for each factor set over the sample period. Column (2) shows alternative full sample results that assume tangency portfolio weights are equal. This can be viewed as an extreme shrinkage estimator of the tangency portfolio weights. Column (3) of Table 2 shows alternative Sharpe ratios computed using the Kan and Zhou (2007) unbiased estimator for the squared Sharpe ratio. The Sharpe ratios in the second and third columns fall relative to those in the first column, but not by very much. Consequently, in-sample overfitting of tangency portfolio weights does not appear to be the dominant explanation for excessively large Sharpe ratios associated with popular multifactor models.

The final two columns of Table 2 report estimates using out-of-sample designs that estimate tangency portfolio weights for each model with either an expanding or rolling window of historical data (the initial window is set to 10 and 15 years, respectively). In each period, given the current estimated tangency portfolio weights, the OOS return of the portfolio is recorded. The Sharpe ratio is then computed as the plug-in estimate using the resulting time series of OOS returns. In Section 2, we argue that pseudo-OOS designs of this type are unlikely to fully control for implicit look-ahead bias because the OOS evaluation period overlaps with portions of the sample periods analyzed in the literature documenting the anomalous return patterns. Columns (4) and (5) of Table 2 show that, although OOS Sharpe ratio estimates tend to be lower than full-sample analogs, they exceed one in many cases, especially for more recently proposed models such as the Fama-French five-factor model and the Q4 model. In some cases, the OOS Sharpe ratio estimates are close to 1.5, which is nearly three times the market Sharpe ratio measured over the OOS period.

4.1 Post-publication declines in factor Sharpe ratios

Given that classic finite-sample bias does not resolve the Sharpe ratio puzzle, we consider the potential role of biases driven by data snooping and publication-induced learning about mispricing. To emphasize the relevance of bias in factor model assessments associated with academic research patterns, we first highlight several empirical features of the anomaly and factor data.

In Table 3, we document the decay in average returns following publication for many prominent characteristics-based factors and anomaly portfolios, in the spirit of, e.g., Schwert (2003) and McLean and Pontiff (2016). All eleven characteristics-based factors exhibit positive and statistically significant premia over the full sample period, although the significance of the size premium is somewhat marginal. The annualized market risk premium is approximately 6.8% per year. Among the non-market factors, the momentum and betting against beta factors exhibit the highest premiums and the long-term reversal factor has the smallest estimated premium of approximately 2.3% per year. Columns (5) and (6) show that virtually all of the non-market factors earn substantially higher premia prior to publication. Among the individual factors, post-publication average returns are insignificantly different from zero for all but the betting against beta, profitability, and momentum factors. The final column shows that post-publication premium reductions are economically significant for all factors except the operating profitability factor (roughly 4.5% per year), and statistically significant at conventional levels for seven among the eleven factors. The bottom two rows present pooled results for the eleven non-market factors and for an alternative set of 44 characteristics-based anomaly hedge portfolios. The pooled mean return pre-publication is roughly 0.5% per month for both sets, falls to around 0.15% post-publication for both sets, and the decrease is statistically significant in both cases. This reflects a substantial post-publication reduction in average returns of approximately 70%, confirming the relevance of optimistic bias in factor model assessments related to both data snooping and publication-induced learning.[17]

Next, we provide evidence that popular out-of-sample research designs are not immune from look-ahead bias. Figure 2 shows the fraction of in-sample factor return observations that are included in OOS evaluation periods for a variety of models and anomaly portfolios. Factor-month or anomaly-month return observations are defined as in-sample for all months from the beginning of the sample until the publication year of the original paper documenting the anomaly signal. Out-of-sample evaluation periods are defined as the periods over which rolling or expanding window OOS tangency portfolio returns can be constructed, where the rolling or expanding windows are defined using 10 or 20 years of data and the total sample period extends from 1963–2020 (e.g., a 10-year window implies that the OOS evaluation period extends from 1973–2020).[18]

The proportion of in-sample observations in 'out-of-sample' evaluation periods is large for many multifactor models. For an OOS evaluation period that extends from 1983–2020, over 30% of the observations for the Fama-French five-factor model overlap with samples

[17] In the Internet Appendix, we show that qualitatively similar results obtain when we use the end of the sample period analyzed in the original paper documenting the characteristic rather than the publication year.

[18] As a concrete example, the size and value factors are considered in-sample until 1981 and 1985, respectively. Under a 10-year rolling or expanding window design where data begins in 1963, this implies that tangency portfolio returns deemed 'out-of-sample' are not truly out-of-sample from 1973–1981 for the size factor and 1973–1985 for the value factor. For a sample that ends in 2020, this implies that nearly 20% of factor-month observations (excluding the market factor) for the Fama and French (1993) three-factor model that appear in the out-of-sample period are not truly out-of-sample.

examined in earlier studies documenting the underlying characteristics. For a 1973–2020 OOS evaluation period, 55–60% of the observations for models that incorporate the BAB and QMJ factors are not truly out-of-sample. For the sample of 44 anomaly portfolios obtained from Serhiy Kozak's website, 45-55% of OOS observations are not truly out-of-sample.

Finally, in Figure 3, we report Sharpe ratios computed pre- and post-publication for many popular factors and long-short portfolios. We focus on characteristics with a minimum of 20 years of post publication data so that we can obtain reasonably precise post-publication Sharpe ratio estimates. For each of the factors or anomaly-based hedge portfolios, the figure contrasts Sharpe ratios computed using pre-publication annualized monthly returns for a long-short portfolio based on the corresponding characteristic (blue bars) with the analogous Sharpe ratio computed using post-publication monthly returns (orange bars). Factors and anomalies are ordered according to the magnitude of full sample Sharpe ratios and vertical black bars indicate a 95% confidence interval. The figure indicates an economically significant post-publication reduction in the Sharpe ratio for nearly all factors and anomaly hedge portfolios. In the majority of cases, post-publication Sharpe ratios are less than half the magnitude of pre-publication Sharpe ratios, and post-publication Sharpe ratios are typically less than 0.25 on an annualized basis.[19]

4.2 Alternative factor model Sharpe ratios

The Fama-French three factor and Fama-French-Carhart four factor models represent classic alternatives to the CAPM. Because both models were proposed in the 1990s, and the factors in the models are associated with characteristics proposed between 1981 and 1993, it is possible to construct a post-publication 'validation' dataset that encompasses a period over which all factors in the model are publicly available and that is sufficiently long to permit reasonably precise Sharpe ratio estimates. Table 4 presents estimates over various validation samples for these two models. The validation samples begin in the year in which the final

[19] Pre- and Post-Publication and Pre- and Post-Sample End Sharpe ratios for all factors and long-short portfolios are reported in Figure 1 in the Online Appendix. Results are qualitatively similar to Figure 3.

characteristic (value in Panel A, and momentum in Panel B) was published, the sample end year of the study documenting the final characteristic, or the sample end year of the paper that proposed the factor model (Fama and French (1993) in Panel A, and Carhart (1997) in Panel B). Sharpe ratios are then computed from January in the year following the 'public' year through December 2020, or December 2018 in cases where we drop COVID years.

The first column of Table 4 indicates the validation sample over which Sharpe ratios are estimated. These samples range from 27 to 36 years in length. The second column presents plug-in Sharpe ratios computed over each evaluation period with standard errors (assuming i.i.d. returns) reported in parentheses. Z-scores for tests of differences between reported Sharpe ratios and the market Sharpe ratio (reported in the final column) are calculated following Jobson and Korkie (1981) and reported in brackets. Sharpe ratios computed using ex post tangency portfolio weights for the validation sample period equal around 0.6 for the Fama-French three-factor model, and around 0.8 for the Fama-French-Carhart four-factor model. Five out of the six reported plug-in Sharpe ratio estimates are not statistically significantly greater than the market Sharpe ratio. The third and fourth columns of Table 4 report Kan and Zhou (2007) bias-corrected Sharpe ratios and Sharpe ratios for equally-weighted tangency portfolios, respectively. In Panel A, Sharpe ratio estimates fall to around 0.5-0.55. These estimates are just below the market Sharpe ratio computed over the same period. In Panel B, Sharpe ratio estimates fall to around 0.65-0.73. These estimates exceed the market Sharpe ratio measured over the same period but are not statistically significantly greater than the market Sharpe ratio.

Although Sharpe ratio estimates in columns (2) and (3) are computed over validation samples, the tangency portfolio weights are estimated using all data up to the end of 2020 and would thus not be feasible to investors in real-time. Columns (5) and (6) of Table 4 report Sharpe ratios for tangency portfolios where feasible portfolio weights are estimated using rolling or expanding windows. Because rolling window tangency portfolio weights become extremely erratic during the COVID crisis, we report rolling window Sharpe ratios for samples that exclude 2019 and 2020.[20] In all cases, expanding and rolling window Sharpe ratio estimates fall considerably relative to the infeasible, full sample estimates and either fail to exceed, or only marginally exceed, the market Sharpe ratio. None of the feasible Sharpe ratio estimates are statistically greater than the market Sharpe ratio.

Results in Table 4 focus on two prominent multifactor models, the Fama-French three-factor model and the Fama-French-Carhart four-factor model. However, several other characteristics-based factors and anomalies were proposed during the 1980s and 1990s that could reasonably be considered as candidate factors for an asset pricing model (e.g., see Figure 1). We therefore consider the subset of 21 anomaly portfolios from the Kozak dataset with publication years prior to the year 2000, and we evaluate the performance of four factor models based on all possible combinations of three anomaly factors with the market factor. A total of 1,330 four-factor models can be generated in this fashion. We evaluate each four-factor model's performance over a validation period spanning January 2000 – December 2019, which is out-of-sample with respect to the original studies proposing all 21 characteristics. Figure 4 presents the estimated Sharpe ratios (left-hand figures) for all 1,330 models as well z-scores for tests of differences between each model's Sharpe ratio and the value-weighted market Sharpe ratio (right-hand figures). We focus on results for feasible tangency portfolios, and report Sharpe ratios and z-scores for expanding window models in Panel A and for rolling window models in Panel B. Models are ordered according to the magnitude of their estimated Sharpe ratio or z-score.

Expanding window Sharpe ratios range from 0.3-0.8 for the majority of models, while rolling window Sharpe ratios generally range from 0.2-0.6. Dashed-red horizontal lines in the left-hand figures indicate the Sharpe ratio of the value-weighted market factor, which is approximately 0.38 during this period. Around one-half of rolling window models have estimated Sharpe ratios that exceed the market factor, while around two-thirds of expanding

[20] Alternatively, we can apply a covariance shrinkage estimator to obtain less erratic SDF weights during the COVID crisis. In unreported results, we confirm that a covariance shrinkage estimator in the spirit of Kozak et al. (2020) also leads to more stable SDF weights during the years 2019 and 2020. However, because the factor variances are extremely large during this period, a *very* extreme degree of shrinkage is required.

window models have Sharpe ratios that exceed the market factor. Z-scores reported in right-hand figures indicate that almost none of the models have Sharpe ratios that are statistically greater than the market Sharpe ratio. Dashed-orange lines correspond to the 95% confidence critical values of 1.96 and -1.96. Among the expanding window models, less than 2% have z-scores over 1.9, and the vast majority fall between 1 and -1. Among the rolling window models, none have z-scores over 1.9 and approximately 2.3% have z-scores below -1.96.

As a point of comparison, we repeat this exercise where all models are evaluated over the pre-validation period from 1964–1999. Models evaluated over this period, which overlaps with samples examined in earlier studies documenting the underlying characteristics, produce much larger Sharpe ratio estimates. Around one-third of these estimates exceed one on an annualized basis and around 20% are statistically greater than the market Sharpe ratio measured over the same period. Explicit results appear in the Internet Appendix.

The evidence in Table 4 and Figure 4 indicates that Sharpe ratios for a large variety of multifactor models estimated using post-publication 'validation samples' that are relatively free of pre-testing bias do not violate conventional good deal bounds. Collectively, this indicates that pre-test biases associated with academic research patterns can be substantial and that estimates of Sharpe ratios that correct for this bias frequently fall to a similar magnitude as the market Sharpe ratio.

4.3 Adaptive model Sharpe Ratios

Many currently popular multifactor models involve characteristics that were only relatively recently proposed. To address models that include such characteristics, we evaluate 'adaptive' versions of characteristics-based models that gradually incorporate new factors as the underlying characteristics are discovered via the research process.[21]

[21]The conceptual standard we adopt in this regard is whether an investment professional might plausibly be aware of the factor based on public academic research output. Importantly, our standard does not preclude the possibility that certain investment professionals adopt portfolios or strategies based on signals that have *not* been previously documented in academic research. But we consider such portfolios or 'factors' as non-public information.

We consider several alternative specific procedures. A first method measures public status via the publication year of the original paper establishing evidence of a cross-sectional relation between expected returns and the firm characteristic that serves as the basis for the factor. The second approach instead uses the ending year of the data sample that is analyzed in the original paper documenting the characteristics. This typically occurs one or more years prior to publication of the paper in an academic journal.

Focusing on the Fama-French three-factor model as an example, the cross-sectional relation between expected returns and size is typically attributed to Banz (1981), and the value (market-book) anomaly to Rosenberg et al. (1985). Using publication years, the adaptive variation of the three-factor model would therefore add the size factor starting in 1981 and the value factor starting in 1985. The alternative approach adds the size factor starting in 1975 and the value factor starting in 1984, as these correspond to the sample ending years in those papers. It is worth emphasizing that either approach results in an adaptive model that includes the size and value factors well prior to the study of Fama and French (1993) that formally proposes the model. Our approach is deliberately intended to be generous, in the sense of permitting the inclusion of a factor as soon as a cross-sectional relation between the underlying characteristic and expected returns appears in academic literature.

Table 5 reports 'adaptive' estimates of Sharpe ratios for many popular characteristics-based models. Columns (1) and (2) show estimated Sharpe ratios for non-adaptive models for reference. Columns (3)–(6) report estimated Sharpe ratios for adaptive versions of the models. For comparability with standard in-sample methods, in columns (3) and (4) we set the tangency portfolio weights in each period equal to the full-sample weights for the set of factors included in the model at that time. This implies that, once the model has evolved to completion (the last discovered factor is added), the tangency portfolio returns from that point onward are identical to the conventional case. Column (3) uses the publication year of the original paper documenting each characteristic as a measure of public status. Column (4) is more conservative and instead uses the (earlier) ending year of the sample analyzed in the original paper. Plug-in estimates of optimal portfolio weights are known to be imprecise even for portfolio decisions involving a relatively small number of assets (factors in our context). Thus, in columns (2), (5), and (6) we consider a simple equal-weighted proxy for the tangency portfolio applied to either the conventional or adaptive versions of factor models.

The decline in Sharpe ratios under the real-time approach is precipitous. Perhaps most strikingly, *none* of the models achieve a Sharpe ratio greater than the market portfolio under the real-time approach using publication year as a measure of public status. Under the alternative sample year-end measure, even the best performing models achieve only small Sharpe ratio increases relative to the market (on the order of 0.02–0.05). Columns (5) and (6) show that similar reductions in Sharpe ratios occur when it is assumed that the tangency portfolio is equal-weighted. Notably, the reduction in Sharpe ratios associated with adopting a real-time evaluation approach is economically much larger than the reduction in Sharpe ratios associated with moving from optimized in-sample tangency portfolio weights to equal weights. This illustrates that, despite the well-known problem of estimation noise in mean-variance optimal portfolio weights, the practice of requiring factors to be in the public information set prior to incorporation in models has a much greater impact on the magnitude of estimated Sharpe ratios for these models.

Table 6 reports Sharpe ratios estimated using OOS designs, where tangency portfolio weights are updated over time using an expanding or rolling window of data. Column (1) shows non-adaptive Sharpe ratio estimates based on this OOS approach using an expanding window that begins with 10 years of data (the initial evaluation period occurs in 1973). Column (2) shows similar results for a 15-year rolling window (the initial evaluation period occurs in 1978). In these cases, the market Sharpe ratio estimate is around 0.47 and 0.54, respectively. Columns (3) and (4) show that incorporating the adaptive real-time filter within the expanding window OOS design produces very different results relative to the non-adaptive model version. Using the publication year as the measure of public information, no model generates an OOS Sharpe ratio greater than 0.5 and all multifactor models fail to outperform the Sharpe ratio of the value-weighted market portfolio. Using the more conservative sample end year measure, some multifactor models do outperform the market, but no model achieves a Sharpe ratio greater than around 0.6.

The final two columns of Table 6 show adaptive results when tangency portfolio weights are updated using an alternative 15-year rolling window. This allows weights to be more flexible over time and incorporates a longer initial window used to estimate weights.[22] The relatively sparse three- and four-factor models do very poorly in this OOS design even before adopting the real-time filter. Columns (5) and (6) show that OOS Sharpe ratios generally fall significantly for adaptive versions of models. Under the 15-year rolling window design, there is more variation in the magnitude of reductions in Sharpe ratios for the adaptive models, and with respect to the publication year versus sample end year measure of public information. Two adaptive models achieve OOS Sharpe ratios in excess of 0.8. However, the sixth row in the panel ($FF5 - HML + UMD$) shows that these higher Sharpe ratios are largely attributable to the role of a single factor (HML) within the models as well as to the inclusion of this factor in the year 1984, one year prior to the publication of the study of Rosenberg et al. (1985). Thus, the favorable results are not particularly robust.

The Internet Appendix provides additional robustness checks and discussion. One important design choice involves the rolling window length or initial sample length used to form tangency portfolio estimates. Qualitatively similar results obtain for window lengths ranging from 10 to 20 years. Tangency portfolio estimates for relatively large factor models (e.g., the Fama-French five- and six-factor models) might suffer deleterious effects from the noise associated with standard plug-in estimates of the factor covariance matrix and factor means. However, we find qualitatively similar results upon using prominent shrinkage estimators for the mean and covariance matrix of factor returns.

Adaptive versions of models add new factors over time. Under the null hypothesis that

[22] We also consider a 10-year rolling window design; however, in this case we find that the portfolio weights become erratic for some models. Despite this, results are qualitatively similar to those in Table 6 and it remains the case that the real-time evaluation approach produces substantially lower Sharpe ratios relative to the conventional OOS values. The Internet Appendix provides further details and explicit results.

the new factors are priced, and holding constant factor means and covariances, the Sharpe ratio of the tangency portfolio for the adaptive model should increase over time relative to the market Sharpe ratio as the model is augmented with additional priced factors. This implies that the adaptive estimates of model Sharpe ratios presented in Tables 5 and 6 may be downward biased relative to the 'true' out-of-sample model Sharpe ratio.

Figure 5 provides visual evidence concerning this hypothesis. The figure plots rolling estimates of the difference between Sharpe ratios for the CAPM and for adaptive versions of two benchmark multifactor models using a 10-year rolling window. Panel A contrasts the CAPM Sharpe ratio with that of the Fama-French six-factor model, while Panel B contrasts the CAPM with the Q4 model. SMB is the first factor added in the adaptive versions of both multifactor models. There is little difference between the Sharpe ratio of the adaptive Q4 model and the CAPM until just after the year 2000, when the Q4 model's Sharpe ratio increases significantly relative to the market Sharpe ratio. After this, however, the Sharpe ratio difference begins to steadily fall, even as the final (investment) factor is added, and the difference ultimately becomes negative, such that the rolling market Sharpe ratio exceeds that of the Q4 model toward the end of the sample. The pattern for the Fama-French six-factor model is qualitatively similar. During the 1990s, the Sharpe ratio of the adaptive model increases significantly relative to the market Sharpe ratio, but again the difference reverses after the early 2000s and is negative by the end of the sample.

Collectively, the evidence suggests that the Sharpe ratios associated with tangency portfolios for 'truly' out-of-sample models either 1) do not increase over time relative to the market Sharpe ratio, 2) increase only marginally, or 3) increase only temporarily and subsequently fall below the market Sharpe ratio by the late 2010s. This suggests that adaptive Sharpe ratio estimates are not substantially downward biased relative to the 'true' out-of-sample Sharpe ratios.

4.4 Alternative model performance metrics

We consider the impact of an adaptive approach on alternative prominent asset pricing performance metrics based on the magnitude of pricing errors. Under this "left hand side" (LHS) approach, models are assessed based on the size of intercepts (unexplained average returns) in time series regressions of test asset returns on the model's factors. We focus on a large set of 275 quintile-sorted portfolios based on size and, independently, beta, book-to-market, operating profitability, investment growth, momentum, long-term reversal, short-term reversal, accruals, net issuance, stock return variance, and residual variance as test assets. This is similar to the set of test assets considered in, e.g., Fama and French (2018).[23] Explicit results appear in the Internet Appendix.

Consistent with prior literature, full sample, non-adaptive results indicate that popular multifactor models substantially reduce pricing errors relative to the market model. The Fama-French five-factor and further extended models perform best; however, model rankings do not line up directly with Sharpe ratio comparisons, similar to findings in Fama and French (2018). In contrast, under a real-time evaluation approach, multifactor models that adaptively incorporate new factors as they are discovered generally only marginally reduce average pricing errors, if they reduce them at all. Results are qualitatively similar for alternative pricing metrics, as well as under alternative rolling or expanding window approaches. Multifactor models generally improve upon the market model, but the economic significance of the improvements is much lower under the real-time adaptive evaluation approach.

5 Factor Models Extracted from the Anomaly Zoo

In the previous section, we focused on the performance of traditional ad-hoc characteristics-based models in the spirit of Fama and French (1993). Kozak et al. (2020) term such models

[23] A drawback of this approach is that results are potentially sensitive to the specific choice of test assets. Therefore, we also consider an alternative set of test assets that consists of the 48 Fama-French industry portfolios, as well as portfolios sorted into deciles based on accruals, net issuance, stock return variance, and residual variance (characteristics not used to form factors in the models we consider).

as “characteristics-sparse,” in the sense that the non-market factors are constructed from a relatively small number of firm characteristics, such as market capitalization and book-to-market ratios. However, several recent studies distill factors from large sets of characteristics or related portfolios by applying forms of regularization or related machine learning methods. Examples of such approaches include Kelly et al. (2019), Freyberger et al. (2020), Kozak et al. (2020), and Lettau and Pelger (2020). Estimates of Sharpe ratios reported in the literature for these approaches are large, even using out-of-sample designs. These approaches typically construct SDFs (or equivalently factors) using an underlying ‘basis set’ of a potentially large set of firm characteristics or portfolios sorted with respect to these characteristics. The basis characteristics are often selected to be those characteristics associated with cross-sectional return patterns in prior literature, and it is this aspect of the research design that introduces the potential of optimistically biased assessments of the performance of the resulting models.

As an illustration, we consider the performance of models based on principle components analysis (PCA) applied to portfolios constructed as sorts on a large number of characteristics from the anomaly literature. We present evidence that Sharpe ratios and related performance metrics for PCA-based models are subject to optimistic biases similar to conventional ad hoc characteristics-based models.

We begin by replicating large Sharpe ratios for statistical factor PCA models constructed using sorts on a large number of characteristics from the anomaly literature. The first and second set of models in Figure 6 summarize the Sharpe ratios associated with PCA models extracted from the set of 88 extreme decile portfolios associated with 44 underlying characteristics from the anomaly literature. Results are presented for models where the number of factors, K, ranges from 1–6. The dashed-black, horizontal line corresponds to the annualized Sharpe ratio of the value-weighted market factor. Consistent with many prior studies, Sharpe ratios for multi-factor PCA models are large both in-sample and using standard out-of-sample designs. The in-sample five-factor PCA model distilled from the set of 88 anomaly portfolios generates an annualized Sharpe ratio of over 1.35, which is over three times the Sharpe ratio of the value-weighted market factor during this period. This estimate falls only marginally to around 1.2 under a standard out-of-sample design.

The first two sets of models presented in Figure 6 are estimated by conditioning the input set of test assets on a set of 88 anomaly portfolios from which factors are distilled. From a real-time perspective, it does not seem realistic to pre-condition (only) on a set of characteristics-based portfolios. Indeed, a subtle manifestation of look-ahead bias involves *omitting* from the set of portfolios used to construct factors various portfolios that, from an ex ante perspective, might reasonably capture variation in stock returns and/or priced risks. In light of this, we augment the set of input tests assets to include not only portfolios sorted on characteristics studied in the academic literature, but on other portfolios as well. Industry-sorted portfolios seem a compelling candidate for inclusion in the set, especially given attention to industry factors in earlier literature. Motivated by Chen et al. (1986) and related studies that explore macroeconomic factors, we also consider portfolios sorted according to estimated betas with respect to a set of macroeconomic factors. (See the Appendix for details concerning the construction of these factors.) The expanded set of basis portfolios thus consists of 48 Fama-French industry-sorted portfolios, 50 macro-risk-based portfolios, and the original 88 extreme decile portfolios based on 44 anomaly characteristics.

The third, fourth, and fifth sets of models in Figure 6 summarize the Sharpe ratios associated with PCA models extracted from the expanded set of portfolios. In-sample Sharpe ratio estimates are even larger, and approach 1.5 for the five- and six-factor models. However, out-of-sample estimates fall by nearly 50% for models with four–six factors and are well under one on an annualized basis. This demonstrates that attributes of factors selected using machine learning and related methods can be sensitive to the choice of included basis portfolios or characteristics. Restricting attention only to a set of characteristics identified in the anomaly literature optimistically biases Sharpe ratios associated with resulting models. When additional, ex ante plausible characteristics such as industry membership dummies and/or betas with respect to macroeconomic risk factors are included in the basis set, Sharpe ratios fall significantly in out-of-sample contexts.

The final set of models in Figure 6 reports Sharpe ratio estimates for PCA models that impose real-time filtering with respect to included firm anomaly characteristics. Sharpe ratio estimates for these models fall further relative to the non-adaptive, out-of-sample models estimated using the expanded basis characteristics, but only by around 0.05-0.1. Consequently, constructing the set of basis portfolios to better reflect ex ante agnosticism concerning whether industry-based, macroeconomic, or characteristics-based factors will better explain the cross-section of returns accounts for the majority of the optimistic bias associated with estimates of Sharpe ratios for statistical factor models.

In the Online Appendix, we consider two alternative, recently proposed methods for estimating statistical factor models. These include the 'risk premium principal components analysis' (RP-PCA) approach proposed by Lettau and Pelger (2020) and the 'shrinking the cross-section' approach proposed by Kozak et al. (2020). We replicate very large Sharpe ratios associated with RP-PCA models distilled from the set of 88 anomaly portfolios. We then show that model Sharpe ratios fall substantially under out-of-sample designs when the basis set of portfolios is augmented to include both industry and macro-risk-based portfolios. However, in contrast to the results presented in Figure 6, non-adaptive multifactor RP-PCA models estimated using the expanded set of input test assets continue to produce out-of-sample Sharpe ratios between 1 and 1.2. This is consistent with the notion that the RP-PCA estimation procedure is better able to identify weak factors compared to standard PCA. When RP-PCA factors are estimated using the expanded set of input test assets and the adaptive filtering procedure, model Sharpe ratios fall further and are generally below 0.75. Thus, inclusion of the characteristics-based anomaly portfolios prior to the publication of the underlying characteristics is critical to the performance of the RP-PCA models.

The Kozak et al. (2020) method of estimating multifactor models shrinks SDF coefficients (tangency portfolio weights) under a prior that implicitly imposes economically-motivated beliefs regarding limitations to the maximum obtainable Sharpe ratio. Interestingly, we show that with a sufficient degree of shrinkage, this approach produces models that are relatively resilient to look-ahead bias and do not violate conventional good deal bounds. Out-of-sample Sharpe ratios are similar across models regardless of the set of input test assets as well as with and without the adaptive filtering procedure.

6 Conclusion

In this paper, we examine the sources of excessively large Sharpe ratios associated with popular multifactor asset pricing models. We show that Sharpe ratios that violate 'good deal bounds' continue to obtain after applying simple, robust estimates of tangency portfolio weights, as well as under conventional pseudo-out-of-sample research designs that rely only on past data. We argue that the most compelling explanation behind excessive Sharpe ratios involves a subtle form of look-ahead bias such that factors included in models, or alternatively the characteristics and portfolios from which factors are extracted, are selected based on prior research outcomes linking such characteristics with cross-sectional variation in returns.

We introduce alternative methods of evaluating models that mitigate the underlying look-ahead bias concerns. First, we consider adaptive versions of ad hoc models that introduce additional factors only after the new factors are 'discovered' by academic researchers. Sharpe ratios for adaptive versions of popular factor models fall dramatically relative to the conventional versions of such models, and often no longer exceed the historical Sharpe ratio associated with the market portfolio. Second, we re-examine the performance of recently proposed methods of extracting factors from a large set of characteristics or portfolios. In this case, we find that the Sharpe ratios associated with the resulting models fall dramatically when the underlying set of characteristics is specified so as to include not only common characteristics from the large stock return anomaly literature, but also ex ante plausible characteristics such as industry membership, betas with respect to macroeconomic shocks, and so forth.

Our results have a variety of implications. Perhaps most importantly, we interpret the much smaller Sharpe ratios associated with popular multifactor models that do *not* violate standard notions of good deal bounds as good news. This is because real-time investors who 'factor invest' using these models after they are proposed do not achieve exorbitant Sharpe ratios. Our results do, however, pose challenges. We find that the performance gap between the CAPM and popular multifactor models becomes much smaller upon adopting evaluation methods that mitigate the central look-ahead biases we highlight. This exacerbates the already challenging problem of factor model specification, and leaves considerable room for future research on this topic.

References

Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected return. Journal of Finance 61, 259–299.

Asness, C., Frazzini, A., 2013. The devil in hml's details. Journal of Portfolio Management 39, 49–68.

Asness, C.S., Frazzini, A., Pedersen, L.H., 2019. Quality minus junk. Review of Accounting Studies 24, 34–112.

Banz, R.W., 1981. The relationship between return and market value of common stocks. Journal of Financial Economics 9, 3–18.

Barbee, W.C., Mukherji, S., Raines, G.A., 1996. Do sales-price and debt-equity explain stock returns better than book-market and firm size? Financial Analysts Journal 52, 56–60.

Barillas, F., Shanken, J., 2017. Which alpha? The Review of Financial Studies 30, 1316–1338.

Barillas, F., Shanken, J., 2018. Comparing asset pricing models. The Journal of Finance 73, 715–754.

Barry, C.B., Brown, S.J., 1984. Differential information and the small firm effect. Journal of Financial Economics 13, 283–294.

Basu, S., 1977. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. Journal of Finance 32, 663—682.

Bessembinder, H., Burt, A., Hrdlicka, C., 2022. Factor returns and out-of-sample alphas: Factor construction matters. Working Paper .

Bessembinder, H., Burt, A., Hrdlicka, C.M., 2021. Time series variation in the factor zoo. Working Paper .

Bhandari, L.C., 1988. Debt/equity ratio and expected common stock returns: Empirical evidence. Journal of Finance 43, 507—528.

Blume, M., Husic, F., 1973. Price, beta, and exchange listing. Journal of Business 28, 283–299.

Campbell, J.Y., Cochrane, J.H., 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. Journal of political Economy 107, 205–251.

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

Chen, A.Y., Zimmermann, T., 2022. Open source cross-sectional asset pricing. Critical Finance Review, 11, 207–264.

Chen, L., Novy-Marx, R., Zhang, L., 2010. An alternative three-factor model. Working Paper .

Chen, N.F., Roll, R., Ross, S.A., 1986. Economic forces and the stock market. Journal of Business , 383–403.

Chordia, T., Goyal, A., Saretto, A., 2020. Anomalies and false rejections. The Review of Financial Studies 33, 2134–2179.

Cochrane, J.H., Saa-Requejo, J., 2000. Beyond arbitrage: Good-deal asset price bounds in incomplete markets. Journal of Political Economy 108, 79–119.

Connor, G., Korajczyk, R.A., 1993. A test for the number of factors in an approximate factor model. the Journal of Finance 48, 1263–1291.

Cooper, M.J., Gulen, H., Schill, M.J., 2008. Asset growth and the cross-section of stock returns. Journal of Finance 63, 1609–1651.

Da, R., Nagel, S., Xiu, D., 2022. The statistical limit of arbitrage. Working Paper .

Da, Z., Liu, Q., Schaumurg, E., 2014. A closer look at the short-term reversal. Management Science 60, 658–674.

Daniel, K., Titman, S., 2006. Market reactions to tangible and intangible. Journal of Finance 61, 1605–1643.

Datar, V.T., Naik, N.Y., Radcliffe, R., 1998. Liquidity and stock returns: An alternative test. Journal of Financial Markets 1, 203–219.

De Bondt, W.F.M., Thaler, R., 1985. Does the stock market overreact? Journal of Finance 40, 793–805.

Dechow, P.M., Sloan, R.G., Soliman, M.T., 2004. Implied equity duration: A new measure of equity risk. Review of Accounting Studies 9, 197–228.

Fairfield, P.M., Whisenant, J.S., Yohn, T.L., 2003. Accrued earnings and growth: Implications for future profitability and market mispricing. Accounting Review 78, 353–371.

Fama, E.F., French, K.R., 1992. The cross-section of expected returns. Journal of Finance 46, 427–466.

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

Fama, E.F., French, K.R., 2006. Profitability, investment and average returns. Journal of Financial Economics 82, 491–518.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E.F., French, K.R., 2018. Choosing factors. Journal of Financial Economics 128, 234–252.

Fama, E.F., MacBeth, J., 1973. Risk, return, and equilibrium: Empirical tests. Journal of Political Economy 81, 607–636.

Frazzini, A., Pedersen, L.H., 2014. Betting against beta. Journal of Financial Economics 111, 1–25.

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Review of Financial Studies 33, 2326–2377.

Haddad, V., Kozak, S., Santosh, S., 2020. Factor timing. Review of Financial Studies 33, 1980–2018.

Hansen, L.P., Jagannathan, R., 1991. Implications of security market data for models of dynamic economies. Journal of political economy 99, 225–262.

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

Harvey, C.R., Liu, Y., 2021. Lucky factors. Journal of Financial Economics 141, 413–435.

Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. Review of Financial Studies 29, 5–68.

Haugen, R.A., Baker, N.L., 1996. Commonality in the determinants of expected stock returns. Journal of Financial Economics 41, 401–439.

Heston, S.L., Sadka, R., 2008. Seasonality in the cross-section of stock returns. Journal of Financial Economics 87, 418–445.

Hirshleifer, D., 2004. Do investors overvalue firms with bloated balance sheets? Journal of Accounting and Economics 38, 297–331.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28, 650–705.

Hou, K., Xue, C., Zhang, L., 2018. Replicating Anomalies. Review of Financial Studies 33, 2019–2133.

Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. Journal of Finance 45, 881–898.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and losers: Implications for stock market efficiency. Journal of Finance 48, 65–91.

Jensen, T.I., Kelly, B.T., Pedersen, L.H., 2022. Is there a replication crisis in finance? Journal of Financial Economics, forthcoming, .

Jobson, J.D., Korkie, B.M., 1981. Performance hypothesis testing with the sharpe and treynor measures. Journal of Finance 36, 889–908.

Kan, R., Robotti, C., Shanken, J., 2013. Pricing model performance and the two-pass cross-sectional regression methodology. The Journal of Finance 68, 2617–2649.

Kan, R., Zhou, G., 2007. Optimal portfolio choice with parameter uncertainty. Journal of Financial and Quantitative Analysis 42, 621–656.

Kelly, B., Pruitt, S., Su, Y., 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics 134, 501–524.

Kozak, S., Nagel, S., Santosh, S., 2018. Interpreting factor models. Journal of Finance 73, 1183–1223.

Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. Journal of Financial Economics 135, 271–292.

Lakonishok, J., Shleifer, A., Vishny, R., 1994. Contrarian investment, extrapolation, and risk. Journal of Finance 44, 1541–1578.

Leamer, E.E., 1978. Specification searches: Ad hoc inference with nonexperimental data. volume 53. John Wiley & Sons Incorporated.

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

Linnainmaa, J.T., Roberts, M.R., 2018. The history of the cross-section of stock returns. Review of Financial Studies 31, 2606—2649.

Lintner, J., 1965. Security prices, risk, and maximal gains from diversification. Journal of Finance 20, 587–615.

Litzenberger, R.H., Ramaswamy, K., 1979. The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence. Journal of Financial Economics 7, 163—195.

Lo, A.W., MacKinlay, A.C., 1990. Data-snooping biases in tests of financial asset pricing models. Review of Financial Studies 3, 431–467.

Lopez-Lira, A., Roussanov, N., 2021. Do common factors really explain the cross-section of stock returns? Working Paper .

Lyandres, E., Sun, L., Zhang, L., 2008. The new issues puzzle: Testing the investment-based explanation. Review of Financial Studies 21, 2825–2855.

MacKinlay, A.C., 1995. Multifactor models do not explain deviations from the capm. Journal of Financial Economics 38, 3–28.

Martin, I.W.R., Nagel, S., 2021. Market efficiency in the age of big data. Working Paper .

McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? Journal of Finance , 5–31.

Moskowitz, T., Grinblatt, M., 1999. Do industries explain momentum? Journal of Finance 54, 1249–1290.

Mossin, J., 1966. Equilibrium in a capital asset market. Econometrica 34, 768–783.

Novy-Marx, R., 2013. The other side of value: The gross profitability premium. Journal of Financial Economics 108, 1–28.

Novy-Marx, R., 2014. Quality investing. Working Paper .

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

Pesaran, M.H., Timmermann, A., 1995. Predictability of stock returns: Robustness and economic significance. Journal of Finance 50, 1201–1228.

Piotroski, J., 2000. Value investing: the use of historical financial statement information to separate winners from losers. Journal of Accounting Research 38, 1—-41.

Pontiff, J., Woodgate, A., 2008. Share issuance and cross-sectional returns. Journal of Finance 63, 921—-945.

Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. Journal of Portfolio Management 11, 9–16.

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

Schwert, G.W., 2003. Anomalies and market efficiency. Handbook of the Economics of Finance , 937–972.

Sharpe, W.F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance 19, 425–442.

Sloan, R., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? The Accounting Review 71, 289–315.

Smith, S., Timmermann, A., 2021. Have risk premia vanished? Journal of Financial Economics 145, 553–576.

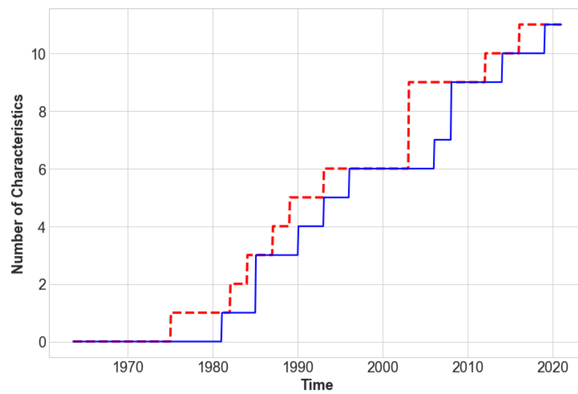
Soliman, M.T., 2008. The use of dupont analysis by market participants. Accounting Review 83, 823—-853.

Xing, Y., 2008. Interpreting the value effect through the q-theory: An empirical investigation. Review of Financial Studies 21, 1767–1795.

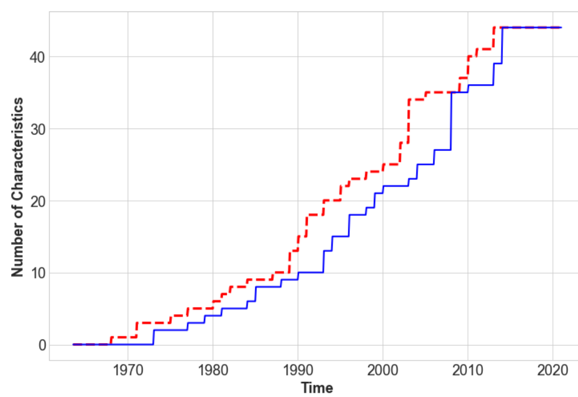
Figure 1: **Number of Publicly Available Anomaly Signals**

This figure reports the number of publicly available factors and anomaly portfolios, where public status is defined using publication dates (solid blue lines) and sample end dates (dashed red lines) for each set of factors and anomaly portfolios. In Panel A, summary statistics are based on the 11 characteristics-based factors reported in Panel A of Table 1. In Panel B, summary statistics are based on the 44 anomaly signals from Serhiy Kozak's website and reported in Panel B of Table 1. In Panel C, summary statistics are based on the 326 anomaly signals documented by Chen and Zimmermann (2022).

(a) Characteristics-Based Factors



(b) Anomaly Decile Portfolios from Serhiy Kozak Data



(c) Anomaly Signals Documented by Chen and Zimmermann (2022)

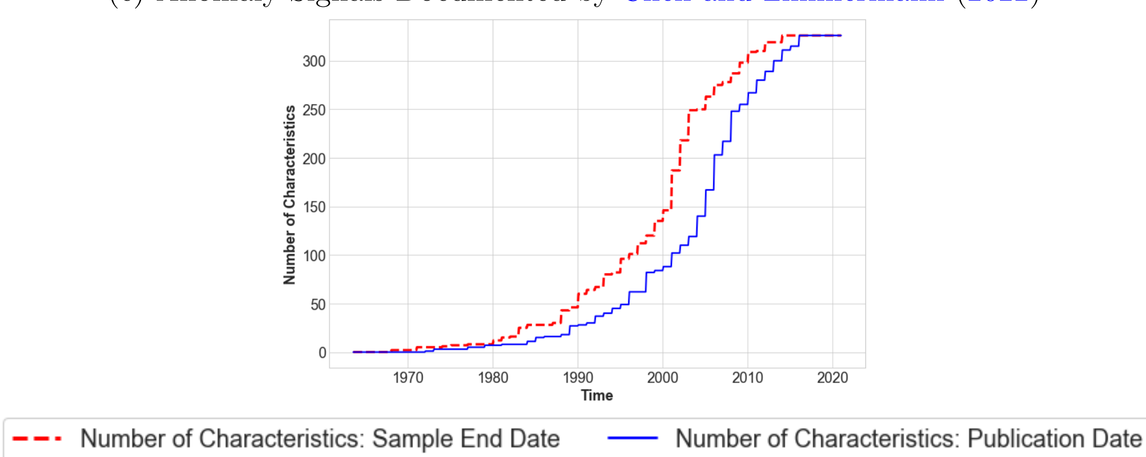


Figure 2: **In-Sample Observations in 'Out-of-Sample' Evaluation Periods**

This figure reports the fraction of in-sample observations in 'out-of-sample' evaluation periods for a variety of characteristics-based factor models and for the sample of 44 anomaly characteristics described in Panel B of Table 1. Factor-month or anomaly-month observations are defined as in-sample for all months from the beginning of the sample until the publication year of the original paper documenting the anomaly signal. 'Out-of-sample' evaluation periods are defined as the periods over which rolling or expanding window 'out-of-sample' (OOS) tangency portfolio returns can be constructed, where the rolling or expanding windows are defined using 10 or 20 years and the total sample period extends from 1963–2020 (e.g., a 10-year window implies that the OOS evaluation period extends from 1973–2020). Data for the 44 anomaly portfolio returns extends from 1964–2019, and therefore the Out-of-Sample Evaluation Periods are 1974–2019 (red bar) and 1984–2019 (blue bar) for this subset of characteristics.

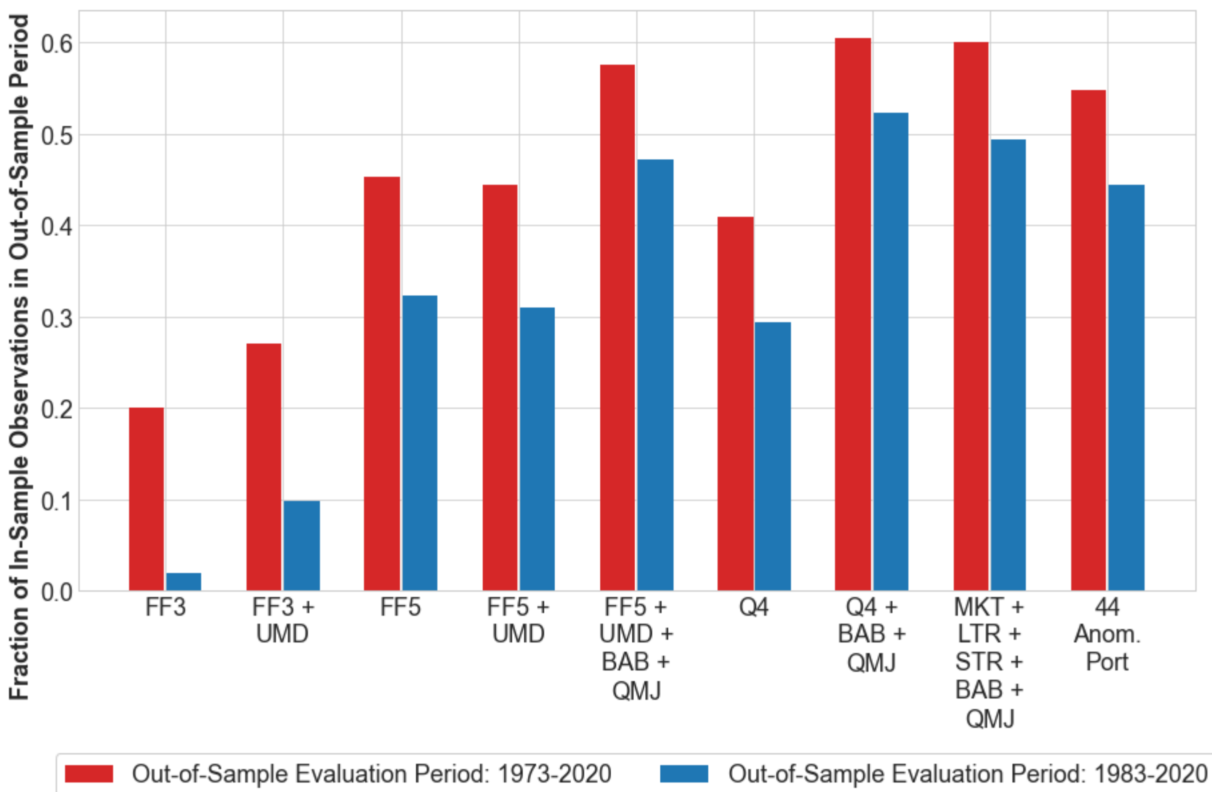


Figure 3: **Pre- and Post-Publication Sharpe Ratios for Popular Characteristics-Based Factors and Anomaly Decile Portfolios**

This figure reports annual Sharpe Ratios pre- and post-publication for popular characteristics-based factors and long-short anomaly decile portfolios with at least 20 years of post-publication data. The total sample period from the characteristics-based factor returns is 07.1963 – 12.2020. The total sample period from the anomaly decile portfolio returns is 07.1964 – 12.2019. Publication dates are listed in Table 1. Blue bars indicate the pre-publication Sharpe Ratio. Orange bars indicate the post-publication Sharpe Ratio. Vertical black lines indicate a 95% confidence interval. Factors and long-short anomaly portfolios are sorted according to their full sample Sharpe ratios.

(a) Popular Characteristics-Based Factors

(b) Anomaly Decile Portfolios

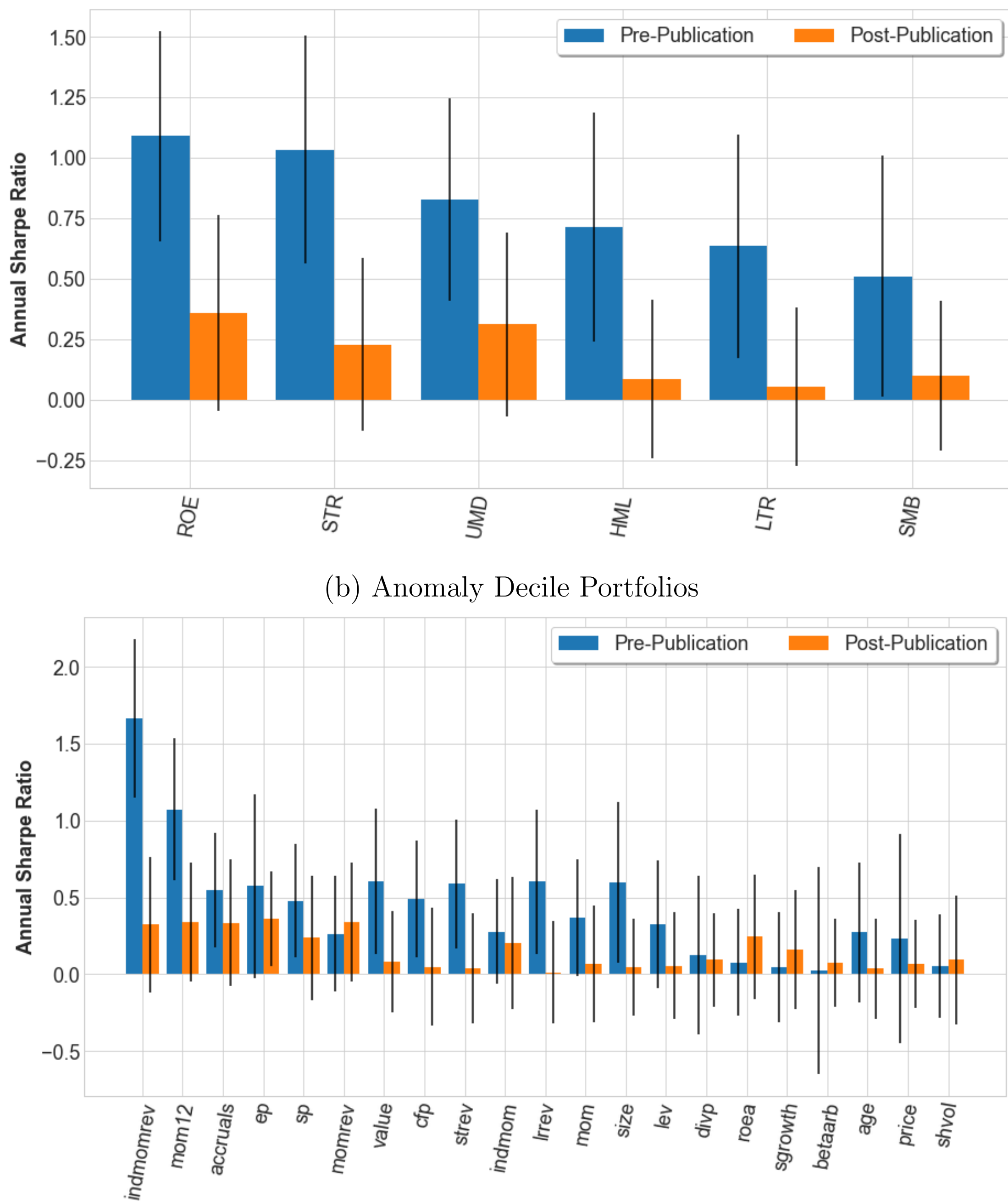
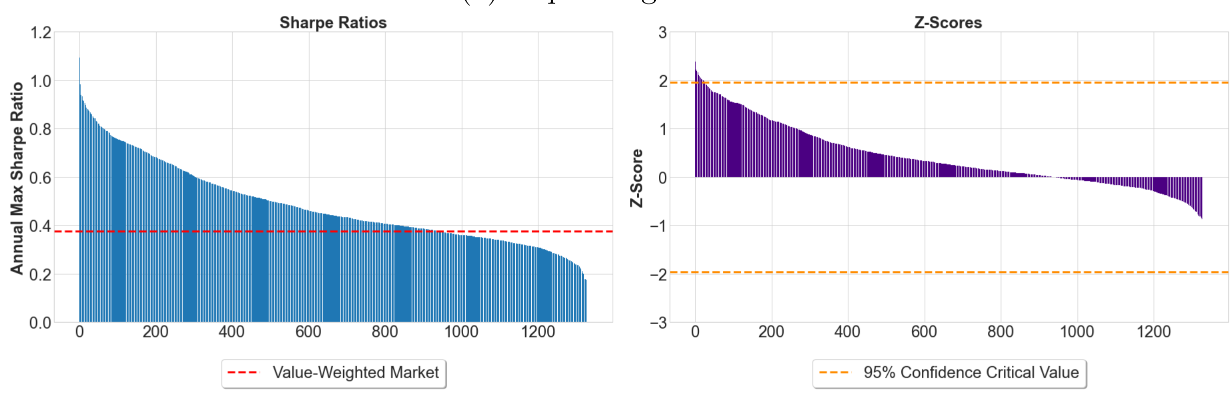


Figure 4: **Validation Data Sharpe Ratios and Z-Scores for Differences Between Model Sharpe Ratios and the Market Sharpe Ratio**

This figure reports model performance statistics for all combinations of three-factor models that can be constructed from the set of 21 anomalies in the Kozak dataset with post-publication data spanning a minimum of 20 years (i.e., all anomalies that were published before the year 2000). This implies a total of 1,330 unique models. Factors in the models are constructed as long-short decile portfolios. Each three-factor model is augmented with the value-weighted Market factor, and is evaluated over a validation period spanning January 2000 – December 2019, which is out-of-sample with respect to the original studies proposing all 21 characteristics. Left-hand figures report the maximum Sharpe ratio associated with each model, in descending order. Dashed-red lines correspond to the Sharpe ratio of the value-weighted market factor measured over the evaluation period, 2000–2019. Right-hand figures report the z-score for a test of the difference between each model's Sharpe ratio and the market Sharpe ratio, in descending order. Dashed-orange lines correspond to the 95% confidence critical value of 1.96. In Panel A, tangency portfolio weights are constructed using an expanding window, where mean and covariance estimates for all factors measured over an expanding window from July 1964 though month $m-1$. In Panel B, tangency portfolio weights are computed using a rolling window, where mean and covariance estimates for all factors measured over a 15-year rolling window from months $m-180 : m-1$. For both rolling and expanding windows, OOS month m portfolio returns begin in January 2000. Sharpe ratios are estimated using monthly factor returns, and are then annualized by multiplying by $\sqrt{12}$.

(a) Expanding Window



(b) Rolling Window

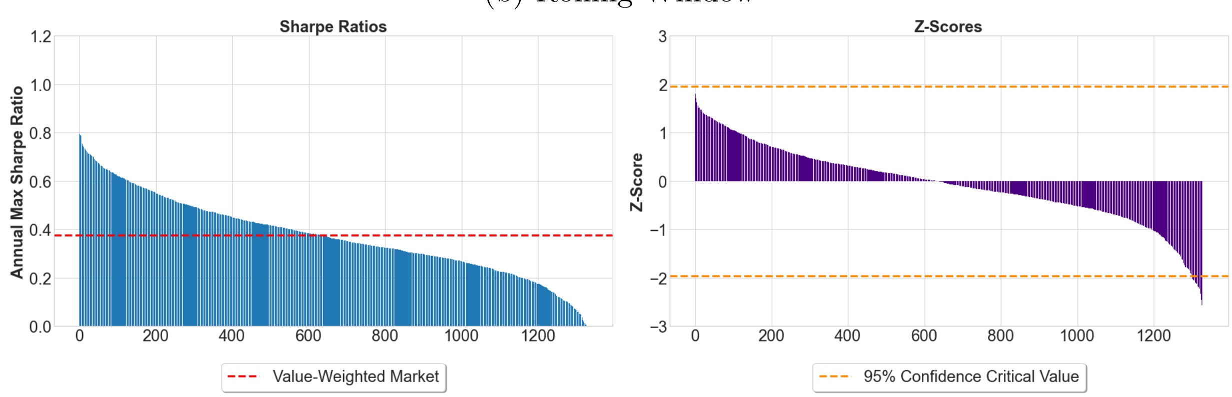
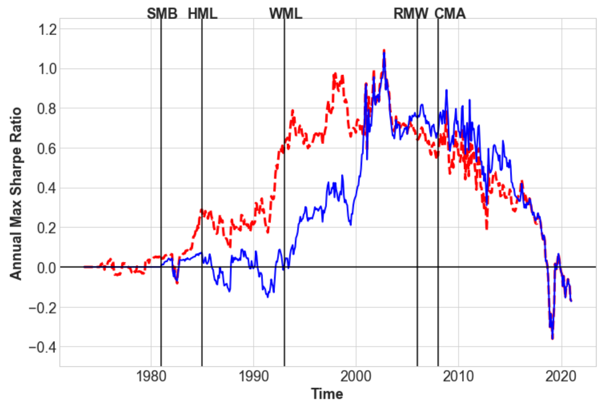


Figure 5: **Differences Between Maximum Sharpe Ratios Associated with Adaptive Versions of Characteristics-Based Factor Models and the Market Sharpe Ratio**

This figure reports differences between rolling estimates of annual maximum Sharpe Ratios for adaptive factor models and a rolling estimate of the market Sharpe ratio. Panel A compares the $FF5 + UMD$ model Sharpe ratio to the market Sharpe ratio. Panel B compares the $Q4$ model Sharpe ratio to the market Sharpe ratio. The time-varying Sharpe ratio for the market model is computed using 10-year rolling mean and covariance estimates of the value-weighted market return. The 'Matched Weight' tangency portfolios are constructed using mean and covariance estimates for all factors in the model measured over the full sample period. Weights for the available factors in each month m are estimated to maximize the Sharpe Ratio using the full-sample measures of the means and covariances, and are applied to available factor returns in m. The time-varying Sharpe ratios for the 'Matched Weight' are then computed using 10-year rolling mean and covariance estimates of this tangency portfolio. Dashed red lines report the difference between the time-varying Sharpe ratio for the 'Matched Weight' tangency portfolio that incorporates each factor beginning in its publication year and the time-varying market Sharpe ratio. Solid blue lines report the difference between the time-varying Sharpe ratio for the 'Matched Weight' tangency portfolio that incorporates each factor beginning in its sample end year and the time-varying market Sharpe ratio. Vertical black lines indicate the publication year for each factor in the model.

(a) FF5 + UMD



(b) Q4

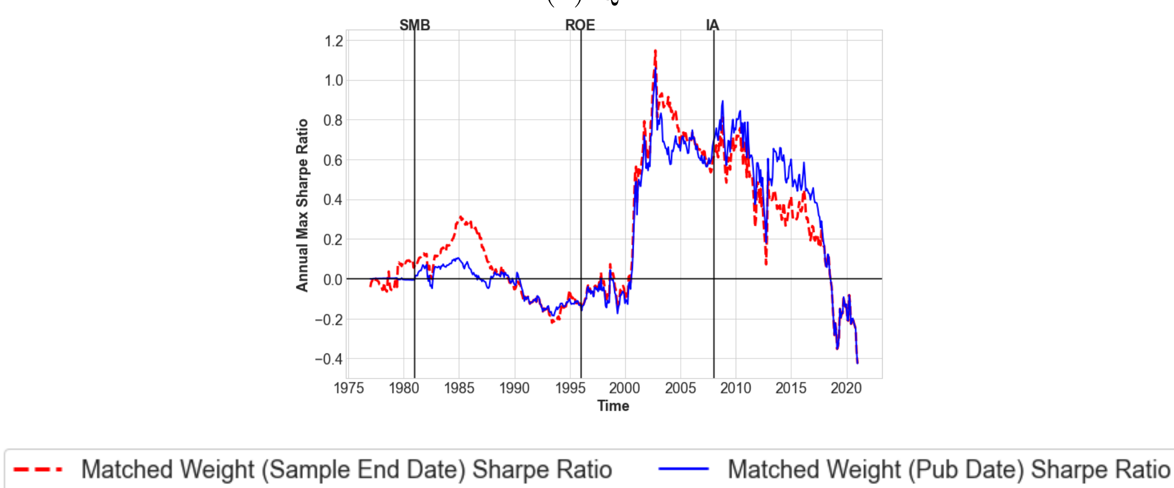


Figure 6: **Annual Maximum Sharpe Ratios for Statistical Factor Models**

This figure reports annual maximum Sharpe ratios for various non-adpative and adaptive PCA models, where the number of factors K varies from 1-6. The input set of tests assets for the 'Anomaly Portfolios Only' models includes the 88 anomaly extreme decile portfolios. The input set of tests assets for the 'Anomaly, Industry, and Macro Portfolios' models includes the 88 anomaly extreme decile portfolios, the 48 Fama-French industry portfolios, and the 50 macro portfolios. The sample period in all cases extends from 07.1964 – 12.2019. All factors are estimated using standard PCA. SDF portfolio weights are equal to their mean-variance values. All out-of-sample results are estimated using 15-year rolling windows. X-axis labels indicate whether models are estimated using the full sample of data or a 15-year rolling window, whether rolling window models are non-adaptive or adaptive, and whether adaptive models use the publication date filter or the sample end date filter. The dashed horizontal black line indicates the annualized Sharpe ratio for the value-weighted market factor estimated over the same period.

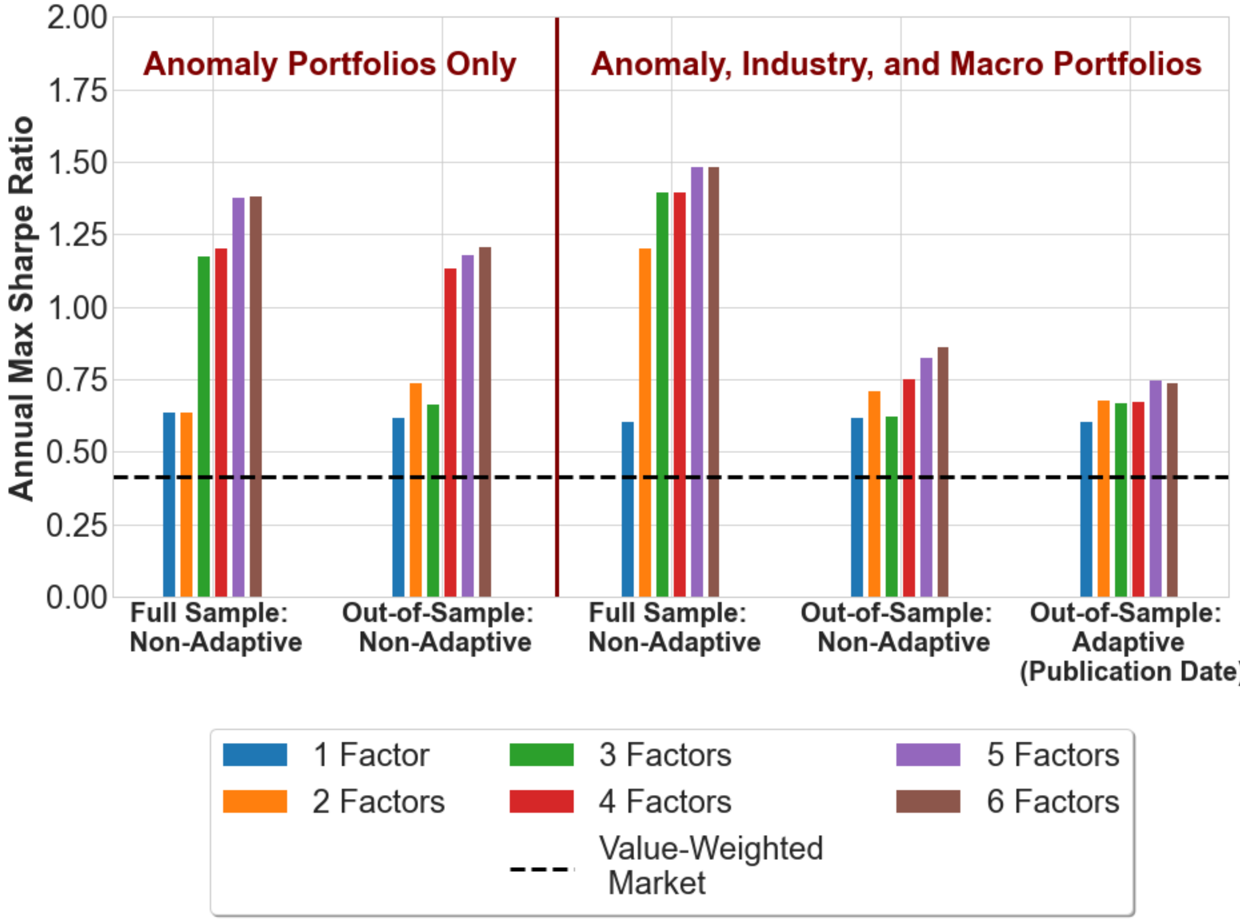


Table 1: **Factor Sources**

This table reports citations and descriptions for all factors and anomaly signals. Original citations and sample periods are obtained from Chen and Zimmerman's Open Source Asset Pricing. In Panel A, the 'Source' column indicates where we obtain the (monthly) factor returns. 'KFD' refers to the Ken French Data library. 'G-q' refers to Global-q data library. 'AQR' refers to AQR's data library. In Panel B, all decile portfolio returns are obtained from Serhiy Kozak's website.

Panel A: Characteristics-Based Factors

Factor Acronym	Characteristic Description	Original Citation	Sample	Original Source
BAB	Betting Against Beta, defined as the daily beta from rolling regressions of excess returns on market excess returns	Frazzini and Pedersen (2014)	1929–2012	AQR
CMA	Investment, defined as the annual change in total assets divided by one-year-lagged total assets. Factor constructed from a 2-by-3 sort on size and investment.	Cooper, Gulen and Schill (2008)	1968–2003	KFD
HML	Book-to-Market, defined as the book value of equity from the t-1 fiscal year end divided by market equity es of the end of December in year t-1.	Rosenberg, Reid and Lanstein (1985)	1973–1984	KFD
IA	Investment, defined as the annual change in total assets divided by one-year-lagged total assets. Factor constructed from a triple 2-by-3-by-3 sort on size, I/A, and ROE.	Cooper, Gulen and Schill (2008)	1968–2003	G-q
LTR	Long-Term Reversal, defined as the prior return measured over months -60:-13.	De Bondt and Thaler (1985)	1929–1982	KFD
QMJ	Quality, defined as the average of profitability, safety, and growth scores.	Asness, Frazzini and Pedersen (2019)	1957–2016	AQR
RMW	Operating Profitability, defined as annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses divided by book equity for the last fiscal year end in t-1.	Fama and French (2006)	1977–2003	KFD
ROE	Return on Equity, defined as quarterly income before extraordinary items divided by one-quarter-lagged book equity.	Haugen and Baker (1996)	1979–1993	G-q
SMB	Size, defined as market equity as of the end of June.	Banz (1981)	1926–1975	KFD
STR	Short-Term Reversal, defined as the prior return measured over month -1.	Jegadeesh (1990)	1934–1987	KFD
UMD	Momentum, defined as the prior return measured over months -12:-2.	Jegadeesh and Titman (1993)	1964–1989	KFD

Table 1 cont.

Panel B: Anomaly Decile Portfolios from Serhiy Kozak Data

	Acronym	Characteristic	Original Citation	Original Sample
1.	accruals	Accruals	Sloan (1996)	1962–1991
2.	age	Firm Age	Barry and Brown (1984)	1931–1980
3.	aturnover	Asset Turnover	Soliman (2008)	1984–2002
4.	betaarb	Beta Arbitrage	Fama and MacBeth (1973)	1929–1968
5.	cfp	Cash Flow / Market Equity	Lakonishok, Shleifer and Vishny (1994)	1968–1990
6.	ciss	Composite Issuance	Daniel and Titman (2006)	1968–2003
7.	divp	Dividend Yield	Litzenberger and Ramaswamy (1979)	1936–1977
8.	dur	Cash Flow Duration	Dechow, Sloan and Soliman (2004)	1962–1998
9.	ep	Earnings / Price	Basu (1977)	1957–1971
10.	fscore	Piotroski's F-score	Piotroski (2000)	1976–1996
11.	gltnoa	Growth in Long-term NOA	Fairfield, Whisenant and Yohn (2003)	1964–1993
12.	gmargins	Gross Margins	Novy-Marx (2013)	1963–2010
13.	growth	Asset Growth	Cooper et al. (2008)	1968–2003
14.	igrowth	Investment Growth	Xing (2008)	1964–2003
15.	indmom	Industry Momentum	Moskowitz and Grinblatt (1999)	1963–1995
16.	indmomrev	Industry Momentum-Reversal	Moskowitz and Grinblatt (1999)	1963–1995
17.	indrrev	Industry Relative Reversal	Da, Liu and Schaumurg (2014)	1982–2009
18.	indrrevlv	Ind. Rel. Reversal (Low Volatility)	Da, Liu and Schaumurg (2014)	1982–2009
19.	inv	Investment	Lyandres et al. (2008)	1970–2005
20.	invcap	Investment-to-Capital	Xing (2008)	1964–2003
21.	ivol	Idiosyncratic Volatility	Ang, Hodrick, Xing and Zhang (2006)	1963–2000
22.	lev	Leverage	Bhandari (1988)	1952–1981
23.	lrrev	Long-term Reversal	De Bondt and Thaler (1985)	1929–1982
24.	mom	Momentum (6-month)	Jegadeesh and Titman (1993)	1964–1989
25.	momrev	Momentum-Reversal	Jegadeesh and Titman (1993)	1964–1989
26.	mom12	Momentum (1-year)	Jegadeesh and Titman (1993)	1964–1989
27.	nissa	Share Issuance (annual)	Pontiff and Woodgate (2008)	1970–2003
28.	nissm	Share Issuance (monthly)	Pontiff and Woodgate (2008)	1970–2003
29.	noa	Net Op. Assets	Hirshleifer (2004)	1964–2002
30.	price	Price	Blume and Husic (1973)	1932–1971
31.	prof	Gross Profitability	Novy-Marx (2013)	1963–2010
32.	roaa	Return on Assets (annual)	Chen, Novy-Marx and Zhang (2010)	1972–2010
33.	roea	Return on Equity (annual)	Haugen and Baker (1996)	1979–1993
34.	season	Seasonality	Heston and Sadka (2008)	1965–2002
35.	sgrowth	Sales Growth	Lakonishok, Shleifer and Vishny (1994)	1968–1990
36.	shvol	Share Volume	Datar, Naik and Radcliffe (1998)	1962–1991
37.	size	Size	Banz (1981)	1926–1975
38.	sp	Sales-to-Price	Barbee et al. (1996)	1979–1991
39.	strev	Short-term Reversal	Jegadeesh (1990)	1934–1987
40.	valmom	Value-Momentum	Novy-Marx (2014)	1963–2013
41.	valmomprof	Value-Momentum-Profitability	Novy-Marx (2014)	1963–2013
42.	valprof	Value-Profitability	Novy-Marx (2014)	1963–2013
43.	value	Value (annual)	Rosenberg, Reid and Lanstein (1985)	1973–1984
44.	valuem	Value (monthly)	Asness and Frazzini (2013)	1950–2011

Table 2: **Maximum Sharpe Ratios for Characteristics-Based Factor Models**

This table reports maximum Sharpe Ratios for various full-sample and alternative 'out-of-sample' factor models. Plug-In tangency portfolio weights are computed using the full sample (1963–2020) mean and covariance estimates for all factors in the row labels. Equally Weighted tangency portfolios are constructed by equally weighting all factors in the row labels. KZ Bias Corrected Sharpe ratios are estimated using the Kan and Zhou (2007) biased-corrected estimator and the plug-in mean-variance tangency portfolio returns. Rolling Window tangency portfolio weights are computed using mean and covariance estimates for all factors measured over a 15-year rolling window from months $m - 180 : m - 1$. Optimal portfolio weights are calculated to maximize the Sharpe Ratio, and applied to returns in month m to form the 'out-of-sample' optimal portfolio. Expanding Window tangency portfolio weights are constructed using mean and covariance estimates for all factors measured over an expanding window up to month $m - 1$. Optimal portfolio weights are computed to maximize the Sharpe Ratio, and applied to returns in month m to form the 'out-of-sample' optimal portfolio. A minimum of 10 years is required. The reported Sharpe Ratios for the rolling (expanding) windows are computed from the month m portfolio returns, which are available over the 1978–2020 (1973–2020) sample. The sample period begins in 1967, instead of 1963, for all models denoted with a *. Sharpe ratios are estimated using monthly factor returns, and are then annualized by multiplying by $\sqrt{12}$.

	Plug-In	Equally Weighted	KZ Bias Correction	Expanding Window	Rolling Window
MKT	0.4406	0.4406	0.4194	0.4776	0.5447
$FF3$	0.6189	0.5802	0.5728	0.5513	0.2082
$FF3 + UMD$	0.9550	0.8746	0.9136	0.8321	0.3008
$FF5$	1.0598	0.8775	1.0123	0.9451	1.0254
$FF5 + UMD$	1.2230	1.1328	1.1722	1.0553	1.0770
$FF5 - HML + UMD$	1.2218	1.0438	1.1793	1.1130	1.0354
$FF5 + UMD + BAB + QMJ$	1.5786	1.3433	1.5220	1.3817	1.4926
$Q4$*	1.3611	1.0686	1.3271	1.2269	1.1443
$Q4 + BAB + QMJ$*	1.6140	1.4782	1.5690	1.5171	1.5883
$MKT + LTR + STR + BAB + QMJ$	1.4558	1.2399	1.4181	1.3262	1.4070

Table 3: **Post-Publication Reduction in Factor and Anomaly Portfolio Returns**

This table reports summary statistics for all long-short characteristics-based factors. Both Global-q factors (IA and ROE, denoted with a *) are available from 01.1967 – 12.2020. The sample period for all other characteristics-based factors extends from 07.1963 – 12.2020. The final row ('44 Anomaly Long-Short Decile Portfolios') reports summary statistics for the 44 long-short anomaly decile portfolios. The sample period for the 44 anomaly returns extends from 07.1964 – 12.2019. All average monthly returns are expressed in percentage points, and standard errors are reported in parentheses. ***, **, and * represent statistical significance at the 1%, 5%, and 10% levels, respectively.

Factor Acronym	Characteristic	Publication Year	Mean Return	Pre-Publication Return	Post-Publication Return	Difference
MKT	Market		0.5679 ***			
			(0.170)			
BAB	Betting Against Beta	2014	0.8040 ***	0.8282 ***	0.6290 **	-0.1992
			(0.125)	(0.137)	(0.272)	(0.305)
CMA	Investment	2008	0.2563 ***	0.3359 ***	-0.0165	-0.3524 **
			(0.076)	(0.091)	(0.120)	(0.150)
HML	Book-to-Market	1985	0.2510 **	0.5484 ***	0.0734	-0.4749 **
			(0.109)	(0.165)	(0.143)	(0.218)
IA*	Investment	2008	0.3356 ***	0.4622 ***	-0.0635	-0.5257 ***
			(0.074)	(0.088)	(0.129)	(0.156)
LTR	Long-Term Reversal	1985	0.1963 **	0.4609 ***	0.0383	-0.4226 **
			(0.096)	(0.156)	(0.121)	(0.197)
QMJ	Quality Minus Junk	2019	0.3579 ***	0.3917 ***	-0.5792	-0.9709 *
			(0.086)	(0.087)	(0.564)	(0.570)
RMW	Operating Profitability	2006	0.2487 ***	0.2517 **	0.2403 **	0.0114
			(0.082)	(0.104)	(0.117)	(0.157)
ROE*	Return on Equity	1996	0.5059 ***	0.6785 ***	0.3056 *	-0.3729 *
			(0.101)	(0.115)	(0.170)	(0.206)
SMB	Size	1981	0.2020 *	0.4677 **	0.0858	-0.3819
			(0.116)	(0.218)	(0.136)	(0.257)
STR	Short-Term Reversal	1990	0.4802 ***	0.7681 ***	0.2341	-0.5340 **
			(0.120)	(0.144)	(0.184)	(0.234)
UMD	Momentum	1993	0.6370 ***	0.8227 ***	0.4414 *	-0.3813
			(0.160)	(0.183)	(0.267)	(0.323)
All Characteristics-Based Factors			0.3883 ***	0.5332 ***	0.1647 ***	-0.3685 ***
(Excluding MKT)			(0.032)	(0.039)	(0.057)	(0.069)
44 Anomaly Long-Short Decile Portfolios			0.4038 ***	0.5488 ***	0.1588 ***	-0.3900 ***
			(0.027)	(0.031)	(0.049)	(0.058)

Table 4: **Validation Data Sharpe Ratios for Early Fama-French Factor Models**

This table reports maximum Sharpe Ratios for the Fama-French three-factor model and the Fama-French-Carhart four-factor model. All models are evaluated over a validation period that is out-of-sample with respect to the original studies proposing the characteristics or factor models. The 'Evaluation Period' column reports the validation data period. Model public status is defined as: the year in which the final characteristic (value in Panel A, and momentum in Panel B) was published, the sample end year of the study documenting the final characteristic, or the sample end year of the paper that proposed the model (Fama and French (1993) in Panel A, and Carhart (1997) in Panel B). Sharpe ratios are evaluated from January in the year following the 'public' year through December 2020, or December 2018 in cases where we drop COVID years. Plug-in tangency portfolio weights are computed using the full evaluation period mean and covariance estimates. KZ Bias Corrected Sharpe ratios are estimated using the Kan and Zhou (2007) biased-corrected estimator and the plug-in tangency portfolio weights. Equally Weighted tangency portfolios are constructed by equally weighting all factors. Rolling Window tangency portfolio weights are computed using mean and covariance estimates for all factors measured over a 15-year rolling window from months $m-180 : m-1$. Expanding Window tangency portfolio weights are constructed using mean and covariance estimates for all factors measured over an expanding window from July 1963 though month $m-1$. For both rolling and expanding windows, OOS month m portfolio returns begin in January in the first year of the evaluation period. Plug-In weights are 'infeasible' because they are estimated using data through the end of 2020, and therefore an investor could not have obtained them in real-time. Rolling and expanding window weights are 'feasible' because they are estimated using only historical data, and therefore an investor could have obtained them in real-time. The 'Market' column reports the market Sharpe ratio measured over the evaluation period. Sharpe ratios are estimated using monthly factor returns, and are then annualized by multiplying by $\sqrt{12}$. I.I.D. standard errors are reported in parentheses. Z-scores for tests of differences between reported Sharpe ratios and the market Sharpe ratio are reported in brackets.

Panel A: Fama-French Three-Factor Model Sharpe Ratios

		Infeasible Weights		Feasible Weights			
	Evaluation Period	Plug-In	KZ Bias Correction	Equally Weighted	Expanding Window	Rolling Window (Drop COVID)	Market
Final Characteristic Publication Date	1986–2020	0.5883	0.5063	0.4896	0.3975	0.2532	0.5604
		(0.1702)	(0.1699)	(0.1699)	(0.1696)	(0.1743)	(0.1869)
		[0.5285]		*[-0.6328]*	*[-0.9546]*	*[-1.4627]*	
Final Characteristic Sample End Date	1985–2020	0.6162	0.5403	0.5114	0.4014	0.2605	0.5864
		(0.1680)	(0.1677)	(0.1676)	(0.1672)	(0.1717)	(0.1870)
		[0.5659]		*[-0.6773]*	*[-1.0844]*	*[-1.5951]*	
Factor Model Sample End Date	1992–2020	0.6109	0.5142	0.5504	0.4705	0.2804	0.5788
		(0.1871)	(0.1867)	(0.1869)	(0.1865)	(0.1928)	(0.1870)
		[0.5247]		*[-0.2263]*	*[-0.5996]*	*[-1.2674]*	

Panel B: Fama-French-Carhart Four-Factor Model Sharpe Ratios

		Infeasible Weights		Feasible Weights			
	Evaluation Period	Plug-In	KZ Bias Correction	Equally Weighted	Expanding Window	Rolling Window (Drop COVID)	Market
Final Characteristic Publication Date	1994–2020	0.7704	0.6591	0.6541	0.5472	0.4990	0.5706
		(0.1948)	(0.1942)	(0.1942)	(0.1936)	(0.2010)	(0.1870)
		[1.4149]		*[0.4382]*	*[-0.0933]*	*[-0.0239]*	
Final Characteristic Sample End Date	1990–2020	0.8232	0.7333	0.7293	0.6339	0.5757	0.5641
		(0.1821)	(0.1816)	(0.1816)	(0.1811)	(0.1870)	(0.1869)
		[1.7833]		*[0.9399]*	*[0.3097]*	*[0.3309]*	
Factor Model Sample End Date	1994–2020	0.7704	0.6591	0.6541	0.5472	0.4990	0.5706
		(0.1948)	(0.1942)	(0.1942)	(0.1936)	(0.2010)	(0.1870)
		[1.4149]		*[0.4382]*	*[-0.0933]*	*[-0.0239]*	

Table 5: **Full Sample Maximum Sharpe Ratios for Adaptive Versions of Characteristics-Based Factor Models**

This table reports full sample maximum Sharpe Ratios for adaptive and non-adaptive versions of factor models. Plug-In tangency portfolio weights are computed using the full sample (1963–2020) mean and covariance estimates for all factors in the row labels. Equally Weighted tangency portfolios are constructed by equally weighting all factors in the row labels. Matched Weight tangency portfolio weights are computed using full sample plug-in mean and covariance estimates for all factors in the row labels. Weights for the available factors in each month m are estimated to maximize the Sharpe Ratio using the full-sample measures of the means and covariances, and are applied to available factor returns in m. The sample period begins in 1967, instead of 1963, for all models denoted with a *. In the 'Non-Adaptive' columns, all factors in the row labels are included at all time periods. In the Adaptive 'Pub Date' ('Sample End') columns, factors in the row labels do not enter the sample until their publication year (their original sample end year). Sharpe ratios are estimated using monthly factor returns, and are then annualized by multiplying by $\sqrt{12}$.

	Non-Adaptive		Adaptive			
			Matched Weight		Equally Weighted	
	Plug-In	Equally Weighted	Pub Date	Sample End	Pub Date	Sample End
MKT	0.4406	0.4406	0.4406	0.4406	0.4406	0.4406
$FF3$	0.6189	0.5802	0.3317	0.3721	0.3336	0.3798
$FF3 + UMD$	0.9550	0.8746	0.3702	0.4435	0.3686	0.4468
$FF5$	1.0598	0.8775	0.3933	0.4086	0.3436	0.3738
$FF5 + UMD$	1.2230	1.1328	0.4142	0.4763	0.3739	0.4411
$FF5 - HML + UMD$	1.2218	1.0438	0.4172	0.4533	0.3802	0.4248
$FF5 + UMD + BAB + QMJ$	1.5786	1.3433	0.4172	0.4998	0.3833	0.4611
$Q4$*	1.3611	1.0686	0.3745	0.3982	0.3844	0.4159
$Q4 + BAB + QMJ$*	1.6140	1.4782	0.3887	0.4275	0.3945	0.4368
$MKT + LTR + STR + BAB + QMJ$	1.4558	1.2399	0.3184	0.3534	0.3332	0.3605

Table 6: **Out-of-Sample Maximum Sharpe Ratios for Adaptive Versions of Characteristics-Based Factor Models**

This table reports out-of-sample maximum Sharpe Ratios for adaptive and non-adaptive versions of factor models. Rolling Window tangency portfolio weights are computed using mean and covariance estimates for all factors measured over a 15-year rolling window from months $m-180:m-1$. Expanding Window tangency portfolio weights are constructed using mean and covariance estimates for all factors measured over an expanding window up to month $m-1$. A minimum of 10 years is required. Rolling and expanding window optimal portfolio weights are applied to returns in month m to form the 'out-of-sample' optimal portfolio. The reported Sharpe Ratios for the rolling (expanding) windows are computed from the month m portfolio returns, which are available over the 1978–2020 (1973–2020) sample. The sample period begins in 1967, instead of 1963, for all models denoted with a *. In the 'Non-Adaptive' columns, all factors in the row labels are included at all time periods. In the Adaptive 'Pub Date' ('Sample End') columns, factors in the row labels do not enter the sample until their publication year (their original sample end year). Sharpe ratios are estimated using monthly factor returns, and are then annualized by multiplying by $\sqrt{12}$.

	Non-Adaptive		Adaptive			
	10-Year Expanding Window	15-Year Rolling Window	10-Year Expanding Window		15-Year Rolling Window	
			Pub Date	Sample End	Pub Date	Sample End
MKT	0.4776	0.5447	0.4776	0.4776	0.5447	0.5447
$FF3$	0.5513	0.2082	0.3161	0.3859	0.2083	0.2083
$FF3+UMD$	0.8321	0.3008	0.3789	0.5099	0.3007	0.3008
$FF5$	0.9451	1.0254	0.4109	0.4755	0.6200	0.6702
$FF5+UMD$	1.0553	1.0770	0.4433	0.5765	0.6972	0.8407
$FF5-HML+UMD$	1.1130	1.0354	0.3483	0.4328	0.4345	0.4840
$FF5+UMD+BAB+QMJ$	1.3817	1.4926	0.4432	0.6023	0.6817	0.8560
$Q4$*	1.2269	1.1443	0.3972	0.5556	0.5259	0.4805
$Q4+BAB+QMJ$*	1.5171	1.5883	0.4155	0.5970	0.5445	0.5263
$MKT+LTR+STR+BAB+QMJ$	1.3262	1.4070	0.3155	0.3943	0.3932	0.4608